Large-scale Function Prediction Using Semantic Similarity of Weighted Gene Ontology Terms


(1) Department of Biological Chemistry, University of Padova, Padova, Italy
(2) Department of Histology, Microbiology, and Medical Biotechnologies University of Padova, Padova, Italy
(3) Department of Information Engineering, University of Padova, Padova, Italy
(4) Istituto Agrario San Michele all’Adige Research and Innovation Centre, Foundation Edmund Mach, Trento, Italy

Motivation
Thanks to the advent of the Next Generation sequencing technologies, we have assisted to an exponential increase in sequence data generation. The task to assign a curator-reviewed function to every single sequence is unworkable and new methods based on computational approaches are necessary. This is now eased by the advent of Gene Ontology (GO) and biological databases as UniProtKB-GOA which have revolutionized the way knowledge data are accessed. The GO arranges concepts in a controlled vocabulary and a hierarchical Directed Acyclic Graph, whereas UniProtKB-GOA is a databank of proteins described with GO terms. These projects greatly facilitate the mining of biological information by means of computer algorithms. In this scenario, we have developed Argot2, a web-based protein function predictor for large-scale sequencing projects. It is based on a weighting scheme and a semantic similarity grouping strategy to annotate proteins with Gene Ontology terms that, presently, derive from BLAST and HMMER searches vs. Uniprot and Pfam databases respectively. Argot2 reports a confidence score for each prediction and has been designed to scale up the annotation of entire genomes.

Methods
GOA databank and Pfam HMM models are downloaded and stored locally in a MySQL database. GO terms occurrences in GOA are pre-calculated and saved in a separate table. We extract the GO annotations of all proteins belonging to each Pfam entry to enrich GO assignments. For each GO term, we calculate its occurrence within the Pfam model and divide it by the total number of the proteins in the model. Logistic curve transformation is applied to reward highly abundant GO terms and to penalize those that are sparse and likely false positives. The GO terms associated to proteins and Pfam models retrieved by BLAST and HMMER, respectively, are weighted according to the e-value score of the hits. These lists of weighted GO terms are given in input to Argot2.
Argot2 takes into consideration the GO structure released by the GO consortium. All the paths to the root are reconstructed starting from the GO input list. The nodes that have not been visited are pruned and a final GO slim is obtained. The Information Content (IC) of each GO node in the slim is calculated according to Resnik formula. The farther to the root the higher the IC, as the GO term is more informative. Before proceeding with the computation, a first filter is applied to exclude isolated terms that can be eventually unrelated to the query. Assuming a Gaussian distribution of the node weights (W), the Z-score is used to trim poorly populated branches of the graph. The remaining nodes in the GO-slim are, then, grouped according to their semantic similarity, using the Lin formula. The GO terms that share a strong biological relationship form a unique informative group. A second filter, the Group Score (G-score), is further applied to prune isolated groups. For each of the remaining groups the most specific and high scoring annotation is reported. After the filtering phase, the algorithm assigns the TS score to the culled GO hits. TS takes into account the information content IC, the weight W and the group score G-score of the node, rewarding those hits that are particularly significative and specific. The GO terms with TS over a chosen threshold are extracted and reported.

Results
We have assessed Argot2 over 6187 annotated proteins from Yeast genome and we have compared it with Blast2GO tool. Preliminary results showed that our tool greatly outperforms Blast2go in terms of speed and annotation accuracy. In our testing conditions, reasonably good performance in both precision and recall has been reached. The F-measure, $2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall})$, calculated at the default parameters, are the following: i) Molecular Function, 0.89 and 0.64 ii) Biological Process, 0.81 and 0.55 for Argot2 and Blast2GO respectively. Argot2 has proven to effectively discriminate among false and true positives. This is critical when annotating very large genome data sets: reducing the false positives rate is definitely desirable, since it can prevent biased information to impact negatively on post-genomic studies. In future releases new sources of information will be added, trying to give an answer to non trivial cases that lie beyond sequence similarity based evidences.

Availability
http://www.medcomp.medicina.unipd.it/Argot

Contact email
marco.falda@unipd.it