Chapter 1

Using PFP and ESG Protein Function Prediction Web Servers

Qing Wei, Joshua McGraw, Ishita Khan, and Daisuke Kihara

Abstract

Elucidating biological function of proteins is a fundamental problem in molecular biology and bioinformatics. Conventionally, protein function is annotated based on homology using sequence similarity search tools such as BLAST and FASTA. These methods perform well when obvious homologs exist for a query sequence; however, they will not provide any functional information otherwise. As a result, the functions of many genes in newly sequenced genomes are left unknown, which await functional interpretation. Here, we introduce two webservers for function prediction methods, which effectively use distantly related sequences to improve function annotation coverage and accuracy: Protein Function Prediction (PFP) and Extended Similarity Group (ESG). These two methods have been tested extensively in various benchmark studies and ranked among the top in community-based assessments for computational function annotation, including Critical Assessment of Function Annotation (CAFA) in 2010–2011 (CAFA1) and 2013–2014 (CAFA2). Both servers are equipped with user-friendly visualizations of predicted GO terms, which provide intuitive illustrations of relationships of predicted GO terms. In addition to PFP and ESG, we also introduce NaviGO, a server for the interactive analysis of GO annotations of proteins. All the servers are available at http://kiharalab.org/software.php.

Keywords Protein function prediction, Genome annotation, BLAST, Gene Ontology, Automated function prediction, Sequence analysis

1 Introduction

Functional interpretation of novel proteins is a central problem in molecular biology and bioinformatics. As genome sequencing and proteomic technologies advance at a striking pace, an overwhelming amount of sequence data awaits to be analyzed and assigned with functional interpretations. Since performing biological experiment for such purposes does not scale up in terms of time, effort and expense, automatic function prediction (AFP) methods have been pursued and have become one of the important problems in bioinformatics. There are many AFP algorithms developed in the past years in order to achieve accurate annotation and wider coverage to replace the conventional function prediction methods which use homology as the source of information [1, 2]. A review by

Daisuke Kihara (ed.), *Protein Function Prediction: Methods and Protocols*, Methods in Molecular Biology, vol. 1611,
DOI 10.1007/978-1-4939-7015-5_1,

Hawkins & Kihara summarizes several categories of AFP methods beyond traditional sequence similarity, which leverage sequence, structural, genomic, cellular and metabolic context-based information [3]. A review by Sael et al. [4] focuses on AFP methods for non-homologous proteins in the sequence and structure-based categories.

For the advancement of such computational techniques, it is very important that there are community-wide efforts for objective evaluation of prediction accuracy. Among several efforts carried out in the protein function prediction community in the past, a recent notable one is CAFA (Critical Assessment of Function Annotation) [5]. The first round of CAFA was held in 2010–2011 [5], and the second round, CAFA2, was held in 2013–2014 [6]. CAFA3 is planned in 2016–2017.

Here, we introduce two publicly available webservers for function prediction methods: Protein Function Prediction (PFP) [7, 8] and Extended Similarity Group (ESG) [9]. Both webservers take a list of query sequences and output a list of predicted Gene Ontology (GO) terms [10, 11]. The servers have been maintained over years and extensively benchmarked in the past [12, 13]. In both CAFA1 and CAFA2, PFP and ESG were ranked among the top function prediction methods. In the CAFA1 experiment, ESG was ranked fourth in the molecular function (MF) GO category among 54 participating groups [5], while PFP did well in all the three categories in CAFA 2 [6]. In an earlier community-based assessment, the function prediction category of (CASP) held in 2006, PFP was ranked the top [14].

PFP and ESG were designed to achieve complementary goals: PFP is for achieving a large prediction coverage by retrieving annotations widely including from weakly similar sequences. On the other hand, ESG is for improving specificity by accumulating contribution of consistently predicted GO terms in an iterative search. The interactive webserver of PFP and ESG [15] is developed to assist in the sequence-based function prediction and to enhance the understanding of predicted functions by an effective visualization of the predictions in a hierarchical GO topology. In addition, we also describe NaviGO, a newly developed web-based tool for interactive analysis of GO term annotations of proteins. All the servers are available at http://kiharalab.org/software.php.

2 Function Prediction Algorithms in PFP and ESG

In this section, we briefly explain the main idea of PFP and ESG algorithms. For more details, please refer to the original papers [7–9].

2.1 The PFP Algorithm

The PFP algorithm uses PSI-BLAST [1] to obtain sequence hits for a target sequence and computes the score for GO term f_a as follows:

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

where N is the number of sequence hits considered in the PSI-BLAST hits; Nfunc(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 jth annotation of the sequence hit i; and constant b takes value 2 ($= \log_{10} 125$) to keep the score positive when retrieved sequences up to an E-value of 125 are used. The conditional probabilities $P(f_a|f_j)$ are used to consider co-occurrence of GO terms in a single sequence annotation, which are computed as the ratio of the number of proteins co-annotated with GO terms f_a and f_j as compared with ones annotated only with the term f_j. To take into account the hierarchical structure of GO, PFP transfers the raw score to the 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. The score of a GO term computed as the sum of the directly computed score by Eq. 1 and the ones from the parental propagation is called the raw score.

Compared to the conventional usage of PSI-BLAST that uses a strict E-value cutoff, e.g., 0.001, for transferring function annotations, the characteristic of PFP is that it collects GO annotations even from very weakly sequences up to an E-value of 125. Individual weakly similar sequences do not contribute much to a raw score, but a GO term can accumulate a substantially large score and be predicted with confidence if the GO term appears in many sequences.

2.2 The ESG Algorithm

ESG recursively performs PSI-BLAST searches from sequence hits obtained in the initial search from the query sequence Q, which will retrieve N sequence hits (N is "the number of hits per stage" parameter in the ESG input page as shown in the next section), S_1, S_2,...S_N, each with E-value E_1, E_2,...E_N, respectively. Each sequence hit in a search is assigned a weight W_i that is computed as the proportion of the $-\log(E\text{-value})$ of the sequence relative to the sum of the $-\log(E\text{-value})$ from all the sequence hits considered in the search of the same level:

$$W_i = \frac{-\log(E_i) + b}{\sum_{j=1}^{N} \{-\log(E_j) + b\}} \qquad (2)$$

where score $-\log(E_i)$ is shifted by a constant value b, which makes the score a nonnegative value. This weight is assigned for GO terms annotating the sequence hit and the probability of the GO term f_a annotating the query sequence Q is defined as the sum of weights of f_a that come from sequences annotated with f_a:

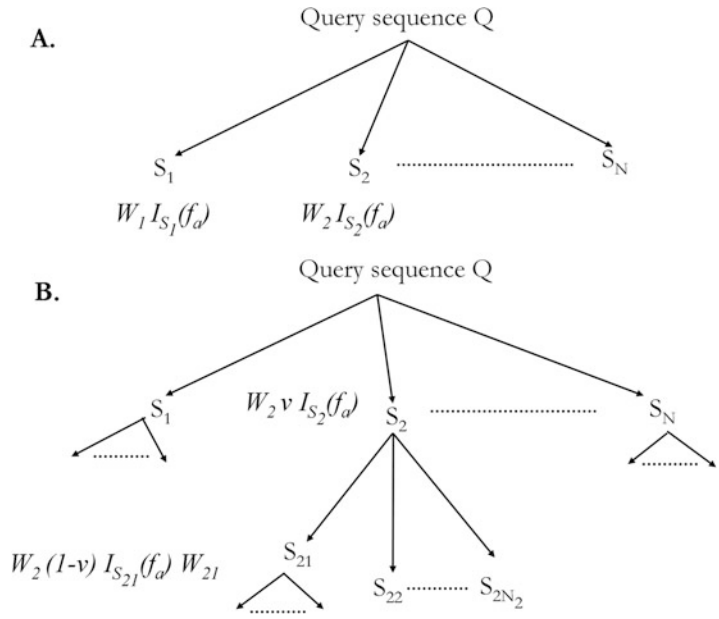

Fig. 1 Computing the ESG score. (**a**) For a single-layer search, a score of a function f_a is computed as a sum of the weight of sequences that have f_a in their GO annotation. (**b**) When a two-layer search is performed, a score comes from a weighted combination of the second level search and the first level search. This figure is adopted from the original paper of ESG (Chitale, Hawkins, Park, & Kihara, Bioinformatics, 25: 1739–1745, 2009) with permission from the publisher

$$P_{\mathrm{Q}}^{\mathrm{d}}(f_{\mathrm{a}}) = \sum_{i=1}^{N} W_i \cdot I_{S_i}(f_{\mathrm{a}}) \tag{3}$$

The function I indicates whether the given sequence S_i has annotation f_a:

$$I_{S_i}(f_{\mathrm{a}}) = \begin{cases} 1 \text{ if } S_i \text{ has } f_{\mathrm{a}} \text{ annotation} \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

The index d on the left side of Eq. 3 indicates that function information comes from direct annotations to sequences. Additionally, multilevel exploration ("the number of stages" parameter in the ESG input page) of the sequence-similarity space (PSI-BLAST) shown in Fig. 1 is performed around the target protein by sharing the weights between levels using a weight parameter v. In the second round, each of the sequences S_1, S_2,. . .S_N retrieved in the first round is in turn used as a query. Suppose sequence S_i obtains N_i sequences by a PSI-BLAST run, each referred to as S_{ij}. The weights for S_{ij}, W_{ij} can be computed in a similar manner to Eq. 2. Combining the two levels of searches:

$$P_{\mathrm{Q}}^{\mathrm{d}}(f_{\mathrm{a}}) = \sum_{i=1}^{N} W_i \cdot P_{S_i}^{\mathrm{d}}(f_{\mathrm{a}}) \tag{5}$$

$$P_{\mathrm{S}}^{\mathrm{d}}(f_{\mathrm{a}}) = v \cdot I_{S_i}(f_{\mathrm{a}}) + (1 - v) \cdot \sum_{j=1}^{N_i} \mathrm{W}_{ij} \cdot I_{S_{ij}}(f_{\mathrm{a}}) \tag{6}$$

Equation 5 is a variation of Eq. 3, representing that the score of a GO term f_{a} for the query Q is contributed by sequences retrieved at the first level (S_1 to S_N). The weights for GO terms found in the second level search are computed similarly, where Eq. 2 defines the weight W_i. Eq. 6 defines the score for f_{a} for sequence S_i as a combination of $I_{S_i}(f_{\mathrm{a}})$, which is sequence S_i's annotation, and the second level search. The first and the second terms are weighted by a factor v. Moreover, the equations can be recursively extended to multiple levels of searches to explore broader space around the query sequence. The score for each GO term ranges from 0.0 to 1.0.

ESG predicts a GO term with a high score if it appears many times consistently in the multiple searches including the initial search and the second level searches. In general, the number of GO terms predicted by ESG is smaller (5–10 GO terms) than PFP (often over 50 terms), and terms predicted by high scores by ESG are usually highly accurate.

3 Input and Output of the Servers

3.1 Query Input Page of PFP and ESG

In Subheading 3, we explain how to use the webservers with an example. PFP is available at http://kiharalab.org/pfp.php and ESG is at http://kiharalab.org/esg.php. Query sequences can be submitted to both PFP and/or ESG from the combined submission page at http://kiharalab.org/web/pfp_esg.php. Please also refer to a detailed instruction at http://kiharalab.org/web/pfp_tutorial.php and http://kiharalab.org/web/esg_tutorial.php for PFP and ESG, respectively. Both the servers may be used without making an account; however, users are encouraged to create their account on the servers. With an account, users may automatically keep and refer to prediction results that have been processed earlier.

PFP and ESG accept query inputs of FASTA formatted protein sequences. Users may submit sequences separated by line breaks in the text box titled "Enter Query Sequence(s)" or upload a FASTA file containing multiple sequences (Fig. 2). To view a sample of the format, users may click on "Load Sample" to fill the field with an example sequence. Selecting "Clear" will remove all inputs sequences including uploaded files. Currently, up to 100 sequences

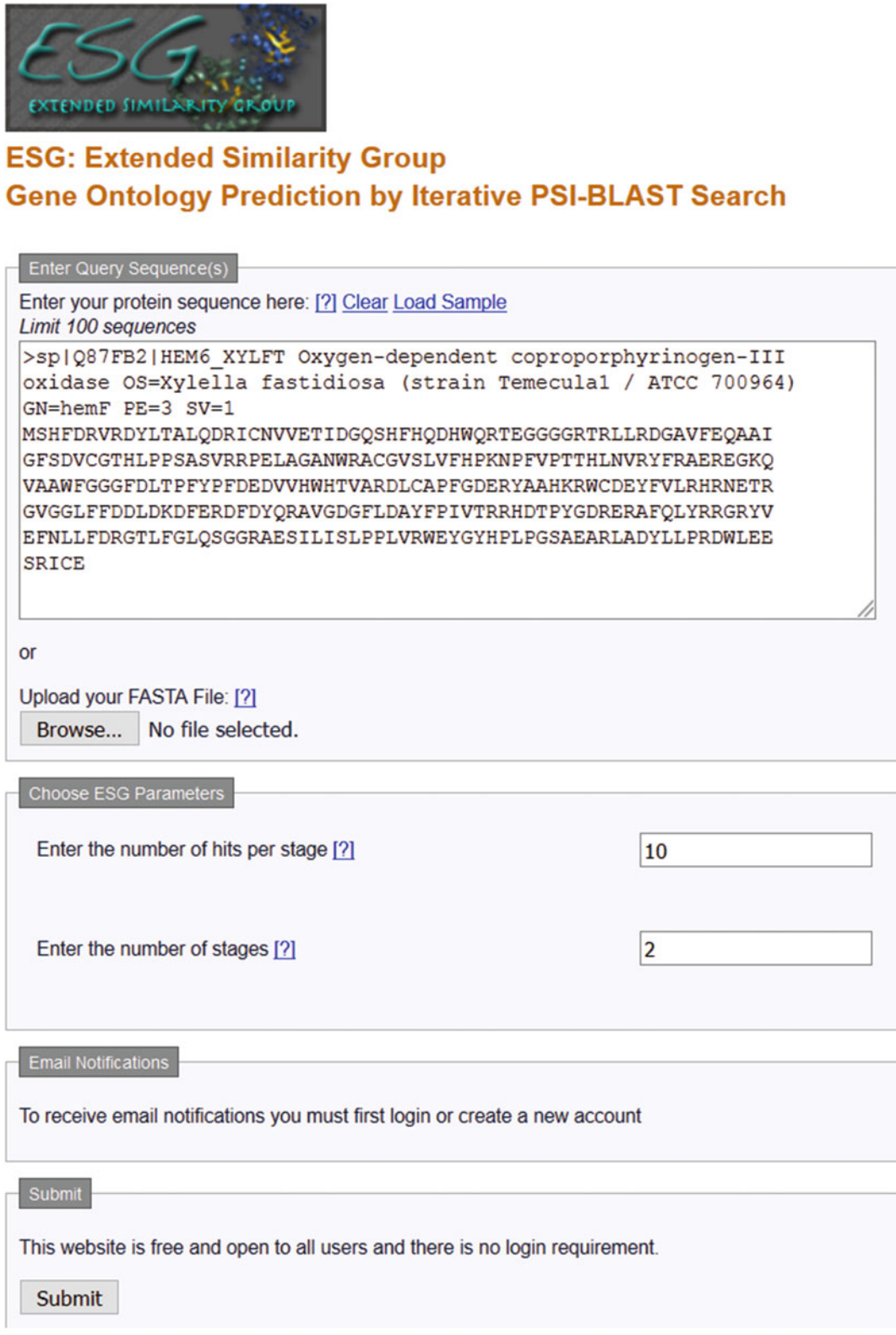

Fig. 2 Query input page of ESG. Query sequences can be pasted in the submission window or a sequence file can be uploaded. The query page of PFP is essentially the same, except that it does not have the number of hits and the number of stages parameters

may be submitted at a time to avoid overloading the computer server by the job queue.

For ESG, there are two more parameters that must be entered: "Number of hits" and "Number of stages." "Number of hits" indicates the number of PSI-BLAST hits to be considered at each level of ESG. The default value of this parameter is set to 10 in our web server. "Number of stages" indicates the level of searches to be performed by ESG. The default value for this parameter is chosen as 2. We recommend not changing the "Number of stages" parameter to a larger value as the computational time will suffer exponentially and we did not observe an improvement during benchmark in the original paper [9]. As for the "Number of hits" parameter, it can be increased if a prediction result by the default value is not satisfactory. For example, we used 50 for this value since it performed well during the benchmark [9]. However, if the parameter value is increased from 10 to 50, it requires roughly five times more computational time (with the two-stage setting).

3.2 Output Page with Case Studies

After selecting the submit button at the bottom section of the page, users will be directed to the job page displaying the status of that job. The job will be queued and assigned CPU time when available. You may refresh the page manually to check the status. Average computational time for PFP and ESG is 40.1 s and 7.5 min [15], respectively. When the job is completed, clicking on the job ID will display the predicted GO terms for the query sequences. Below we explain in detail how the results are presented.

3.2.1 PFP Output Page

The PFP results page shows the input sequences at the top section followed by the predicted terms for each GO category (Molecular Function (MF), Biological Process (BP), and Cellular Component (CC)), which have confidence greater than 5% of score of the top hit (Fig. 3). The results page also provides a link to the results in the XML format, which users may download for further processing. Selecting "Visualization of Predicted GO terms" will allow users to view the predicted terms in an interactive GO hierarchy. This tool allows users to pan and zoom through sub-nodes of related branches and is color mapped based on their assigned probability. Alternatively, users may select to color the nodes based on the number of child nodes under predicted terms. There are three different layouts users may choose (tree, radial, and circle) for visualizing the GO hierarchy as well as configurable layouts and interactive nodes in the Cytoscape [16] (Fig. 4).

Three links are provided below the visualization redirect links, which allow users to download static images of the GO hierarchy visualization. Selecting to download the image will render the SVG image and generate a figure. At the top of each static image is also a link to download the PNG image file. Users may also save the SVG image by bookmarking the static page for future reference.

List of Predicted GO Terms (the raw score can range 0 to over 20K)

Very high confidence : > 20K
High confidence : > 10K
Moderate confidence : > 500
Low confidence: >= 100 (But worthwhile to examine)
Below low confidence: < 100

Molecular Function Terms

PFP Score	Term	Description
36238.68	GO:0004109 [+]	coproporphyrinogen oxidase activity
36184.18	GO:0042803 [+]	protein homodimerization activity
30816.91	GO:0042802	identical protein binding
12008.03	GO:0046983	protein dimerization activity
10454.21	GO:0016634	oxidoreductase activity, acting on the CH-CH group of donors, oxygen as acceptor

5 Predictions
3 Predictions > 20K; **2** Predictions > 10K; **0** Predictions > 500; **0** Predictions >= 100

Biological Process Terms

PFP Score	Term	Description
68236.29	GO:0006779 [+]	porphyrin-containing compound biosynthetic process
67396.63	GO:0033014	tetrapyrrole biosynthetic process
66437.40	GO:0006778	porphyrin-containing compound metabolic process
65630.69	GO:0033013	tetrapyrrole metabolic process
46970.40	GO:0042168	heme metabolic process
36243.03	GO:0055114 [+]	oxidation-reduction process
36015.28	GO:0006782 [+]	protoporphyrinogen IX biosynthetic process
35920.00	GO:0046501	protoporphyrinogen IX metabolic process
29808.98	GO:0046148	pigment biosynthetic process
29014.25	GO:0042440	pigment metabolic process
24438.22	GO:0006783 [+]	heme biosynthetic process
15670.98	GO:0051188	cofactor biosynthetic process
11611.37	GO:0051186	cofactor metabolic process
7816.74	GO:0008152 [+]	metabolic process
7519.68	GO:0018130	heterocycle biosynthetic process

15 Predictions
11 Predictions > 20K; **2** Predictions > 10K; **2** Predictions > 500; **0** Predictions >= 100

Cellular Component Terms

PFP Score	Term	Description
34761.10	GO:0005737 [+]	cytoplasm
31279.85	GO:0044424	intracellular part
31040.66	GO:0005622 [+]	intracellular
25705.33	GO:0044464	cell part
25704.85	GO:0005623	cell

5 Predictions
5 Predictions > 20K; **0** Predictions > 10K; **0** Predictions > 500; **0** Predictions >= 100

Fig. 3 An example of predicted GO terms by PFP is shown in the PFP output page. The query used is hemF, oxygen-dependent coproporphyrinogen-III oxidase (UniProt ID: Q87FB2). Each category of GO terms is separated by Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). Prediction confidence is annotated by the color of the PFP Score, whereas *red* is very high confidence (>20 K) and *blue* is low confidence (100–500)

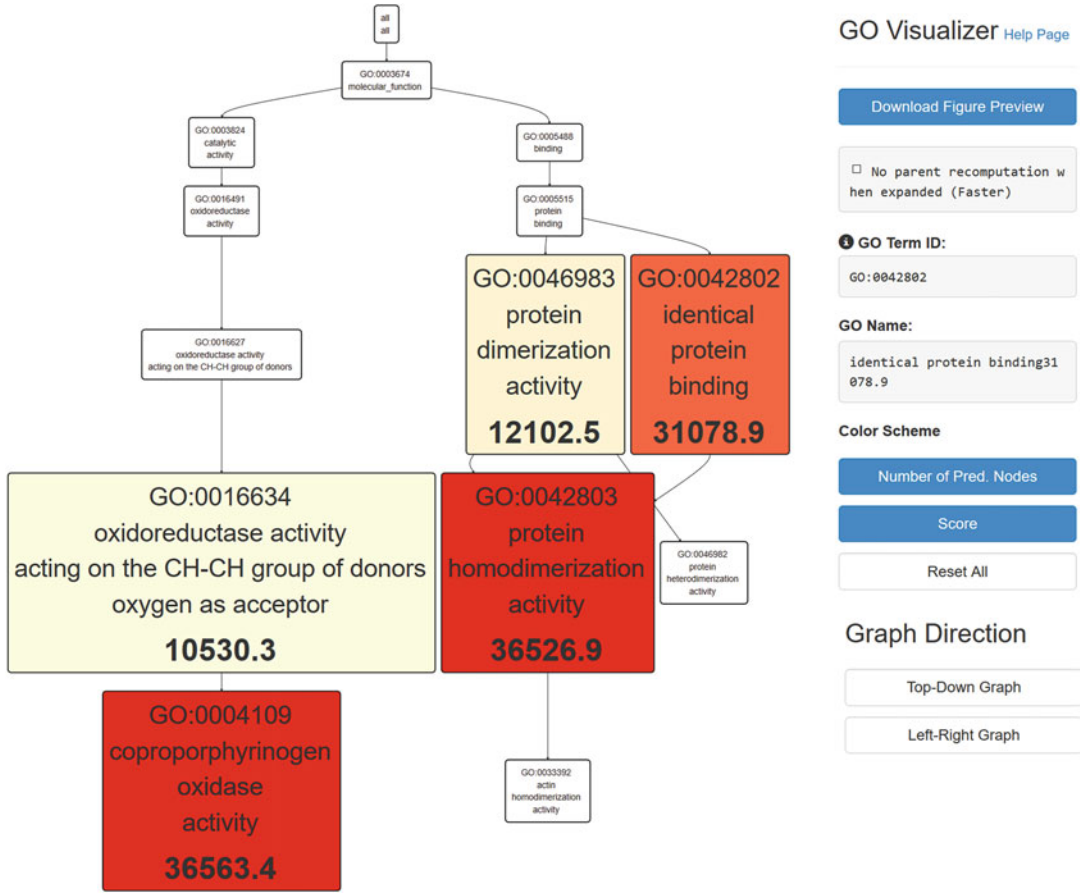

Fig. 4 Cytoscape output demonstrating a hierarchical Tree Layout of the PFP prediction. Each node represents a predicted GO term. *Red shades* in this figure indicate the prediction confidence

At the bottom section of the output page, the predicted results are categorized by MF, BP, and CC GO terms including the confidence, term ID, and term description. GO terms are colored as red, orange, green, and black, whereas red indicates high confidence of prediction (>70%) and black represents a low confidence (<30%). PFP allows users to trace the origin of the predicted GO terms through a dropdown list. Since the PFP algorithm often retrieves GO annotations from distantly related sequences that may not be obvious homologs, this tool provides useful insights as to how predictions are computed and the function of the query sequence. For each predicted GO term, clicking the [+] sign will open a dropdown list of sequence IDs which contributed toward the prediction. The contribution of each sequence is shown as the percentage of the score that originates from similar sequences (Fig. 5).

Molecular Function Terms

PFP Score	Term	Description
36238.68	GO:0004109 [-]	coproporphyrinogen oxidase activity
	B7MHT7 0.5%	
	B7MY86 0.5%	
	C4ZX09 0.5%	
	A8A2T4 0.5%	
	Q0TF33 0.5%	

Fig. 5 Example of the PFP GO term dropdown box displaying several links to other UniProt proteins that conferred the prediction, as well as the percent of their contribution. This list is shown for a GO term, GO:0004109, predicted for a query protein, Q87FB2

As an example, here we discuss prediction by PFP for oxygen-dependent coproporphyrinogen-III oxidase (UniProt ID: Q87FB2) (Fig. 3). This protein is involved in the first step of the protoporphyrinogen-IX from coproporphyrinogen-III synthesis pathway during heme biosynthesis. According to the EMBL-EBI database, this protein contains four MF, four BP, and one CC GO terms. PFP correctly predicts two of the four MF terms with medium to high confidence: GO:0004109 (coproporphyrinogen oxidase activity) and GO:0042803 (protein homodimerization activity). By expanding the dropdown list of GO:0004109 (coproporphyrinogen oxidase activity), we can trace the proteins that confer this prediction (Fig. 5). Proteins include hemF of *Escherichia coli O6:K15:H31* (UniProt ID: Q0TF33) (the protein in the bottom of Fig. 5) in the list serve to catalyze the aerobic oxidative decarboxylation of propionate groups of rings A and B of coproporphyrinogen-III to yield the vinyl groups in protoporphyrinogen-IX, and thus have the annotation of GO:0004109.

All four BP terms are predicted by PFP with very high confidence, which are GO:0006779 (porphyrin-containing compound biosynthetic process), GO:0006782 (protoporphyrinogen IX biosynthetic process), GO:0006783 (heme biosynthetic process), and GO:0055114 (oxidation-reduction process). Expanding the dropdown of GO:0006779 (porphyrin-containing compound biosynthetic process) reveals other hemF proteins such as (UniProtID: B7M6U5) of *Escherichia coli O8 (strain IA11)* which support this prediction. PFP also correctly predicts the only CC term, GO:0005737 (cytoplasm), with very high confidence (Fig. 3).

3.2.2 ESG Output Page

Both ESG and PFP display identically formatted results page. To understand ESG's output page, refer to Subheading 3.2.1. PFP Output page.

3.3 GO Term Analysis Using NaviGO

In the last section, we introduce NaviGO, a recently developed web-based tool for Gene Ontology visualization and similarity quantification, which is useful for understanding the relationships between predicted GO terms. It is accessible at http://kiharalab.org/web/navigo.

To enable a quantitative analysis of GO terms and gene functions from various aspects, on NaviGO, users can compute similarity of GO terms using six different scoring schemes that incorporate a variety of information ranging from GO topological structure, contextual association, and GO annotation frequency. There are four major functionalities, which are accessible through tabs on the top bar of the web site, i.e., GO Parents, GO Set, GO Enrichment, and Protein Set.

In the GO Parents page, users are able to retrieve parental GO terms in the GO hierarchy (Directed Acyclic Graph, DAG) for a list of query GO terms. It uses a lite version of GO Visualizer [15] to help users understand relationships of GO terms topologically in the GO DAG. Results are rendered in an interactive DAG where query GO terms are circled with bold black outlines. Additionally, parental GO terms will be listed in the text area below the visualization.

In the GO Set page, the tool will compute pairwise GO similarity scores for a list of input GO terms and output them as three formats (Fig. 6): a table, a network graph, and a bubble chart.

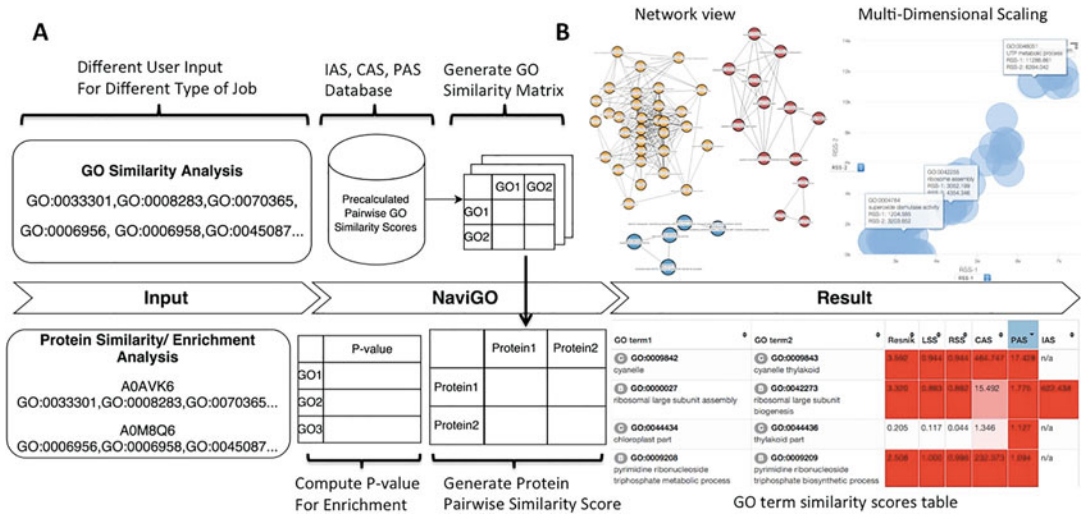

Fig. 6 (**a**) Workflow for NaviGO. Two types of input data are accepted, a set of GO terms or a set of genes with GO annotations. Similarity of GO terms is computed with six different GO scores including IAS, CAS, and PAS. If input data is a list of genes, then pairwise similarity scores for each pair of genes are computed. If GO enrichment analysis is selected, statistical significance of enrichment of GO terms is computed. (**b**), Presentation of results in NaviGO. Results are provided by a network view where similar GO terms or genes are connected; and in a bubble chart where similarity of GO terms is shown in a 2D plot of multi-dimensional scaling, or in a tabulated fashion, where significance of score similarity is indicated by a color scale

In the result table, Resnik's Similarity, Lin's Similarity, Relevance Similarity [17], Co-occurrence Association Score, PubMed Association Score [18], and Interaction Association Score [19] of pairs of input GO terms are colored based on score cutoffs. Table columns are sortable by clicking on score names at top row of the table. Common parents between a pair of GO terms are shown in the last column as well as a link to the interactive visualization, which illustrates parental GO terms in the GO DAG. In the network graph format, we showed an interactive network that summarizes the GO similarity as clusters where nodes are GO terms and edges indicate similarity score above a user-defined cutoff. The bubble chart format uses multidimensional scaling [20] to map the similarity into 2D coordinates and the user is able to choose the scoring schemes for either X or Υ coordinates.

In the GO Enrichment tab, NaviGO will take the NCBI taxonomy ID of the organism and a list of annotated genes in the organism and output the enrichment p-value for each unique GO term in the input annotation. Enriched GO terms are color mapped in GO visualizer. User can also adjust the number of enriched GO terms to visualize.

In the Protein Set tab, users can input a list of annotated proteins and NaviGO will calculate the functional similarity between each pair of input proteins using Funsim score [8, 17] with different similarity schemes similar as in the GO Set tab. The confidence of similarity predictions is classified into five levels: very high, high, moderate, low, and the rest. It indicates the score is within top 1%, 5%, 10%, and 20% relative to the score distribution of all protein pairs of an arbitrary organism specified by the user. The upper section in the result page shows an interactive clustering view based on protein similarity score (Fig. 6). A user-defined cutoff value controls the connectivity of edges between nodes, and scoring schemes can be switched using the bar on the top right-hand corner of the network panel. The computed analysis results can also be download as a table in the CSV format.

Acknowledgments

This work was supported partly by the National Institutes of Health (R01GM097528), the National Science Foundation (IIS1319551, DBI1262189, IOS1127027).

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
2. Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63–98

3. Hawkins T, Kihara D (2007) Function prediction of uncharacterized proteins. J Bioinforma Comput Biol 5(1):1–30
4. Sael L, Chitale M, Kihara D (2012) Structure- and sequence-based function prediction for non-homologous proteins. J Struct Funct Genom 13(2):111–123. doi:10.1007/s10969-012-9126-6
5. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Toronen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DWA, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK, Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kaszner R, Seemayer S, Vicedo E, Schaefer C, Achten D, Auer F, Boehm A, Braun T, Hecht M, Heron M, Honigschmid P, Hopf TA, Kaufmann S, Kiening M, Krompass D, Landerer C, Mahlich Y, Roos M, Bjorne J, Salakoski T, Wong A, Shatkay H, Gatzmann F, Sommer I, Wass MN, Sternberg MJE, Skunca N, Supek F, Bosnjak M, Panov P, Dzeroski S, Smuc T, Kourmpetis YAI, van Dijk ADJ, Braak CJF, Zhou Y, Gong Q, Dong X, Tian W, Falda M, Fontana P, Lavezzo E, Di Camillo B, Toppo S, Lan L, Djuric N, Guo Y, Vucetic S, Bairoch A, Linial M, Babbitt PC, Brenner SE, Orengo C, Rost B, Mooney SD, Friedberg I (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10(3):221–227. http://www.nature.com/nmeth/journal/v10/n3/abs/nmeth.2340.html supplementary-information
6. Jiang Y, Ronnen Oron T, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, Koo E, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraeian SM, Martelli PL, Profiti G, Casadio R, Cao R, Zhong Z, Cheng J, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, Salakoski T, Ginter F, Fang H, Smithers B, Oates M, Gough J, Törönen P, Koskinen P, Holm L, Chen C-T, Hsu W-L, Bryson K, Cozzetto D, Minneci F, Jones DT, Chapman S, Dukka BKC, Khan IK, Kihara D, Ofer D, Rappoport N, Stern A, Cibrian-Uhalte E, Denny P, Foulger RE, Hieta R, Legge D, Lovering RC, Magrane M, Melidoni AN, Mutowo-Meullenet P, Pichler K, Shypitsyna A, Li B, Zakeri P, ElShal S, Tranchevent L-C, Das S, Dawson NL, Lee D, Lees JG, Sillitoe I, Bhat P, Nepusz T, Romero AE, Sasidharan R, Yang H, Paccanaro A, Gillis J, Sedeño-Cortés AE, Pavlidis P, Feng S, Cejuela JM, Goldberg T, Hamp T, Richter L, Salamov A, Gabaldon T, Marcet-Houben M, Supek F, Gong Q, Ning W, Zhou Y, Tian W, Falda M, Fontana P, Lavezzo E, Toppo S, Ferrari C, Giollo M, Piovesan D, Tosatto S, del Pozo A, Fernández JM, Maietta P, Valencia A, Tress ML, Benso A, Di Carlo S, Politano G, Savino A, Rehman HU, Re M, Mesiti M, Valentini G, Bargsten JW, van Dijk AD, Gemovic B, Glisic S, Perovic V, Veljkovic V, Veljkovic N, Almeida-e-Silva DC, Vencio RZ, Sharan M, Vogel J, Kansakar L, Zhang S, Vucetic S, Wang Z, Sternberg MJ, Wass MN, Huntley RP, Martin MJ, O'Donovan C, Robinson PN, Moreau Y, Tramontano A, Babbitt PC, Brenner SE, Linial M, Orengo CA, Rost B, Greene CS, Mooney SD, Friedberg I, Radivojac P (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol 17(1):184. doi:10.1186/s13059-016-1037-6
7. Hawkins T, Luban S, Kihara D (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 15(6):1550–1556. doi:10.1110/ps.062153506
8. Hawkins T, Chitale M, Luban S, Kihara D (2009) PFP: automated prediction of Gene Ontology functional annotations with confidence scores using protein sequence data. Proteins 74(3):566–582. doi:10.1002/prot.22172
9. Chitale M, Hawkins T, Park C, Kihara D (2009) ESG: extended similarity group method for automated protein function prediction. Bioinformatics 25(14):1739–1745. doi:10.1093/bioinformatics/btp309
10. Seok YJ, Sondej M, Badawi P, Lewis MS, Briggs MC, Jaffe H, Peterkofsky A (1997) High affinity binding and allosteric regulation of *Escherichia coli* glycogen phosphorylase by the histidine phosphocarrier protein, HPr. J Biol Chem 272(42):26511–26521
11. D'Ari L, Rabinowitz JC (1991) Purification, characterization, cloning, and amino acid sequence of the bifunctional enzyme 5,10-methylenetetrahydrofolate dehydrogenase/5,10-methenyltetrahydrofolate cyclohydrolase from *Escherichia coli*. J Biol Chem 266(35):23953–23958
12. Khan IK, Wei Q, Chapman S, Kc DB, Kihara D (2015) The PFP and ESG protein function prediction methods in 2014: effect of database updates and ensemble approaches. GigaScience 4:43. doi:10.1186/s13742-015-0083-4
13. Chitale M, Khan IK, Kihara D (2013) In-depth performance evaluation of PFP and ESG

sequence-based function prediction methods in CAFA 2011 experiment. BMC Bioinform 14(Suppl 3):S2. doi:10.1186/1471-2105-14-S3-S2

14. Lopez G, Rojas A, Tress M, Valencia A (2007) Assessment of predictions submitted for the CASP7 function prediction category. Proteins 69(Suppl 8):165–174. doi:10.1002/prot.21651
15. Khan IK, Wei Q, Chitale M, Kihara D (2015) PFP/ESG: automated protein function prediction servers enhanced with Gene Ontology visualization tool. Bioinformatics 31 (2):271–272. doi:10.1093/bioinformatics/btu646
16. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13 (11):2498–2504. doi:10.1101/gr.1239303
17. Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T (2006) A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinform 7:302. doi:10.1186/1471-2105-7-302
18. Chitale M, Palakodety S, Kihara D (2011) Quantification of protein group coherence and pathway assignment using functional association. BMC Bioinform 12:373–373. doi:10.1186/1471-2105-12-373
19. Yerneni S, Khan I, Wei Q, Kihara D (2015) IAS: interaction specific GO term associations for predicting protein–protein interaction networks. IEEE/ACM Trans Comput Biol Bioinform. doi:10.1109/TCBB.2015.2476809
20. Sánchez J, Mardia KV, Kent JT, Bibby JM (1982) Multivariate analysis. Academic Press, London-New York-Toronto-Sydney-San Francisco 1979. xv, 518 pp., $ 61.00. Biom J 24 (5):502–502. doi:10.1002/bimj.4710240520