ESG: Extended Similarity Group method for automated protein function prediction

Meghana Chitale[1], Troy Hawkins[2], Changsoon Park[4], Daisuke Kihara[1, 2, 3]*
1. Department of Computer Science, Purdue University, West Lafayette, IN, USA 47907,
2. Department of Biological Sciences, Purdue University, West Lafayette, IN USA 47907,
3. Markey Center for Structural Biology, Purdue University, West Lafayette, IN USA 47907,
4. Department of Statistics, Chung-Ang University, Seoul 156-756, Korea.
*To whom correspondence should be addressed: dkihara@purdue.edu

1. INTRODUCTION

We present here the Extended Similarity Group (ESG) method, which annotates query sequences with Gene Ontology (GO) terms by assigning probability to each annotation computed based on iterative PSI-BLAST searches. Conventionally sequence homology based function annotation methods, such as BLAST, retrieve function information from top hits with a significant score (E-values). In contrast, the PFP[1] method, which we have presented previously, goes one step ahead in utilizing a PSI-BLAST result by considering very weak hits even an E-value of up to 100 and also by incorporating the functional association between GO terms (FAM matrix[1]) computed using term co-occurrence frequencies in the UniProt database. PFP is very successful which is evidenced by the top rank in the function prediction category in CASP7 competition[2]. Our new approach, ESG method, further improves the accuracy of PFP by essentially employing PFP in an iterative fashion. An advantage of ESG is that it is built in a rigorous statistical framework: Unlike PFP[1] method that assigns a weighted score to each GO term, ESG assigns a probability based on weights computed using the E-value of each hit sequence on the path between the original query sequence and the current hit sequence.

2. METHOD

ESG performs iterative PSI-BLAST searches beginning from query sequence Q whose annotations are to be predicted using the probability score assigned to different GO terms. S_1, S_2, S_3…S_N be the PSI-BLAST hits for Q each with E-values E_1, E_2, E_3… E_N, respectively. At each PSI-BLAST search we consider fixed number (hit_count) of sequences that satisfy the E-value cutoff. Each sequence thus obtained has a weight associated with it which is given by Equation (1). Further, beginning from each sequence hit S_i obtained at level one, we perform an iterative PSI-BLAST search to get sequence hits at the second level referred as S_{ij}. The weight W_i computed for sequence S_i is distributed between S_i and all its children using a step weight parameter v as shown in Equation (3). We compute the weights for each of the second level sequences similar to the first level and multiply it by (1- step weight factor) to get the net score for each of the second level sequence S_{ij}. Using this weighting scheme we can associate the score with each sequence obtained during iterative PSI_BLAST search as shown in Figure 1 and transfer the score to each GO term which annotates that sequence. Thus we can compute the net probability of sequence Q getting annotation f_a as shown in Equation (2) by summing the weighted scores for each sequence that has f_a in its annotation list. The same concept can be easily scaled to work with multiple levels and to take different number of top PSI-BLAST searches at each level.

$$W_i = \frac{-\ln(E_i)}{\sum_{j=1}^{N} -\ln(E_j)} \quad (1)$$

$$P_Q(f_a) = \sum_{i=1}^{N} W_i \bullet P_{Si}(f_a) \quad (2)$$

$$P_{Si}(f_a) = v \bullet I_{si}(f_a) + (1-v) \bullet \sum_{j=1}^{n_i} W_{ij} \bullet I_{Sij}(f_a) \quad (3)$$

In Equation (1), W_i is weight for sequence S_i, E_i is E-value for sequence S_i, N is number of sequence hits for Q that are considered based on hit_count and E-value cutoff. In Equation (2), $P_Q(f_a)$ is the probability that sequence Q is annotated by GO term f_a, N is number of sequence hits for Q that are considered based on hit_count and E-value cutoff, and $P_{Si}(f_a)$ is the probability that sequence S_i is annotated by GO term f_a. In equation (3), v is the step weight parameter, W_{ij} is the weight computed for sequence S_{ij} which is second level PSI-BLAST hit from sequence S_i, $I_{si}(f_a)$ is a binary function which is one if sequence S_i has annotation f_a in database, n_i is number of sequence hits for S_i that are considered from second level PSI-BLAST search based on hit_count and E-value cutoff, $I_{Sij}(f_a)$ is a binary function as described before.

3. RESULTS AND FUTURE WORK

We have predicted annotations for a small subset of proteins by using the ESG method with a probability threshold of 0.4 for selecting the annotations. The predictions were compared with actual annotations of the proteins in the database using the semantic similarity funsim[3] score. This score ranges between 0 and 1.0 with 1.0 being the perfect prediction for actual annotations of a query protein. We have compared funsim scores obtained using top 5 predictions done by PFP and top PSI-BLAST hits in each of the three basic GO categories and those by ESG. Figure 2 shows that PFP and top PSI-BLAST give on an average 0.6 and 0.45 funsim similarity scores respectively as compared to ESG that gives an average funsim score of 0.8, indicating superior performance of ESG.

We plan to use the GO tree structure to add parental scoring scheme taking into account the is_a relations between GO terms and parents as well as incorporate knowledge about correlation between occurrences of annotation terms as captured by FAM matrix. These components when added to the probability scoring scheme are expected to boost the prediction accuracy by further improving the similarity between predictions and the actual annotations of query sequences.

4. FIGURES

Figure 1: Iterative blast hits and ESG probability score assignment.

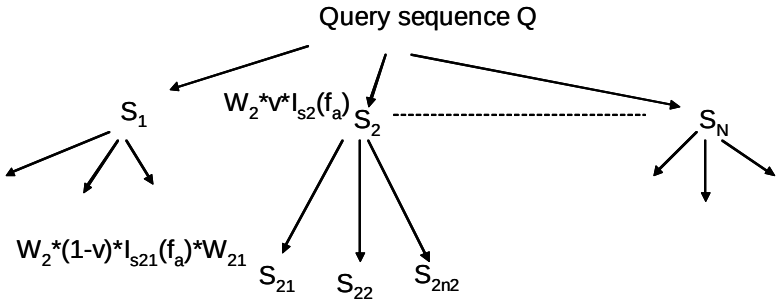
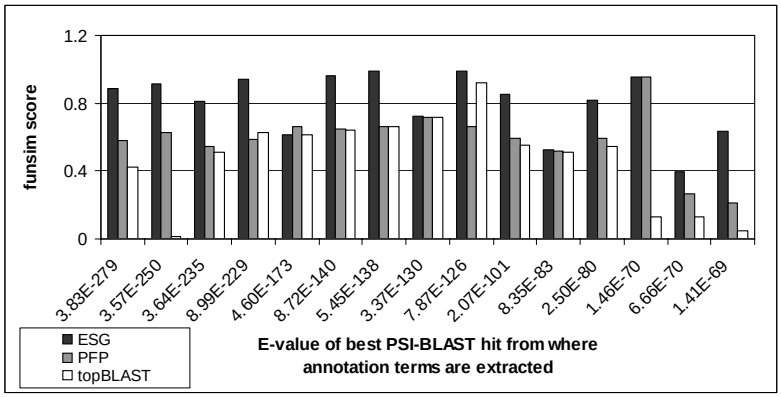


Figure 2: Benchmark results of ESG, PFP and top PSI-BLAST predictions.



5. REFERENCES

1. Hawkins T., Luban S. and Kihara D. 2006. Enhanced Automated Function Prediction Using Distantly Related Sequences and Contextual Association by PFP, Protein Sci, 15(6): 1550-1556.

2. Lopez G, Rojas A, Tress M, and Valencia A. 2007. Assessment of predictions submitted for the CASP7 function prediction category. Proteins. 2007;69 Suppl 8:165-74.

3. Schlicker A., Domingues F.S., Rahnenfuhrer J. and Lengauer T. 2006. A New Measure for Functional Similarity of Gene Products Based on Gene Ontology, BMC Bioinformatics, 7:302.