PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · FEBRUARY 2009

Impact Factor: 2.63 · DOI: 10.1002/prot.22172 · Source: PubMed

4 AUTHORS, INCLUDING:

Troy B Hawkins
Indiana University-Purdue University Indiana…
28 PUBLICATIONS **444** CITATIONS
SEE PROFILE

Daisuke Kihara
Purdue University
130 PUBLICATIONS **2,703** CITATIONS
SEE PROFILE

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data

Troy Hawkins,[1] Meghana Chitale,[2] Stanislav Luban,[3] and Daisuke Kihara[1,2,4]*

[1] Department of Biological Sciences, College of Science, Purdue University, West Lafayette, Indiana 47907
[2] Department of Computer Science, College of Science, Purdue University, West Lafayette, Indiana 47907
[3] Interdisciplinary Bioinformatics Program, University of California San Diego, La Jolla, California 92093
[4] Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, Indiana 47907

ABSTRACT

Protein function prediction is a central problem in bioinformatics, increasing in importance recently due to the rapid accumulation of biological data awaiting interpretation. Sequence data represents the bulk of this new stock and is the obvious target for consideration as input, as newly sequenced organisms often lack any other type of biological characterization. We have previously introduced PFP (Protein Function Prediction) as our sequence-based predictor of Gene Ontology (GO) functional terms. PFP interprets the results of a PSI-BLAST search by extracting and scoring individual functional attributes, searching a wide range of *E*-value sequence matches, and utilizing conventional data mining techniques to fill in missing information. We have shown it to be effective in predicting both specific and low-resolution functional attributes when sufficient data is unavailable. Here we describe (1) significant improvements to the PFP infrastructure, including the addition of prediction significance and confidence scores, (2) a thorough benchmark of performance and comparisons to other related prediction methods, and (3) applications of PFP predictions to genome-scale data. We applied PFP predictions to uncharacterized protein sequences from 15 organisms. Among these sequences, 60–90% could be annotated with a GO molecular function term at high confidence (≥80%). We also applied our predictions to the protein–protein interaction network of the Malaria plasmodium (*Plasmodium falciparum*). High confidence GO biological process predictions (≥90%) from PFP increased the number of fully enriched interactions in this dataset from 23% of interactions to 94%. Our benchmark comparison shows significant performance improvement of PFP relative to GOtcha, InterProScan, and PSI-BLAST predictions. This is consistent with the performance of PFP as the overall best predictor in both the AFP-SIG '05 and CASP7 function (FN) assessments. PFP is available as a web service at http://dragon.bio.purdue.edu/pfp/.

Proteins 2009; 74:566–582.

Key words: protein function prediction; protein-protein interaction network; gene ontology; confidence scores; systems biology; low resolution function.

INTRODUCTION

Advances in proteomics technologies have made possible the collection of large datasets characterizing protein–protein interactions (PPIs) and gene expression on a whole organism scale, and with over 660 complete genome sequences published and 3000 more currently in progress,[1,2] biological sequence data are being produced at a far greater rate than they are experimentally characterized. Thorough, systems-level interpretation of this growing body of proteomics datasets and new genomes relies on the availability of functional annotation of the included proteins. There has been a rush by the computational biology community to produce automated methods for protein function prediction that reflect the paradigm shift from analysis of single sequences to the kinds of large scale analysis and experimentation that are forming the backbone of the genome era of biology.[3–7]

Traditionally, the default method for characterizing a new sequence is essentially to transfer function of existing sequences, which are retrieved by a homology search method, for example, BLAST, with a significant score which exceeds a predefined threshold value (e.g. an *E*-value of 0.01). When universally applied to the large datasets described here, drawbacks in this type of sequence-based function annotation tend to arise from two sources. First, sequence similarity detection algorithms such as BLAST[8] and FASTA/P[9,10] provide function annotation typically only to half of genes in a genome since homologous sequences are not found at accepted significance thresholds.[11–14] Second, auto-

Grant sponsor: National Institute of General Medical Sciences of the National Institutes of Health; Grant numbers: R01GM075004; Grant number: U24 GM077905; Grant sponsor: National Science Foundation; Grant number: DMS 0604776.
*Correspondence to: D. Kihara, Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN 47907. E-mail: dkihara@purdue.edu.
Received 6 March 2008; Revised 3 June 2008; Accepted 5 June 2008
Published online 24 July 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22172

mated methods of annotation transfer between similar sequences can contribute to error propagation in current databases.[15,16] To approach the new problem of functional characterization of this growing body of uncharacterized sequence data, these limitations of traditional methods of protein function prediction must be readdressed.

Hennig et al.[17,18] first applied Gene Ontology (GO) annotations[19] to the results of a BLAST search with their GOblet method, which simply maps sequence matches onto a representation of the GO Directed Acyclic Graph (DAG). Although similarity scores are not combined, users can visualize how multiple sequence matches are associated to common GO terms. Zehetner extended this mapping in the OntoBlast method by weighting the list of retrieved GO terms.[20] Here the weight of any particular term is determined by multiplying the BLAST *E*-values from each sequence hit associated with the term. Khan et al.[21] in the GOFigure method and Martin et al.[22] in the GOtcha method use a similar weighting or ranking scheme, but propagate all scores from GO terms to parent terms in the hierarchy of the GO DAG. The final score given to any predicted term is normalized to the total score of the ontology root, to which every predicted term contributes. Essentially, these tools attempt to simulate the human approach to interpreting the results of a BLAST search with an automated scoring scheme. The significant common feature between all of these methods is the use of consensus among similar sequences retrieved by BLAST, an appropriate response to the growing body of evidence suggesting that the traditional one-to-one approach of function annotation with BLAST, which transfers function from a single sequence hit, is error-prone when applied on a large scale[11,23,24] and across diverse protein families.[25–27] The earlier methods, however, have two significant drawbacks in their application to large scale protein function annotation. First, they limit their use of BLAST to previously defined significance thresholds, only scoring GO terms associated to low *E*-value sequence matches. The failure of BLAST to adequately identify homologous proteins for up to half of the genes in a genome at significant *E*-values is not accounted for. Second, none of these methods output reliability scores to describe real confidence in predicted GO annotations.

We have designed our function prediction method, PFP, with large scale annotation projects and the limitations of previous methods in mind.[28] Similarly to other methods, PFP combines GO terms associated to PSI-BLAST[29] sequence hits using an *E*-value based scoring scheme, propagating scores to parental terms on the GO DAG according to the number of known sequences annotated with parent and child terms. To mine the maximal amount of functional data from this type of approach, we have included three additional novel components. First, PFP utilizes functional information from PSI-BLAST sequence hits up to an *E*-value of higher than 100, well beyond accepted thresholds for direct sequence homology. Second, PFP uses data mining to find closely related GO terms to those that can be predicted directly from sequence matches. These features allow PFP to predict function for those sequences lacking annotated homologs in the database, extracting and inferring functional information not available in the traditionally utilized range of sequence similarity. Third, we have developed a method for assigning confidence scores to GO term predictions based on accuracy evaluations over a set of benchmark sequences. This is a key element for the application of predictions to uncharacterized proteins. Our previous report gave anecdotal evidence of the ability of PFP to accurately predict functional annotations to new proteins on a benchmark dataset of 2000 sequences and in the AFP-SIG '05[30] prediction server assessment. Additional evidence of the success of this method can be taken from the CASP7 (Critical Assessment of Techniques for Protein Structure Prediction) function prediction assessment, where our group was named the best predictor on both previously known and new GO term annotations, outperforming even consensus predictions made by the organizers.[31]

This manuscript provides evidence of the superior predictive power of PFP through a comparison of coverage and accuracy against a more traditional use of PSI-BLAST, the GOtcha method, and InterPro[32] motif scanning on a benchmark set of 120,260 protein sequences (see Fig. 1). We also describe extensive testing of variable input parameters, including source and strength of functional annotations for the BLASTed database, weighting schemes for assigning confidence scores to predictions, and function association scores used for the data mining component. The results show solid evidence that the use of higher *E*-value sequence hits and functional associations recovered by data mining are an appropriate interpretation and extension of PSI-BLAST search results. We also analyze the relevance of sequence based approaches to predictions of different functional subcategories. Lastly, we describe the development of our weighting scheme for assigning confidence scores to GO functional term predictions and apply blind PFP predictions to uncharacterized sequences in several genomes and the *Plasmodium falciparum* (Malaria) PPI network.

METHODS

PFP base

The PFP algorithm uses PSI-BLAST (version 2.2.6) to predict probable GO function annotations in three categories—molecular function, biological process, and cellular component—with statistical significance scores (*P*-value) and expected accuracy within a specified range of edges on the GO directed acyclic graph (DAG). For each

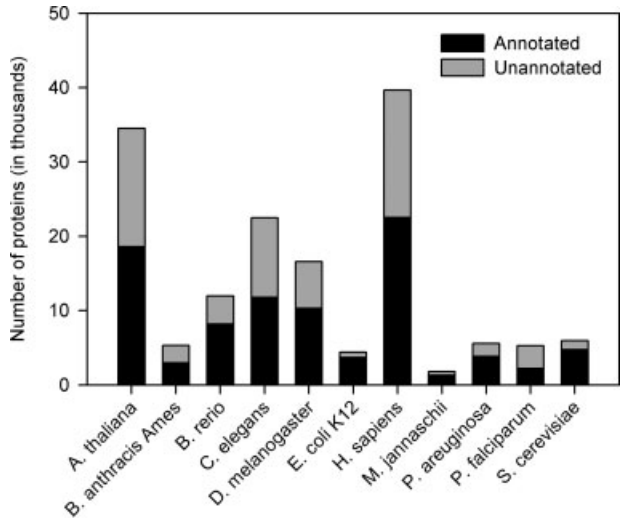

Figure 1

Annotated and unannotated proteins from benchmark organisms. Annotated proteins are associated with at least one GO term in any category by GOA (EBI).

sequence hit retrieved by a PSI-BLAST search, associated GO terms are scored according to the alignment expectation value (*E*-value) provided by PSI-BLAST. The scores for terms associated to several sequence hits are combined by summation. This scoring system ranks GO terms according to both (1) their frequency of association to similar sequences and (2) the degree of similarity those sequences share with the query. A GO term, f_a, is scored as follows:

$$s(f_a) = \sum_{i=1}^{N} \sum_{j=1}^{N\text{func}(i)} \left((-\log(E_{\text{value}(i)}) + b)\delta_{f_j,f_a}\right) \quad (1)$$

where $s(f_a)$ is the final score assigned to the GO term, f_a, N is the number of the similar sequences retrieved by PSI-BLAST, $N\text{func}(i)$ is the number of GO terms assigned to sequence j, $E_{\text{value}(i)}$ is the *E*-value given to the sequence i, and f_j is a GO term assigned to the sequence i. δ_{f_j,f_a} returns 1 when f_j equals to f_a, and 0 otherwise. To maintain the integrity of the PSI-BLAST search, we use the default *E*-value threshold for inclusion in multiple iterations ($-h$ 0.005) and set the maximum number of iterations to three ($-j$ 3). By shifting the scoring space by a constant (b), individual annotations from weakly similar (*E*-value > 1) can be considered and scored. Here we use $b = \log(125)$ to allow the use of sequence matches to an *E*-value of 125.

Term ancestor scoring

Each node in the GO DAG follows the true path rule; that is, any gene associated with a GO term must also be associated with the ancestors of that term leading back to the ontology root. Following this rule, we score ancestors of any predicted term according to the number of genes associated to the predicted term relative to the ancestor term.

$$s(f_p) = \sum_{i=1}^{N_c} \left(s(f_{c_i})\left(\frac{c(f_{c_i})}{c(f_p)}\right)\right) \quad (2)$$

where $s(f_p)$ is the score of the parent term f_p, N_c is the number of child GO term which belong to the parent term f_p, $s(f_{c_i})$ is the score of a child term c_i, and $c(f_{c_i})$ and $c(f_p)$ is the number of known genes, which are annotated with function term f_{c_i} and f_p. For our benchmark evaluation here, we have tested PFP both using ancestral scoring and not using ancestral scoring.

FAM threshold

PFP also uses a novel data mining tool to predict additional GO terms, which are highly associated to those terms associated to sequence hits from PSI-BLAST. This tool, the Function Association Matrix, describes the probability that two GO terms are associated to the same sequence based on the frequency at which they co-occur in UniProt sequences. This allows the FAM to associate function annotations from different GO categories, for example, the biological process "positive regulation of transcription, DNA-dependent" is strongly associated with the molecular function "DNA binding activity" ($P(0045893|0003677) = 0.455$) and the cellular component "nucleus" ($P(0045893|0005634) = 0.296$). Associations can describe parallel functions that may be defined in multiple categories or complementary functions that are defined in one or more categories.

Including associations precalculated by the FAM, the score given to a function f_a is modified as follows:

$$s(f_a) = \sum_{i=1}^{N} \sum_{j=1}^{N\text{func}(i)} \left((-\log(E_{\text{value}(i)}) + b)P(f_a|f_j)\right), \quad (3)$$

$$P(f_a|f_j) = \frac{c(f_a, f_j) + \varepsilon}{c(f_j) + \mu \cdot \varepsilon'} \quad (4)$$

where $P(f_a|f_j)$ is the conditional probability that f_a is associated with f_j, $c(f_a, f_j)$ is number of times f_a and f_j are assigned simultaneously to each sequence in UniProt, and $c(f_j)$ is the total number of times f_j appeared in UniProt, μ is the size of one dimension of the FAM (i.e. the total number of unique GO terms), and ε is the pseudo-count. A pseudo-count is added to each association under the assumption that the annotated proteins used to generate our matrices represent only a subset of all proteins. Note that FAM is asymmetric, i.e. $P(f_a|f_j) \neq P(f_j|f_a)$.

Table I
Percentage and Number of Binary GO Term Associations at Confidence Thresholds Used in Benchmark Evaluation

FAM threshold	Percentage of associations (%)	Number of associations
10	5.86	29,496
25	2.77	13,940
50	1.26	6,319
75	0.40	1,990
90	0.09	441
Intercategory[a]	58.77	295,710

[a]Only includes associations of two GO terms from different categories.

For our benchmark evaluation here, we have tested several thresholds of significance for using FAM association rules predictively. This includes disregarding any function association information [i.e. Eq. (1)] and thresholds of $P(f_a|f_j) \geq 0.10$, 0.25, 0.50, 0.75, and 0.90. Table I shows the percentage of term associations at each confidence level.

Database annotation

This method of transient annotation fully depends on the availability of GO term associations to sequences retrieved by PSI-BLAST. For our benchmark evaluation here, we have used three sets of annotations for the BLAST database. Both GOA annotation sets were retrieved from the Gene Ontology Annotation (GOA) project[33] at the European Bioinformatics Institute (EBI).

GOA Non-IEA

Each association of a GO term to a sequence provided by GOA is accompanied by an evidence code describing the source of the annotation. The GOA Non-IEA annotation set includes GO terms associated to UniProt sequences with all GO evidence codes except Inferred from Electronic Annotation (IEA), which is considered to be the weakest source of evidence.

GOA All

The GOA All annotation set includes all GO terms in the GOA Non-IEA set plus GO terms associated to UniProt sequences with IEA evidence codes.

PFPDB

The PFPDB annotation set includes all GO terms in the GOA All set plus GO terms translated from all other database annotations associated to UniProt sequences. Other database annotations include HAMAP,[34] InterPro,[35] Pfam, PRINTS, ProDom,[36] PROSITE, SMART,[37] and TIGRFam[38] annotations as well as SwissProt Key Words (Table II). These annotations are translated to corresponding GO terms using mappings downloaded from GO.

Benchmark dataset

The benchmark evaluation was performed on annotated proteins from 11 genome sequences. The proteome sets and corresponding gene associations were downloaded from EBI (UniProt GOA Proteome Sets, 1-2007). Species included were *Bacillus anthracis* Ames (Tax ID: 136), *Drosophila melanogaster* (Tax ID: 17), *Escherichia coli* K12 (Tax ID: 18), *Brachydanio rerio* (Tax ID: 20721), *Pseudomonas aeruginosa* (Tax ID: 36), *Homo sapiens* (Tax ID: 25), *Methanococcus jannaschii* (Tax ID: 28), *Arabidopsis thaliana* (Tax ID: 3), *Saccharomyces cerevisiae* (Tax ID: 40), *Plasmodium falciparum* (Tax ID: 493), and *Caenorhabditis elegans* (Tax ID: 9) for a total of 108,591 annotated target sequences.

To test the appropriateness of using weakly similar sequences from PSI-BLAST, we ran PFP ignoring sequence hits under several *E*-value cutoffs [$E \geq 0$ (all-inclusive), 1e−4, 1e−3, 1e−2, 0.1, 1, 10, and 100]. We also varied the use of ancestral scoring, function association with the FAM matrix, and annotations sets for the BLAST database as described earlier to optimize parameters.

To assess the performance of PFP on the benchmark set, we measured sequence coverage (number of sequences for which a correct prediction is made in the top five ranked by expected accuracy within two edges for each category divided by the total number of sequences queried), annotation specificity, S_P (number of correct annotations (true positives, TP) divided by the total number of annotations [true and false positives, TP + FP), Eq. (5)], and annotation sensitivity, S_N [number of correct annotations divided by the total number of target annotations (true positives and false negatives, TP + FN), Eq. (6)].

Table II
Increase in Annotation Coverage of UniProt Sequences and Term Coverage of the GO with Translated Annotations in PFPDB

	Sequence coverage (%)[a]	GO coverage (%)[b]
SwissProt-GO	13.40	35.70
+ HAMAP	46.50	37.10
+ InterPro	82.30	39.50
+ SW-Keywords	85.00	36.30
+ Pfam	76.00	37.90
+ PRINTS	36.40	36.40
+ ProDom	33.10	36.30
+ PROSITE	54.70	36.30
+ SMART	25.70	35.80
+ TIGRFam	44.70	38.00
(+ all) Total	92.90	41.10

[a]Sequence coverage is the percentage of sequences in UniProt annotated with at least one GO term after addition of translated terms from the format in column 1.
[b]GO coverage is the percentage of terms in the GO vocabulary represented in UniProt after addition of translated terms from the format in column 1.

$$S_P = \frac{TP}{TP + FP} \quad (5)$$

$$S_N = \frac{TP}{TP + FN} \quad (6)$$

Z-scores and *P*-values

We applied two measures of statistical significance to the raw scores output for each prediction in the benchmark set. Both measures, *Z*-score and *P*-value, describe the significance of a raw score relative to the distribution of all raw scores for a single term across the benchmark dataset. The *Z*-score indicates the number of standard deviations above or below the mean raw score, whereas the *P*-value is a discrete probability value for any raw score in the distribution.

Confidence scoring

Expected accuracy is calculated for each *P*-value distribution as an empirical measure against the performance of PFP on the benchmark dataset. For each term, actual accuracy is measured as the percentage of correct predictions within 0, 2, or 4 edges of a target term on the GO DAG for *P*-values in bins of 0.005 (200 bins between 1.000 and 0.000). So, for each term we have three standard curves relating a *P*-value significance score of a blind prediction to its expected accuracy in three levels of resolution by edge distance.

Top PSI-BLAST

For comparison purposes, we also collected a list of GO terms associated to the top *N* PSI-BLAST hits for each of the sequences in the benchmark set at each *E*-value cutoff, where *N* is the number of sequences with an *E*-value above the cutoff it takes to find five unique associated GO terms in each of the three category ontologies.

Gotcha

We ran GOtcha[22] with default parameters for each of the sequences in the benchmark set. To run GOtcha for higher *E*-values, we manually removed significant sequences under each cutoff from local BLAST results before re-running. For evaluation purposes, predicted GO terms were ranked by the *P*-score (confidence score) provided in the text results.

InterProScan

InterProScan performs searches of all of the function family and motif databases encompassed by the InterPro database (ProDom, PRINTS, PIR,[39] Pfam, SMART, and TIGR). We ran InterProScan[32] with default parameters and all databases. When possible, GO terms were extracted from text-formatted results. Otherwise, we used translation tables from Gene Ontology [www.geneontology.org] to find the closest possible GO term to the identified family/domain/motif. InterProScan predictions are compared with PFP, GOtcha, and top PSI-BLAST predictions made with an *E*-value cutoff of 0.

Biological context prediction terms

To assess the performance of PFP on a variety of biological contexts, we created a reduced subset of GO containing 47 terms describing significant subsets of the three ontologies. To do this, we translated the MIPS FunCat[40] vocabulary (already a smaller set of terms) into GO, then eliminated terms with product counts below 3000. This appropriately limits the size of the term set we use for biological context performance assessment while still leaving significantly unique terms. For each of the terms in this reduced set, we evaluate each prediction of the term and all of its children with an expected accuracy of $\geq$0.9. A prediction is counted as correct if the common parent it shares with a target is within the subset of terms (the parent term of the subset or deeper).

Semantic similarity

We used a modified (to include CC similarities which were not included in the original implementation) semantic similarity measure from Schlicker et al.[41] to compare predicted sets of GO terms from all of the included methods (PFP, GOtcha, InterProScan, top PSI-BLAST) to the set of known (target) annotations for each sequence in the benchmark set. The similarity of two individual GO terms c_1 and c_2 is

$$\mathrm{sim}(c_1, c_2) = \max_{c \in s(c_1, c_2)} \left(\frac{2 \log p(c)}{\log p(c_1) + \log p(c_2)} \cdot (1 - p(c)) \right), \quad (7)$$

where $p(c)$ is the annotation frequency of term c relative to the frequency of the ontology root, and $S(c_1,c_2)$ is the set of common ancestor terms between terms c_1 and c_2. The similarity of two sets of terms, GO_j^A and GO_j^B, of respective sizes N and M is calculated by constructing an all-by-all similarity matrix s_{ij}.

$$s_{ij} = \mathrm{sim}\left(GO_i^A, GO_j^B\right), \quad \forall i \in \{1, \ldots, N\}, \forall j \in \{1, \ldots, M\} \quad (8)$$

Because PFP, Gotcha, and top PSI-BLAST provide several fold more GO terms for the prediction set than are used for the target GO term set, we used the top *N* terms from each prediction method, where *N* is the number of

terms in the target set (creating a symmetric similarity matrix). We then use normalized averages of row and column vectors from the similarity matrix to correspond to specificity and sensitivity terms. Row vectors compare the similarity of set A (predictions) to set B (targets) and represent specificity, while column vectors compare the similarity of set B (targets) to set A (predictions) and represent sensitivity.

$$\text{Specificity} = \frac{1}{N}\sum_{i=1}^{N} \max_{1\le i\le M} s_{ij} \tag{9}$$

$$\text{Sensitivity} = \frac{1}{M}\sum_{j=1}^{M} \max_{1\le j\le N} s_{ij} \tag{10}$$

To calculate an overall similarity score for the two term sets, we combined the specificity and sensitivity terms for each GO category:

$$\text{GO}_{\text{score}} = \max\{\text{Specificity}, \text{Sensitivity}\}, \tag{11}$$

where GOscore is any of the three category scores (MFscore, BPscore, CCscore). We differentiate from the Schlicker method only to include cellular component similarity into the overall score, which is calculated as

$$\text{funsim} = \frac{1}{3}\left[\left(\frac{\text{MFscore}}{\max(\text{MFscore})}\right)^2 + \left(\frac{\text{BPscore}}{\max(\text{BPscore})}\right)^2 + \left(\frac{\text{CCscore}}{\max(\text{CCscore})}\right)^2\right]. \tag{12}$$

For our evaluation, max(GOscore) = 1 (maximum possible GOscore) and the range of the funSim score is [0,1].

Blind predictions/whole proteome application

We applied PFP predictions to unannotated protein sequences from 14 organisms. Again, the proteome sets and corresponding gene associations were downloaded from EBI. We categorized the results into six groups by the expected accuracy of the top molecular function term predicted and show the relative increase in genome coverage compared with previously annotated proteins.

PPI network enrichment (*P. falciparum*)

We obtained a protein–protein interaction dataset for *P. falciparum*[42] containing over 2500 unique interactions. To evaluate enrichment of the interaction network, we compared the number of fully (both interaction partners annotated) and partially (one of the interaction partners annotated) annotated interactions before and after application of PFP with unannotated proteins in the dataset. We considered only GO biological process predictions with expected accuracy of $\geq$0.9 for node enrichment.

RESULTS

Database coverage

A unique characteristic of PFP is the mining of functional information from divergent sequence hits retrieved by PSI-BLAST, that is, those with *E*-values well above commonly accepted thresholds for significance. This feature is a key extension to other consensus approaches which we expect to increase the number of sequences for which some function can be predicted. To assess the importance of utilizing the information found in high *E*-value sequences, we evaluated sequence coverage while ignoring significant sequence hits with *E*-values below eight cutoff values. Using optimal parameters (see below), when the complete PSI-BLAST results were used (*E*-value cutoff of 0.0, disregarding self-hits), PFP recovered biological process terms correctly for 73% of the benchmark sequences (see Fig. 2) versus the 68% recovered by the Top PSI-BLAST method, which simply transfers the first five GO terms from the best sequence hits above each *E*-value cutoff (molecular function and cellular component term predictions show the same trend). As expected, this coverage drops as the most significant

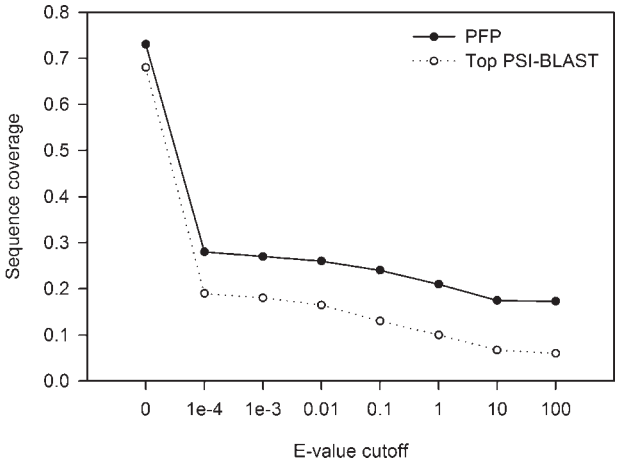

Figure 2

Performance of PFP relative to PSI-BLAST. Sequence coverage (*Y*-axis) is the percentage of benchmark sequences for which a correct GO biological process term was ranked in the top five results. A term is considered correct here if it shares a common ancestor at a GO depth $\geq$1 (GO category root depth = 0) and is within two edges of a known annotation. The *E*-value cutoff (*X*-axis) represents the minimum similarity for sequences from PSI-BLAST considered in the evaluation. PFP predictions (solid line) are ranked by *P*-value significance. PSI-BLAST annotations (dashed line) are the first five unique GO terms above each *E*-value cutoff.

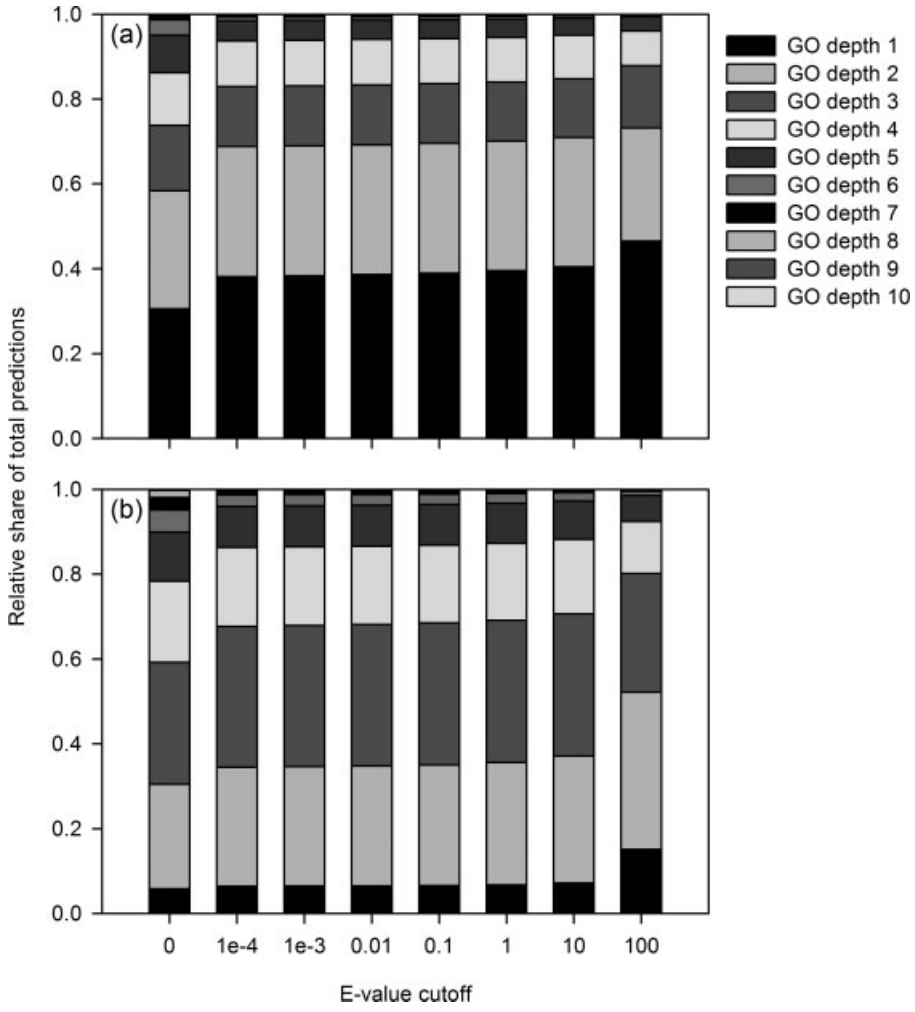

Figure 3

Distribution of (**a**) degree of correctness and (**b**) depth of correct predictions made by PFP. Degree of correctness is the depth of the common parent in the GO DAG shared between a predicted term and the correct answer (GO category root depth = 0).

hits are ignored for both methods. Interestingly, however, even when only higher *E*-value sequences were used, we could still predict correct terms for one in five query sequences while Top PSI-BLAST only predicts correct terms for one in 20 queries. When only sequence hits of *E*-value $\geq$1 were used for making predictions, the coverage of benchmark sequences by PFP more than doubles that of Top PSI-BLAST annotation alone. Sequence hits of *E*-value $\geq$1 are rarely considered to be homologous and are thus ignored by most BLAST-based automated function prediction methods.

Specificity of predicted GO terms

PFP has a distinct advantage in being able to predict more general GO terms when a specific biochemical activity or biological process cannot be predicted from similar sequences. This is apparent when we analyze the average degree of correctness [depth of common parent shared by prediction and target GO annotation, Fig. 3(a)] and depth [Fig. 3(b)] of correct predictions (within two edges) made by PFP at all *E*-value cutoffs. When all sequence hits are utilized, more than 40% of predictions have a common parent with the target annotation at a depth of three or greater, and only about one-third of all correct predictions are made with a common parent depth in the GO of one [Fig. 3(a)]. This contrasts with correct predictions made when all hits below an *E*-value of 100.0 are ignored, where about one-third are made with a common parent depth of three or greater, and more than 40% are made with a common parent depth of one. We consider those predictions of GO terms at shallow depths to be "low-resolution predictions". Interestingly, the same trend is seen when looking at the

depth of the predictions themselves. At an *E*-value cutoff of zero, more than 20% of predicted terms have a depth in the GO DAG of five or greater (very specific), dropping down to 12% at an *E*-value cutoff of 100 [Fig. 3(b)]. These data indicate two very important points. First, that there is still a significant amount of functional similarity between PSI-BLAST hits when taken collectively at higher *E*-values and a query sequence. The significance of this is that somewhat extensive functional information does exist among these sequence hits that previously has not been utilized for predictive purposes. *E*-values of 10 and 100 are well beyond what would normally be considered to infer any reliable or consistent functional relationship between a single sequence hit and the query. Second, that the depth of the correct predictions is somewhat greater than the depth of the common parent they share with target annotations indicates that PFP has a tendency to overpredict (by 1-2 levels in the GO DAG) the functional information contained in the PSI-BLAST results at all *E*-values.

Parameter variation

PFP employs several variable parameters in its scoring scheme, including the source of annotations for the PSI-BLAST database, confidence factor for scoring GO term associations, ancestor term scoring, and the statistical method to determine significance and confidence scores for predictions. To optimize the accuracy of predictions, we tested variations of these parameters in all combinations. Figure 4 shows a comparison of the coverage of our benchmark set when these parameters were adjusted.

We tested three database annotation sets. IEA annotations are assigned to sequences in the database by automated electronic methods without direct experimental evidence [Fig. 4(a)]. GOA Non-IEA is our dataset that excludes these IEA annotations but includes all annotations from GOA supported by other evidence codes. PFPDB incorporates GOA and translations to GO from other functional annotations in UniProt/TrEMBL. Translating functional annotations from other namespaces, such as SwissProt Keywords or Pfam family classifica-

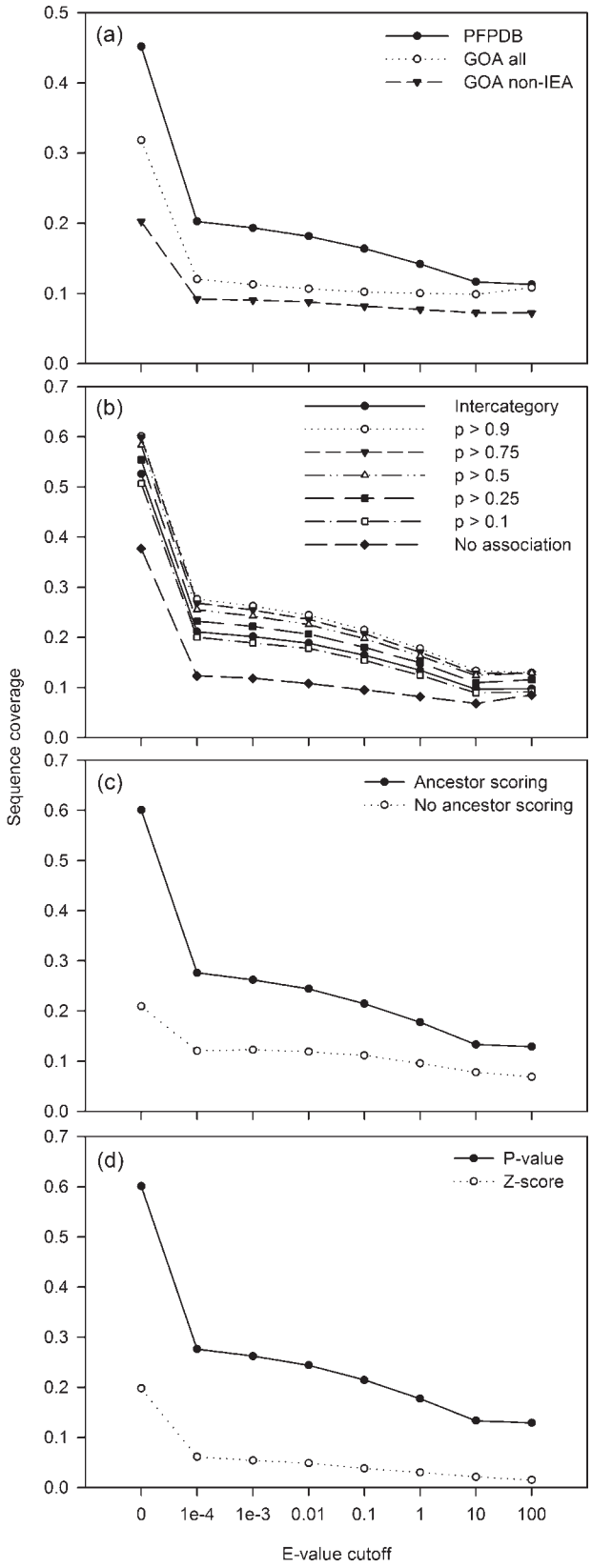

Figure 4

Effects of parameter variation on sequence coverage. (**a**) Effect of different annotation sets for the BLAST database. (**b**) Effect of varying association threshold for annotations retrieved by FAM. (**c**) Effect of scoring GO term ancestors. (**d**) Effect of ranking by significance scores. Sequence coverage (*Y*-axis) is the percentage of benchmark sequences for which a correct GO biological process term was ranked in the top three results (a) or for which a correct GO molecular function term was ranked in the top five results (b-d). A term is considered correct here if it shares a common ancestor at a GO depth $\geq$1 and is within one edge of a known annotation (a) or is an exact prediction of a known annotation (b–d). The *E*-value cutoff (*X*-axis) represents the minimum similarity for sequences from PSI-BLAST considered in the evaluation.

tions, to GO adds usable function annotations to 92.9% of all sequences in UniProt, more than six times the coverage of annotations that originally exist in the database as GO terms (Table II). Figure 4(a) shows clearly that the use of a more comprehensive set of annotations for the PSI-BLAST database results in better prediction coverage of sequence space. Using PFPDB annotations with UniProt, PFP predicts correct biological process terms for 47% of the benchmark sequences as opposed to only 20% using only GOA Non-IEA annotations.

PFP also mines annotated protein sequence databases to find significant associations between GO terms and uses these rules to predict additional terms related to those associated with PSI-BLAST hits [Fig. 4(b)]. We varied the threshold for confidence factor of the association rules used to make these additional predictions. The confidence factor of an association describes the likelihood of the two GO terms having an association truly based on functional relatedness rather than just a happenstantial one. Figure 4(b) shows that as the confidence factor for the association rules represented in our FAM matrix increases, coverage of our benchmark database increases as well, with a threshold of 90% increasing correct molecular function annotations to 60% of the benchmark set. This is a dramatic increase in prediction accuracy compared with the 38% coverage when we omit the data mining component altogether. Using a confidence factor of 90% keeps only the strongest $\sim$ 0.1% of GO term associations (Table I). It is also notable that using only associations across categories contribute to improvement of the accuracy of predicting molecular function [intercategory associations in Fig. 4(b)].

The hierarchical structure of the GO allows us to vary how PFP scores GO terms. We tested the value of scoring a GO term's ancestors leading back up to the category root according to the fraction of sequences annotated with that term over the total number of sequences annotated with the ancestral term. Figure 4(c) shows that this scoring method significantly increases sequence coverage over our benchmark set. Scoring ancestor terms allows consensus of vaguely descript function among sequence hits from PSI-BLAST that may not share specific molecular functions.

Lastly, we tested several methods of assigning significance scores to predictions made by PFP. An essential aspect of PFP for use in blind prediction is the ranking and confidence of predicted GO terms. The distributions of raw scores for each GO term are widely varied (can be seen to a degree in the ranges shown in Fig. 5), so the ranking of predicted terms by raw score is an inadequate representation of the significance of the predictions. Figure 4(d) shows a comparison of sequence coverage using predictions ranked in the top three by two different methods of assigning significance scores. The Z score

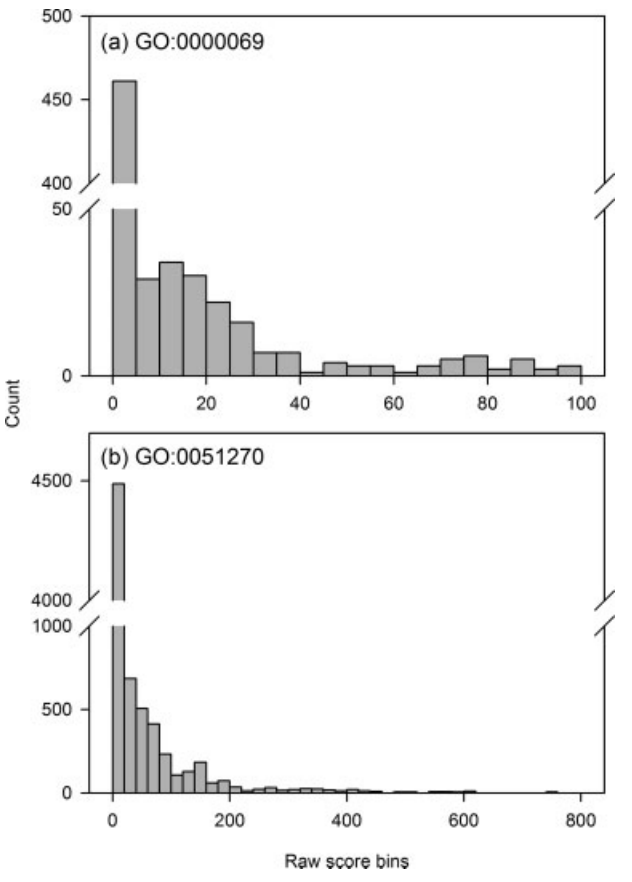

Figure 5

Distributions of raw scores for two Gene Ontology terms. Number of bins is determined by $3+\log(N)\log_2(N)$, where N is the number of data points.

indicates statistical significance of the raw score for each prediction relative to the score distribution for that term across the whole benchmark set. P-value is a probability of the raw score for a predicted term relative to the same distribution. The difference in coverage is up to fivefold between the two methods, showing the impact that significance score assignment has on prediction accuracy. P-value is clearly a more appropriate for determining prediction significance, likely because Z-score works best for distributions closely fitting to the normal distribution whereas P-value can be more generally applied to the distributions of scores we see in this evaluation (see Fig. 5). On the basis of these results, for the performance comparison of PFP to Top PSI-BLAST, GOtcha, and InterProScan, we used the best performing set of parameters here: PFPDB for annotations to UniProt, a confidence factor threshold of 90% ($P \geq 0.90$) for GO term associations, our unique ancestor scoring function, and P-value significance scores for final prediction confidence assessment and ranking.

Statistical significance and confidence scoring

We assigned statistical significance scores to each of the predicted GO terms output by PFP and related those scores to an estimated confidence for blind predictions. For each term we determined the distribution of raw scores, which was used to assign a term-specific *P*-value to each prediction (see Fig. 5). This value represents the significance of a particular score relative to its distribution; however, for any given term the relationship between statistical significance and accuracy is unique [Fig. 6(a–c)]. We therefore constructed standard curves relating *P*-value significance and specificity for each GO term. For the proteins in our benchmark set, some GO terms are generally predicted with higher accuracy [Fig. 6(c)] and would thus be easier to predict in blind application while others are generally predicted with lower accuracy [Fig. 6(a)] and would be harder to predict in blind application. GO molecular function terms are predicted with better accuracy than cellular component and biological process terms, however the significance scores correlate well with accuracy for all three categories [Fig. 6(d–f)]. These averaged category curves are not used in the actual assignment of confidence scores to predictions, but are useful in showing the general trend that our *P*-value significance score translates effectively into real prediction confidence.

Performance comparison against PSI-BLAST, GOtcha, InterProScan

Of existing sequence-based function annotation methods, GOtcha is the most similar, using consensus to interpret BLAST homology search results. We compared the performance of PFP, GOtcha, function assignment by InterProScan, and top PSI-BLAST using a CASP7 style evaluation, substituting a semantic similarity measure for the depth-based measure used in CASP. Performance between the methods was evaluated based on head-to-head prediction on each of the benchmark sequences. We ranked PFP predictions by our estimated accuracy measure and GOtcha predictions by the *P*-score (confidence score). Figure 7 shows the percentage of benchmark sequences for which PFP predictions had the highest semantic similarity scores to the target annotations. PFP significantly outperforms both GOtcha and Top PSI-BLAST at an *E*-value cutoff of 0, winning over 60% of the cases, and this difference is exaggerated as the *E*-value cutoff increases for both molecular function and biological process terms. At distant *E*-values, PFP provides better predictions than other methods 80% of the time. This type of evaluation, as opposed to sequence coverage (see Fig. 2), more clearly shows the all-around superior ability of PFP to predict relevant functional attributes, especially when only high *E*-value sequence hits are used. Figure 8 shows the average FunSim score (total semantic similarity) for all predictions. This score effectively describes the degree of significance of predictions with reference to depth and information content of predicted terms in the GO graph. The top ranked predictions by PFP have an average FunSim score of 0.79 for an *E*-value cutoff of 0 and maintain an advantage over GOtcha and top PSI-BLAST predictions at all *E*-values. Additionally, Table III shows four examples of proteins for which PFP was able to make correct predictions with an *E*-value cutoff of 10.0. The relatively poor performance of InterProScan (for its single data point; *E*-value adjustment is only applicable for the BLAST-based methods) is notable as an indicator of the low coverage of using only strongly conserved functional motifs to predict protein function.

It should be expected that sequence similarity is more likely to be able to predict certain categories of biological functions than others. We assessed the performance of PFP for 47 biological contexts, each represented by a subset of terms from the GO (Fig. 9, see methods section for a description of how these terms were selected). Clearly the accuracy varies between contexts, with predictions of nucleotide binding, transport, and response to stimulus achieving prediction specificity of 90–100%, and predictions of lipid metabolism only achieving only 23%. These more poorly predicted categories represent either areas of biology that are less well understood or those that include a very diverse population of proteins that participate in other biological processes as well. Both of these factors limit the importance of sequence in identifying the role of a protein in a particular biological context.

Functional enrichment of whole proteomes

With established significance scores and a method for relating *P*-value to an estimated accuracy, we were able to assign predictions to unannotated proteins in fifteen genomes and assess the ability of PFP to enrich the functional knowledge of these proteomes at any given confidence level. For each proteome, we counted the number of unknown proteins for which we could make a molecular function prediction with an estimated accuracy of greater than 80%. In each of the fifteen organisms we analyzed, more than two-thirds of the previously unknown proteins could be assigned a GO molecular function term at the highest confidence level, and nearly 100% of these proteins could be assigned a term with an estimated accuracy of 40% or higher (see Fig. 10). The significance of this increased coverage is that after application of PFP, nearly all of the protein content of an organism can be predicted to have some functionality; a computationally derived functional hypothesis is made for each sequence, even those for which only low-resolution predictions can be made.

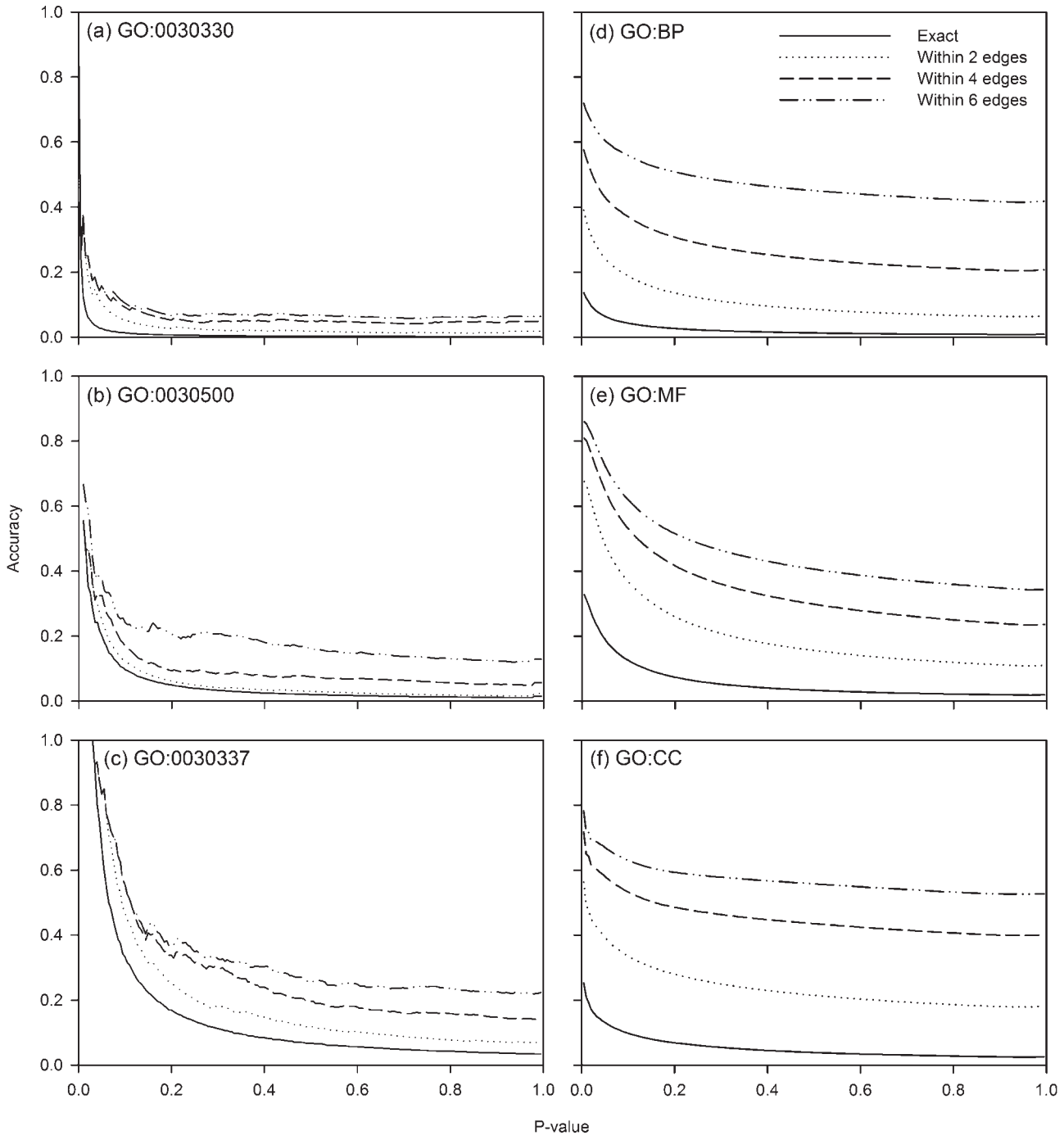

Figure 6

Accuracy as a function of *P*-value significance score. Accuracy (*Y*-axis) is the percentage of annotations sharing a common ancestor at GO depth ≥1 and within zero, two, four, or six edges of the known annotation for three sample GO terms (panels **a** through **c**) and for all annotations in the GO (**d**) biological process, (**e**) molecular function, and (**f**) cellular component categories. It is shown as a function of the *P*-value significance score (*Y*-axis). These are used to create standard curves relating *P*-value to expected accuracy in blind predictions.

Functional enrichment of protein–protein interaction networks

We applied PFP to unknown proteins in the protein interaction network for *P. falciparum* (malaria plasmodium). Interactions here can be divided into three categories: (1) fully enriched, that is, those where both proteins have some known or electronically assigned function, (2) partially enriched, that is, those where only one

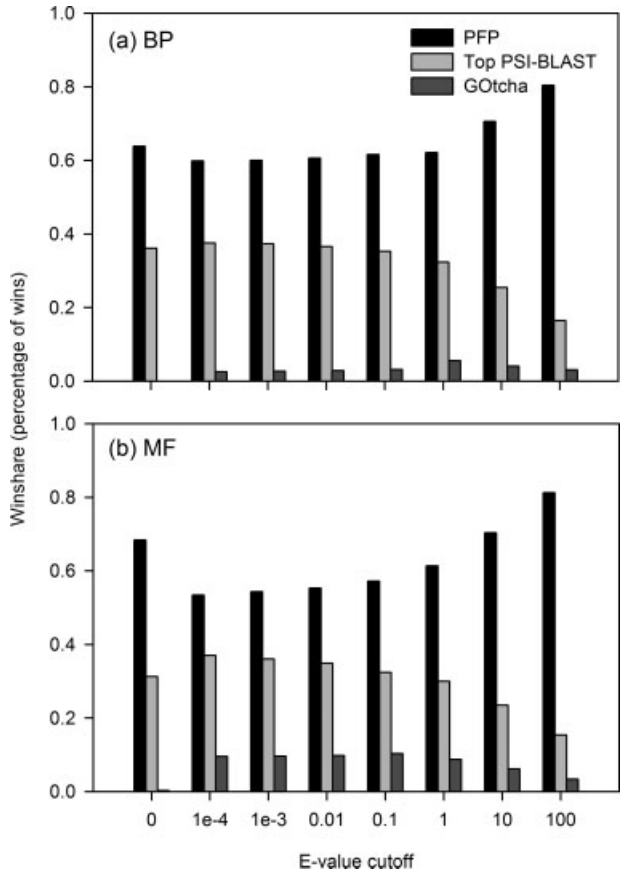

Figure 7

Percentage of wins (*Y*-axis) for head-to-head comparison of PFP, GOtcha, and top PSI-BLAST. Comparisons are made on the basis of the semantic similarity specificity term [Eq. (9)] for GO biological process (**a**) and molecular function (**b**) categories. The *E*-value cutoff (*X*-axis) represents the minimum similarity for sequences from PSI-BLAST considered in the evaluation.

of the two proteins has some known function, and (3) those where neither of the proteins has some known function. For the *P. falciparum* network, we could increase the number of fully enriched interactions by more than fourfold, to over 93% of the total interactions, using GO biological process predictions with an estimated accuracy of over 90% (Table IV). This is an indication of the potential utility of application of PFP prediction to this type of proteomics dataset.

DISCUSSION

PFP was designed to provide a tool for diverse applications of protein function prediction on scales ranging from single sequences to complete genomes independent of the availability of experimental data (other than primary sequence) for the target set. Thus, applications range from functional annotation of new genome sequences to interpretation of large microarray or PPI datasets. Using PSI-BLAST and thorough database functional annotations, PFP predicts GO terms for a query sequence and provides several confidence measures for each prediction. The statistical *P*-value measure relays the relative significance of a prediction score, while the related expected accuracy measure relays the confidence in the prediction accuracy. Prediction confidence scores are key features which should be included in any application of electronic functional annotation to limit propagation of annotation error through sequence databases.[43,44] This presentation of expected accuracy scores is the first application of confidence scores for function predictions that directly relates to an actual reliability score for each predicted GO term, a significant departure from the direct use of BLAST *E*-value for this purpose.

In this manuscript we show the use of Top PSI-BLAST as a baseline for function prediction performance. The comparison is natural, as sequence alignment has been utilized since its inception to infer evolutionary relatedness and subsequently functional similarity. Database searching using FASTA, BLAST, and PSI-BLAST was invented before the new "omics" era in biology, and is often insufficient for obtaining large coverage in function annotation, which is essential for biological interpretation of omics data. PFP greatly improves on this by providing low-resolution function with a statistical significance score when detailed function is not available. The conservation of low resolution function among high *E*-value sequence hits may also be an indication that the evolution of some protein families somewhat follow the structure of the GO vocabulary, that is evolutionary distance may correlate with edge distance in the GO and addi-

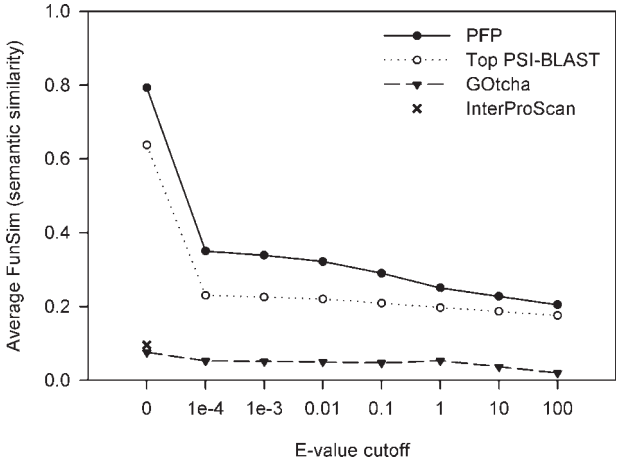

Figure 8

Average funSim scores (*Y*-axis) for PFP, GOtcha, top PSI-BLAST, and InterProScan over all *E*-value cutoffs. The *E*-value cutoff (*X*-axis) represents the minimum similarity for sequences from PSI-BLAST considered in the evaluation.

Table III
Examples of Correction Predictions by PFP Using only BLAST Sequence Hits Above an *E*-Value of 10

Protein ID	GOA Annotations	*E*-Value of Most Distant Homolog[a]	PFP Predictions	Rank[b]
ARGR_BACAN	Cytoplasm (GO:0005737)	0.95	DNA binding (GO:0003677)	1
Arginine repressor	DNA binding (GO:0003677)		Transcription factor activity (GO:0003700)	2
	Transcription factor activity (GO:0003700)			
	Transcription (GO:0006350)			
	Regulation of transcription, DNA-dependent (GO:0006355)			
	Arginine metabolic process (GO:0006525)			
	Arginine biosynthetic process (GO:0006526)			
	Amino acid biosynthetic process (GO:0008652)			
MTNK_BACAN	Kinase activity (GO:0016301)	0.00008	Transferase activity (GO:0016740)	1
Methylthioribose kinase	Transferase activity (GO:0016740)		Kinase activity (GO:0016301)	3
	S-methyl-5-thioribose kinase activity (GO:0046522)			
	Amino acid biosynthetic process (GO:0008652)			
	Methionine biosynthetic process (GO:0009086)			
Q81RY3_BACAN	Copper ion binding (GO:0005507)	100 (next highest is 52.3)	Copper ion binding (GO:0005507)	1
Multicopper oxidase family protein	Oxidoreductase activity (GO:0016491)		Oxidoreductase activity (GO:0016491)	10
ATP6_DROME	Mitochondrion (GO:0005739)	0.14	Proton-transporting ATPase complex, coupling factor F(o) (GO:0045263)	1
ATP synthase 6	Membrane (GO:0016020)		Proton-transporting two-sector ATPase complex (GO:0016469)	2
	Integral to membrane (GO:0016021)		Integral to membrane (GO:0016021)	3
	Proton-transporting two-sector ATPase complex (GO:0016469)		Mitochondrion (GO:0005739)	4
	Proton-transporting ATPase complex, coupling factor F(o) (GO:0045263)		Hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances (GO:0016820)	2
	Hydrogen-exporting ATPase activity, phosphorylative mechanism (GO:0008553)		Hydrogen ion transmembrane transporter activity (GO:0015078)	4
	Hydrogen ion transmembrane transporter activity (GO:0015078)		Proton transport (GO:0015992)	2
	Hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances (GO:0016820)			
	Transport (GO:0006810)			
	Ion transport (GO:0006811)			
	Proton transport (GO:0015992)			

[a]Most distant homolog is the least similar sequence in PSI-BLAST results recognizable as directly homologous to the query (sharing the same protein name). Sequences with *E*-values below (more significant than) 10.0 were ignored in making predictions.
[b]Ranks are given according to the *P*-value.

tionally that low-resolution function may be conserved beyond obviously similar homologous sequences.

We additionally compared the performance of PFP to GOtcha and InterProScan.[32] GOtcha was used because of the similarity of its base algorithm to PFPs. Both use a consensus approach to rank terms assigned to sequences retrieved by a BLAST search. Figures 7 and 8 summarize the results of applying a semantic similarity measure (slightly modified FunSim[41]) as an indicator of prediction correctness. PFP clearly shows increased performance by these measures. This is consistent with previous third party assessments of the predictive ability of PFP. It has been recently noted as the top overall predictor in the CASP7 function category, even outperforming consensus methods used by the evaluators.[31] It should be noted that the performance of GOtcha seems abnormally low. This is a result of GOtcha's ranking scheme, which weights GO terms closest to the root node higher (with more confidence). Thus the highest confidence terms have lower real significance in terms of the information provided to users. Table III provides an excellent example of how PFP's weighting predictions by confidence score can retrieve and highlight broad but significant GO terms from even high *E*-value (low significance) sequence hits.

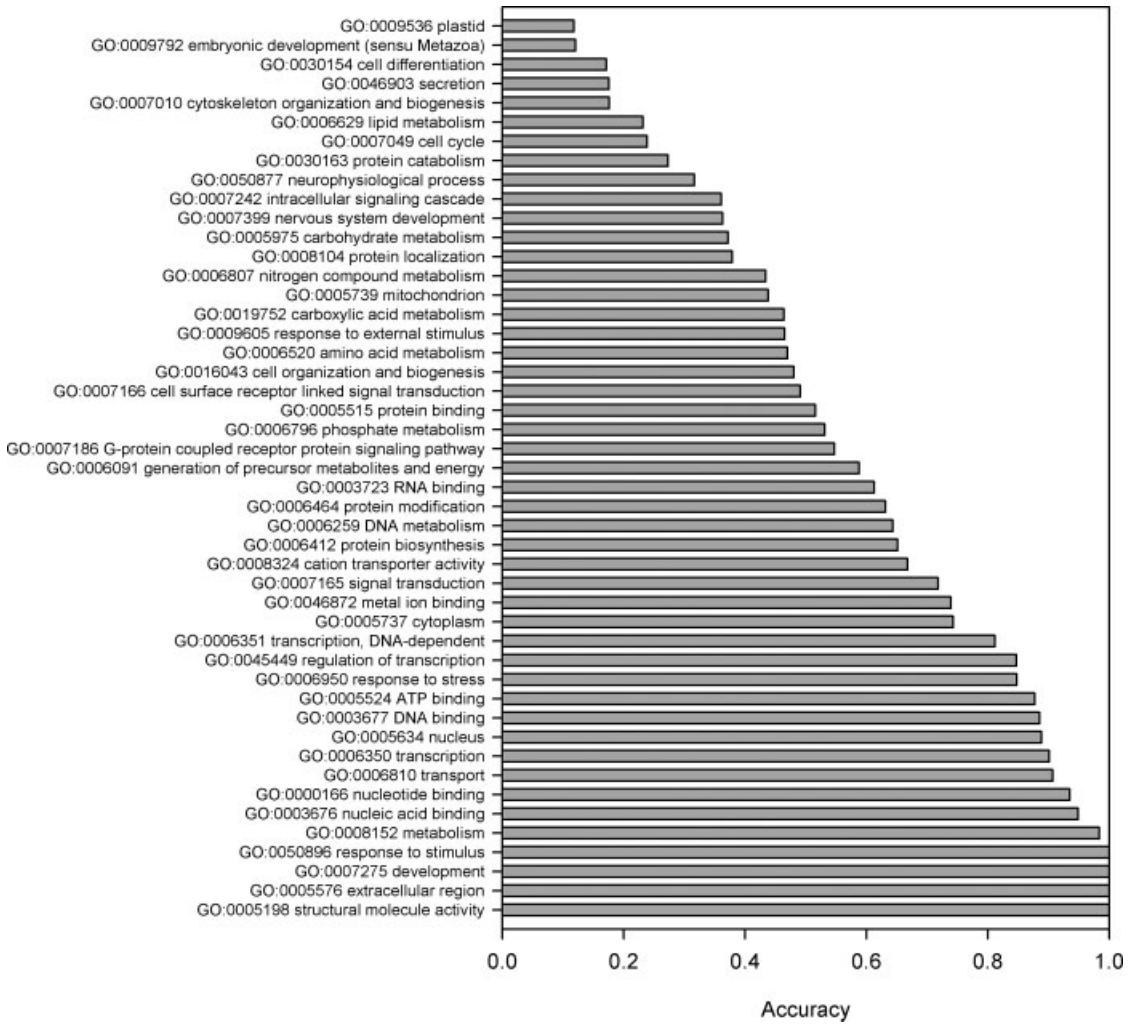

Figure 9

Prediction accuracy by GO category. Accuracy (*X*-axis, percentage of correctly predicted annotations) is shown for 47 biological contexts, represented by a term and all of its descendants in the GO DAG. Predictions are considered to be correct here if they share the root term of the subgraph with a known annotation (e.g. any term evaluated for the context represented by GO:0009805 response to external stimulus is considered correct if it shares that term with a known annotation).

The important functional features are found without relying on close homologs.

Here we have shown two such applications of PFP to large scale datasets. First, we applied PFP predictions to proteins from several genomes. The resulting increase in coverage (see Fig. 10) can be quite dramatic. Although this is a basic application, it is one that has implications for several more complex interpretive algorithms in bioinformatics. For example, consider the related problems of missing gene and metabolic pathway prediction. A common problem in biology is the existence of "missing" genes or pathway elements, that is, reactions or steps in metabolic pathways, which are assumed to exist but are not associated to a particular gene product. In some instances, only one or a few elements are missing from a pathway. However, in the case that a large portion of one pathway is "missing", the possibility exists that the pathway does not occur in the assumed form. Current approaches use the functional makeup of the particular organism's genome to infer the existence of one pathway alternative over another, but rely on BLAST or other sequence similarity methods for the step of functional characterization of the genome.[45–50] This leaves open the real possibility that there is a significant portion of the enzyme content of the organism that is not considered. PFP gives two advantages here, namely that prediction coverage for a target genome will be better with PFP than BLAST and that low-resolution, broad prediction of

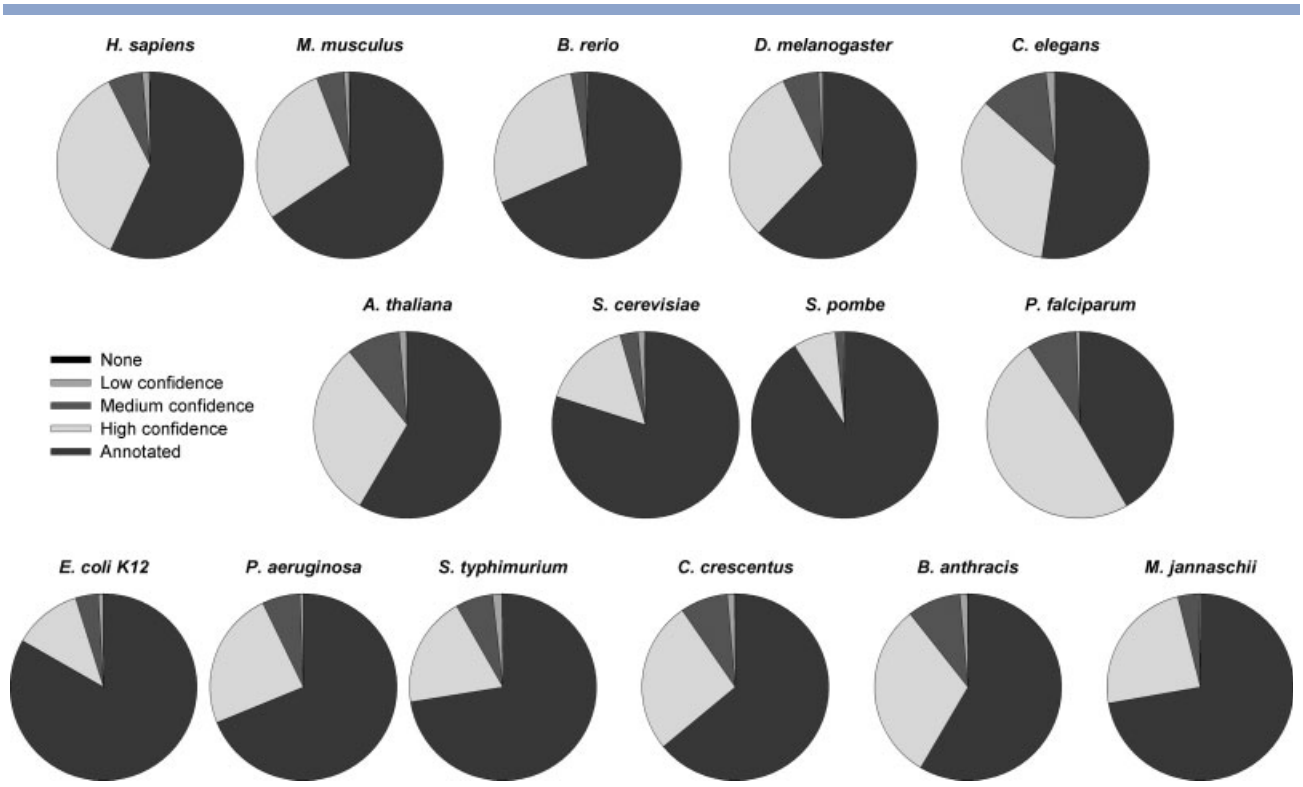

Figure 10

Blind prediction of uncharacterized proteins. Pie pieces represent the percentage of proteome sequences in each category that are previously annotated or predicted to be correct with high (≥80%), medium (≥60%), or low (≥40%) confidence within two edges on the GO DAG. Genomes are organized by phylogeny.

function can still be useful for pathway inference when detailed information about a suspected enzyme cannot be predicted by orthology.

Second, we applied PFP predictions to proteins in the PPI network for *P. falciparum*. We showed that nearly all of the interactions could be fully enriched with function annotation by PFP. Methods in bioinformatics which utilize PPI networks as a primary information source rely on availability of functional information for proteins therein. These methods are used to find clusters of interacting proteins that may be involved in a common biological process[51–54] or to functionally characterize an unknown protein based on the function(s) of its interaction partners.[55,56] Both of these instances will benefit from the ability of PFP to maximize prior functional knowledge and provide reliability scores for each node in the interaction graph.

A future direction is to combine different sources of function information to PFP, as different sources have strength in different biological categories.[57] In summary, PFP is a sequence-based method for protein function prediction, providing a set of GO terms with both *P*-value significance and expected accuracy confidence scores. It is ideally suited for large scale applications, especially omics data analysis, as it provides better coverage in function annotation by providing low-resolution function. PFP is implemented in a web server at http://dragon.bio.purdue.edu/pfp/.

Table IV

Functional Enrichment of the Protein–Protein Interaction Network of *P. falciparum*

	Previously Annotated[a]	Enriched with Predictions (90%)[a]
BOTH	664	2,674
ONE	1,358	168
NEITHER	824	4
TOTAL	2,846	2,846

"BOTH" describes interactions for which both partners are annotated, "ONE" describes interactions for which only either the bait or prey protein is annotated.
[a]GO biological process terms.

REFERENCES

1. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC. The genomes on line database (GOLD) v. 2: a monitor of genome projects worldwide. Nucleic Acids Res 2006;34(Database issue):D332–D334.
2. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 2007;36(Database issue):D475–D479.

3. Friedberg I. Automated protein function prediction—the genomic challenge. Brief Bioinform 2006;7:225–242.
4. Godzik A, Jambon M, Friedberg I. Computational protein function prediction: are we making progress? Cell Mol Life Sci 2007;64: 2505–2511.
5. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. Curr Opin Struct Biol 2005;15:275–284.
6. Hawkins T, Kihara D. Function prediction of uncharacterized proteins. J Bioinform Comput Biol 2007;5:1–30.
7. Hawkins T, Chitale M, Kihara D. New paradigm in protein function prediction for large scale omics analysis. Mol BioSyst 2008;4:223–231.
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
9. Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 1990;183:63–98.
10. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
11. Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. J Mol Evolut 2001;52(6):540–542.
12. Bork P, Koonin EV. Predicting functions from protein sequences—where are the bottlenecks? Nat Genet 1998;18:313–318.
13. Huynen MA, Snel B, von Mering C, Bork P. Function prediction and protein networks. Curr Opin Cell Biol 2003;15:191–198.
14. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003;4:41.
15. Brenner SE. Errors in genome annotation. Trends Genet 1999;15: 132–133.
16. Galperin MY, Koonin EV. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol 1998;1:55–67.
17. Groth D, Lehrach H, Hennig S. GOblet: a platform for Gene Ontology annotation of anonymous sequence data. Nucleic Acids Res 2004;32(Web Server issue):W313–W317.
18. Hennig S, Groth D, Lehrach H. Automated gene ontology annotation for anonymous sequence data. Nucleic Acids Res 2003;31: 3712–3715.
19. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–29.
20. Zehetner G. OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Res 2003;31:3799–3803.
21. Khan S, Situ G, Decker K, Schmidt CJ. GoFigure: automated gene ontology annotation. Bioinformatics (Oxford England) 2003;19: 2484–2485.
22. Martin DM, Berriman M, Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 2004;5:178.
23. Muller A, MacCallum RM, Sternberg MJ. Benchmarking PSI-BLAST in genome annotation. J Mol Biol 1999;293:1257–1271.
24. Agarwal P, States DJ. Comparative accuracy of methods for protein sequence similarity search. Bioinformatics (Oxford England) 1998; 14:40–47.
25. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. Proteins 1993;17:49–61.
26. Rost B. Enzyme function less conserved than anticipated. J Mol Biol 2002;318:595–608.
27. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 2003;333:863–882.
28. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 2006;15:1550–1556.
29. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
30. Friedberg I, Jambon M, Godzik A. New avenues in protein function prediction. Protein Sci 2006;15:1527–1529.
31. Lopez G, Rojas A, Tress M, Valencia A. Assessment of predictions submitted for the CASP7 function prediction category. Proteins 2007;69(Suppl 8):165–174.
32. Zdobnov EM, Apweiler R. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics (Oxford England) 2001;17:847–848.
33. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 2004;32(Database issue):D262–D266.
34. Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, Lima T, Kersey P, Pagni M, Sigrist CJ, Lachaize C, Veuthey AL, Gasteiger E, Bairoch A. Automated annotation of microbial proteomes in SWISS-PROT. Comput Biol Chem 2003;27:49–58.
35. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. New developments in the InterPro database. Nucleic Acids Res 2007;35(Database issue):D224–D228.
36. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D. ProDom: automated clustering of homologous domains. Brief Bioinform 2002;3:246–251.
37. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 2006;34(Database issue):D257–D260.
38. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res 2003;31:371–373.
39. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, Barker WC. The protein information resource. Nucleic Acids Res 2003;31:345–347.
40. Tetko IV, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Fobo G, Ruepp A, Antonov AV, Surmeli D, Mewes HW. MIPS bacterial genomes functional annotation benchmark dataset. Bioinformatics (Oxford, England) 2005;21:2520–2521.
41. Schlicker A, Domingues FS, Rahnenfuhrer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 2006;7:302.
42. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, Fields S, Hughes RE. A protein interaction network of the malaria parasite *Plasmodium falciparum*. Nature 2005;438:103–107.
43. Kyrpides NC, Ouzounis CA. Whole-genome sequence annotation: 'Going wrong with confidence'. Mol Microbiol 1999;32:886–887.
44. Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA. Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics (Oxford England) 2002;18:1641–1649.
45. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 2005;6:R2.

46. Gaasterland T, Selkov E. Reconstruction of metabolic networks using incomplete information. Proceedings/international conference on intelligent systems for molecular biology; ISMB 1995;3: 127–135.
47. Ma H, Zeng AP. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics (Oxford England) 2003;19: 270–277.
48. Tsoka S, Simon D, Ouzounis CA. Automated metabolic reconstruction for *Methanococcus jannaschii*. Archaea (Vancouver BC) 2004;1:223–229.
49. Paley SM, Karp PD. Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. Bioinformatics (Oxford England) 2002;18:715–724.
50. Arakaki AK, Tian W, Skolnick J. High precision multi-genome scale reannotation of enzyme function by EFICAz. BMC Genomics 2006;7:315.
51. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol 2000;18:1257–1261.
52. Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, Jacq B. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol 2003;5:R6.
53. Samanta MP, Liang S. Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl Acad Sci USA 2003;100:12579–12583.
54. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 2003;21:697–700.
55. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003;4:2.
56. Brun C, Herrmann C, Guenoche A. Clustering proteins from interaction networks for the prediction of cellular functions. BMC Bioinformatics 2004;5:95.
57. Myers CL, Troyanskaya OG. Context-sensitive data integration and prediction of biological networks. Bioinformatics (Oxford, England) 2007;23:2322–2330.