

PFP/ESG: automated protein function prediction servers enhanced with Gene Ontology visualization tool

Ishita K. Khan^{1,†}, Qing Wei^{1,†}, Meghana Chitale¹ and Daisuke Kihara^{1,2,*}¹Department of Computer Science and ²Department of Biological Science, Purdue University, West Lafayette, IN 47907, USA

Associate Editor: John Hancock

ABSTRACT

Summary: Protein function prediction (PFP) is an automated function prediction method that predicts Gene Ontology (GO) annotations for a protein sequence using distantly related sequences and contextual associations of GO terms. Extended similarity group (ESG) is another GO prediction algorithm that makes predictions based on iterative sequence database searches. Here, we provide interactive web servers for the PFP and ESG algorithms that are equipped with an effective visualization of the GO predictions in a hierarchical topology.

Availability: PFP/ESG servers are freely available at <http://kiharalab.org/web/pfp.php> and <http://kiharalab.org/web/esg.php>, or access both at http://kiharalab.org/pfp_esg.php

Contact: dkihara@purdue.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 15, 2014; revised on September 6, 2014; accepted on September 25, 2014

1 INTRODUCTION

The rapid growth of genomics and proteomics technologies has posed a major challenge of large-scale automatic annotations of newly sequenced data that awaits functional interpretation. Advanced algorithms can often provide more accurate annotations with a larger coverage than conventional function prediction methods that use homology as the source of information. Here we developed web servers for our two sequence-based function prediction algorithms, protein function prediction (PFP) and extended similarity group (ESG).

PFP extends traditional PSI-BLAST (Altschul *et al.*, 1997) search by extracting and scoring Gene Ontology (GO) annotations from distantly similar sequences and by applying contextual associations of GO terms observed in the annotation database to the scoring scheme (Hawkins *et al.*, 2006, 2009). PFP was ranked the best in the function prediction category in the Critical Assessment of Techniques for Protein Structure Prediction (López *et al.*, 2007).

ESG performs an iterative sequence database search and assigns a probability score to GO terms based on its relative similarity scores to the multiple-level neighbors in a protein

similarity graph (Chitale *et al.*, 2009). ESG was shown to outperform conventional methods in a thorough benchmark study. In the large-scale community-based critical assessment of protein function annotation experiment, ESG was ranked fourth in predicting Molecular Function GO terms among 54 participating groups (Radivojac *et al.*, 2013). Thus, both PFP and ESG have been rigorously benchmarked both in the original papers and in objective assessments by the community. Predictive performance of the two methods is discussed in Supplementary Data.

2 METHODS

2.1 PFP algorithm

The PFP algorithm uses PSI-BLAST to obtain sequence hits for a target sequence and predicts GO terms. PFP computes the score for GO term f_a , $s(f_a)$, as follows:

$$s(f_a) = \sum_{i=1}^N \sum_{j=1}^{N_{func}(i)} ((-\log(E - value(i)) + b)P(f_a|f_j))$$

where N is the number of sequence hits considered in the PSI-BLAST hits, $N_{func}(i)$ is the number of GO annotations for the sequence hit i , $E_value(i)$ is the PSI-BLAST E-value for the sequence hit i , f_j is the j -th annotation of the sequence hit i , and constant b takes value 2 ($= \log_{10}100$). The conditional probabilities $P(f_a|f_j)$ consider co-occurrence of GO terms observed in sequence annotations. To take into account the hierarchical structure of the GO, PFP transfers the raw score to parental terms by computing the proportion of proteins annotated with f_a relative to all proteins that belong to the parental GO term in the database.

2.2 ESG algorithm

The ESG algorithm recursively performs PSI-BLAST searches using sequences that are retrieved by the initial search from the query sequence. Each sequence hit in a search is assigned a weight that is computed as the proportion of the $-\log(E_value)$ of the sequence relative to the sum of the weights from all the sequence hits of the same search round. This weight is assigned to GO terms annotating the sequence hit. Weights for GO terms found in the second iteration search are computed in the same fashion. Finally, the score for a GO term is computed as the sum of weights from the two levels of searches.

PFP and ESG have different characteristics: PFP is designed to have a larger coverage by retrieving annotations widely from even weakly similar sequences, while ESG is for better specificity by taking consistently predicted GO terms in an iterative search. This is further discussed in Supplementary Data.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

