PANNZER 2: Annotate a complete proteome in minutes!

Alan J Medlar[#], Petri Törönen[#], Elaine Zosa, Liisa Holm*
Structural Genomics Group, Institute of Biotechnology
PO Box 56, 00014 University of Helsinki, Finland
[#]These authors contributed equally
*To whom correspondence should be addressed: liisa.holm@helsinki.fi

1. INTRODUCTION

As high-throughput sequencing has become increasingly efficient, downstream analysis has become the main bottleneck in genome sequencing projects. For example, a critical and time-consuming step is the annotation of the organism's proteome: assigning molecular functions, involvement in biological processes and the cellular localizations of all identified protein sequences. The process of protein annotation involves classifying each protein using the Gene Ontology (GO) and selecting suitable free text descriptions, which are required for submission into sequence databases. We refer to these as GO and DE predictions, respectively. Although servers are available that perform functional annotation, none perform both GO and DE predictions concurrently, they tend to be slow and are restricted in the number of queries that can be submitted. Finally, many of the existing methods do not return an estimate of the prediction accuracy.

2. METHODS AND RESULTS

The above shortcomings are remedied by PANNZER 2, an interactive web server and standalone program for protein function prediction. It is, to our knowledge, the only tool that performs both DE and GO prediction. It uses SANS-parallel [1], a high-throughput sequence search program that is thousands of times faster than BLAST, to identify homologous sequences. It is, therefore, capable of analyzing tens of thousands of proteins interactively. Results are displayed as an HTML table summarizing both DE and GO predictions, including a statistical estimate for the reliability of each prediction. Each results page additionally has links to the complete list of SANS hits, allowing users to understand how results were derived. PANNZER 2 provides four alternative scoring functions for GO prediction to highlight which predictions are more robust than others. The scoring functions include PANNZER [2], BLAST2GO [3] and ARGOT2 [4] like scores. PANNZER 2 is written in Python and designed to be a platform for the development novel prediction methods. For this reason, the source code is modular and is based on a spreadsheet data object where new prediction methods and data sources can be added as new columns to the table.

3. ENSEMBLE OF METHODS IN CAFA3 COMPETITION

Using PANNZER 2, our group participated in the recent third critical assessment of function annotation (CAFA3) competition. We combined multiple data sources to make predictions including: sequence similarity, biomedical literature, inter-ontology annotation correlations and protein-protein interaction data. The backbone of our approach was based on PANNZER 2 that uses sequence similarity to generate predictions, which it was able to

do for a majority of query sequences. Each other data source was only available for a subset of queries and, therefore, could only be used where available. For biomedical literature, article abstracts were converted into the bag-of-words feature representation and used to fit a hierarchical classifier based on a binary naïve Bayes classifier per GO term. Similarly, predictions based on inter-ontology annotation correlations were made by training an SVM using a semantic distance function (in our case, weighted Jaccard index) as a kernel function. We also used protein-protein interaction data collated by Uniprot to identify enriched GO classes in the set of interacting proteins. Annotations from each prediction method were calibrated using Platt scaling to ensure that the probabilistic estimates were comparable. Where more than one method predicted a GO term for a given protein, the scores were ensembled by taking the maximum. The idea being, where one data source might contain insufficient signal to differentiate proteins with a given annotation, other data sources might enable a less ambiguous prediction.

4. REFERENCES

1. Somervuo P, Holm L (2015) SANSparallel: interactive homology search against Uniprot. **Nucl. Acids Res.** 43, W24-W29

2. Koskinen P, Törönen P, Nokso-Koivisto J, Holm L (2015) PANNZER - High-throughput functional annotation of uncharacterized proteins in an error-prone environment. **Bioinformatics** 31 (10), 1544-1552

3. Conesa, A; Götz, S; García-Gómez, JM; Terol, J; Talón, M; Robles, M (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. Bioinformatics** 21 (18): 3674–6

4. Falda M, Toppo S, Pescarolo A, Lavezzo E, Di Camillo B, Facchinetti A, Cilia E, Velasco R and Fontana P *(2012)* Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms**. BMC Bioinformatics** 13(Suppl 4):S14