HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2013 September 01.

Published in final edited form as:

Nat Methods. 2013 March ; 10(3): 221–227. doi:10.1038/nmeth.2340.

A large-scale evaluation of computational protein function prediction

A full list of authors and affiliations appears at the end of the article.

Abstract

Automated annotation of protein function is challenging. As the number of sequenced genomes rapidly grows, the overwhelming majority of protein products can only be annotated computationally. If computational predictions are to be relied upon, it is crucial that the accuracy of these methods be high. Here we report the results from the first large-scale community-based Critical Assessment of protein Function Annotation (CAFA) experiment. Fifty-four methods representing the state-of-the-art for protein function prediction were evaluated on a target set of 866 proteins from eleven organisms. Two findings stand out: (i) today's best protein function prediction algorithms significantly outperformed widely-used first-generation methods, with large gains on all types of targets; and (ii) although the top methods perform well enough to guide experiments, there is significant need for improvement of currently available tools.

Introduction

The accurate annotation of protein function is key to understanding life at the molecular level and has great biomedical and pharmaceutical implications. However, with its inherent difficulty and expense, experimental characterization of function cannot scale up to the vast amount of sequence data already available.[1] The computational annotation of protein function has therefore emerged as a problem at the forefront of computational and molecular biology.

Many solutions have been proposed in the last four decades,[2-10] yet the task of computational functional inference in a lab often relies on traditional approaches such as domain identification or finding BLAST[11] hits among proteins with experimentally determined function. Recently, the availability of genomic-level sequence information for thousands of species, coupled with massive high-throughput experimental data, has created new opportunities for function prediction. A number of methods have been proposed to

Contact: Predrag Radivojac (predrag@indiana.edu), Iddo Friedberg (i.friedberg@miamioh.edu).

Author contributions

PR and IF conceived the CAFA experiment, supervised the project, and wrote most of the manuscript. SDM participated in the design of and supervised the method assessment. WTC performed the analysis of feasibility of the experiment, most of the target and performance analysis, and contributed to writing. PR and WTC designed and produced figures. TRO developed the web interface, including the portal for submission and the storage of predictions. TRO and TW verified the assessment code and participated in analysis. AMS designed and performed the analysis of targets. AB, ML, PCB, SEB, CO, and BR steered the CAFA experiment, provided critical guidance, and participated in writing. The remaining authors participated in the experiment, provided writing and data for their methods, as well as contributed comments on the manuscript.

exploit these data, including function prediction from amino acid sequence,[12-16] inferred evolutionary relationships and genomic context,[17-21] protein-protein interaction networks,[22-25] protein structure data,[26-28] microarrays,[29] or a combination of data types.[30-34] With the large number of methods available, an unbiased evaluation can provide insight into the ability of different tools to characterize proteins functionally and guide biological experiments. So far, however, a comprehensive assessment incorporating a large and diverse set of target sequences has not been conducted due to practical difficulties in providing an accurately annotated target set.

In this report, we present the results of the first Critical Assessment of protein Function Annotation (CAFA) experiment, a worldwide effort aimed at analyzing and evaluating protein function prediction methods. Although protein function can be described in multiple ways, we focus on classification schemes provided by the Gene Ontology (GO) Consortium.[35] Over the course of 15 months, 30 teams associated with 23 research groups participated in the effort, testing 54 function annotation algorithms. These methods were evaluated on a target set of 866 protein sequences from eleven species.

Results

Protein function is a concept that can have different interpretations in different biological contexts. Generally, it describes biochemical, cellular, and phenotypic aspects of the molecular events that involve the protein, including how they interact with the environment (e.g. small compounds or pathogens). From the various classification schemes developed to standardize descriptions of protein function, we chose Molecular Function and Biological Process categories from the Gene Ontology (GO). Each category in GO is a hierarchical set of terms and relationships among them that capture functional information so that it facilitates computation and can be interpreted by humans. GO's consistency across species and its widespread adoption make it suitable for large-scale computational studies. In CAFA, given a new protein sequence, the task of a protein function prediction method is to provide a set of terms in GO along with the confidence scores associated with each term.

The experiment was organized as follows. A set of 48,298 proteins lacking experimentally validated functional annotation was provided to the community four months before the submission deadline for predictions (Fig. 1). Proteins were annotated by the predicting groups and these annotations were submitted to the assessors. After the submission deadline, GO experimental annotations for those sequences were allowed to accumulate over a period of eleven months. Methods were then evaluated on 866 targets from eleven species that had accumulated functional annotations during the waiting period (Supplementary Table 1). The Swiss-Prot database[36] was selected as the gold standard because of its relatively high reliability.[37]

The selection of proteins was ineluctably biased due to experimentalist and annotator choice during the evaluation timeframe. Thus, the set of targets was first analyzed to establish that it was representative of those sequences experimentally annotated before the submission deadline. In terms of organismal representation, the eukaryotic targets provided reasonable coverage of taxa (Fig. 1). In contrast, the set of prokaryotic targets was heavily biased

towards *Escherichia coli* K-12, with 43 annotated sequences from other organisms. The distribution of terms over the target sequences was representative of the annotations in Swiss-Prot (data not shown); however, we note that in the Molecular Function category a large fraction of target sequences (38%) were associated with "protein binding" as their most specific term. The distribution of term depths over all targets is shown in Supplementary Fig. 1 for both ontologies.

Overall predictor performance

The quality of protein function prediction can be measured in different ways, which reflect differing motivations for understanding function. In some cases, imprecise experimental characterization means that it is not entirely clear if a prediction is correct or incorrect. For CAFA, we principally report a simple metric, the maximum F-measure ($F_{\max}$; Online Methods), which considers predictions across the full spectrum from high to low sensitivity. This approach, however, has limitations, such as penalizing specific predictions as discussed in Discussion. We note that the choice of evaluation metric differentially impacts different prediction methods, depending upon their application objectives.

Top predictor performance, based on maximum F-measure and calculated over all targets, is shown in Fig. 2 (the precision/recall curves are shown in Supplementary Fig. 2). All methods are compared with two baseline tools: (i) BLAST, where all GO terms of an experimentally annotated sequence (template) from Swiss-Prot were transferred to the target sequence such that the scores equaled pairwise sequence identity between the template and the target (terms with multiple hits retained the highest score) and (ii) Naïve, where each GO term for each target was scored with the relative frequency of this term in Swiss-Prot over all annotated proteins (Online Methods). We also evaluated the quality of PSI-BLAST predictions but found that it did not provide any advantage over BLAST: specifically, $F_{max}^{PSI-BLAST}=F_{max}^{BLAST}=0.38$ for Molecular Function; $F_{max}^{PSI-BLAST}=0.24$ and $F_{max}^{BLAST}=0.26$ for Biological Process. We believe that the improved ability of PSI-BLAST to identify remote homologs has been canceled out by its re-ranking of close hits.

There is a significant performance difference in the ability to predict the two GO categories (Molecular Function vs. Biological Process). This can be partly explained by the topological differences between the ontologies (number of terms: 8728 vs. 18982; branching factor: 5.9 vs. 6.4; maximum depth: 11 vs. 10; number of leaf terms: 7003 vs. 8125, respectively). However, more fundamentally, terms in the Biological Process ontology are associated with a more abstract level of function. Such terms are less likely to be predictable solely from amino acid sequence, which was the data source used by most methods in this experiment and may critically depend on the cellular and organismal context.

Predictor performance on categories of targets

Easy vs. difficult targets—We divided the target sequences into *easy* and *difficult*. A target was considered easy if it had a 60% or higher sequence identity with any experimentally annotated protein. The threshold of 60% was manually chosen after plotting the distribution of sequence identities between targets and annotated proteins (Supplementary Fig. 4). This resulted in 188 easy and 343 difficult targets in the Molecular

Function category and 247 easy and 340 difficult targets in the Biological Process category. Supplementary Fig. 5 shows the precision/recall curves for both categories. Perhaps unsurprisingly, BLAST outperformed Naïve in the easy target category, while their performance was similar for the difficult targets. More importantly, however, because of the similar performance among top-ranked predictors over easy and difficult targets, the sequence identity-based classification of targets does not seem to accurately reflect the uncertainty associated with a protein's true function (except for BLAST). This outcome is surprising and may be caused by the ability of the methods to compensate for the differences in sequence similarity of the best hit by exploiting multiple sequence hits as well as other data sources.

Eukaryotic vs. prokaryotic targets—Supplementary Fig. 6 shows prediction performance for the eukaryotic and prokaryotic targets. Performance is generally similar in the Molecular Function category, with prokaryotic targets exhibiting higher prediction accuracy in the Biological Process category. We believe this is because most prokaryotic targets came from *E. coli* for which reliable experimental data are available, whereas the data for eukaryotic targets come from sources with highly variable coverage and quality. It is important to note that the particular calculation of precision and recall (see Online Methods) has adversely impacted methods that predicted only on eukaryotic targets (BMRF, ConFunc, GOstruct, Tian Lab) and resulted in lower overall performance for these methods. Detailed results for eukaryotic and prokaryotic targets, as well as several individual organisms are shown in Supplementary Figs. 6-7.

Single- vs. multi-domain targets—We further separated targets into sequences containing a single domain vs. sequences containing multiple protein domains, with domains defined according to Pfam-A classification[38] (targets without any Pfam-A hits were grouped together with single domain proteins). Multi-domain proteins were generally longer; however, they were not associated with more functional terms than single-domain proteins. By analyzing the performance of the top ten methods in each category, we found that although the overall accuracy was higher on single-domain proteins, results were significant only in the Molecular Function category and for eukaryotic targets ($P = 1.4 \cdot 10^{-5}$; $n = 10$; paired t-test; Fig. 3). While generally not surprising, the higher performance on single-domain proteins further emphasizes the need for developing methods that can optimally combine sequence information from multiple domains along with other information to produce a relatively small set of predicted terms.

Predictor performance on functional terms

The ability of methods to predict individual GO terms was assessed by calculating the area under the ROC curve (AUC; Online Methods). To more confidently assess the performance in predicting individual terms, we only considered terms for which at least 15 targets were annotated. Average AUC values were then calculated from the top-five performing models in each ontology, excluding those models that only provide single-score predictions.

Using the above criteria we were able to calculate average AUC values for 28 Molecular Function and 223 Biological Process terms (Supplementary Table 2). We found a clear

distinction between the average AUC of Molecular Function terms generally associated with catalytic and transporter activity, and those associated with binding. In general, the prediction of terms associated with binding showed lower AUC values, even though proteins were biased towards being annotated with binding terms. Among the Biological Process terms, we found, as expected, low AUC values associated with less specific terms such as “locomotion”, “cellular process”, and “response to stress”. We also found that prediction of terms associated with cell adhesion, metabolic process, transcription, and the regulation of gene expression showed high performance. We tested whether high predictor AUC on individual terms was due to high levels of sequence similarity among sequences experimentally annotated with those terms and found a moderate level of correlation (data not shown).

Case study

Here we illustrate some challenges associated with computational protein function prediction. We provide a detailed analysis of the human mitochondrial polynucleotide phosphorylase 1 (hPNPase; *PNPT1*), a large (783aa) protein with seven Pfam domains (Fig. 4A). Human PNPase is characterized by several experimentally determined functions, making it an attractive target to evaluate the performance of prediction methods. hPNPase belongs to a family of exoribonulceases, which hydrolyze single-stranded RNA in the 3’-to-5’ direction. In complex with other components of the mitochondrial degradasome it mediates the translocation of small RNAs into the mictochondrial matrix.[39] It is also proposed to be involved in several biological processes including cell-cycle arrest,[40] cellular senescence, and response to oxidative stress.[41]

Due to its involvement in several molecular functions and biological processes, the comprehensive and accurate listing of functions of hPNPase is a challenging task. Furthermore, while polynucleotide phosphorylase 1 is prevalent in bacteria and eukarya, it has accumulated several lineage-specific functions. Specifically, while bacterial and chloroplast PNPase have demonstrated exoribonuclease and polyadenylation activities, hPNPase functions predominantly as an RNA importer,[39] with exoribonuclease activity shown only *in vitro*.[42] Finally, hPNPase is a mitochondrial protein found in the inter-membrane matrix. Taken together with its involvement in the rRNA import process, this suggests the need to predict cellular compartment as part of a comprehensive understanding of function.

Figure 4B shows the experimental GO term annotation of hPNPase as well as the terms predicted by a representative set of the top-ten performing methods. Within the Molecular Function terms, none of the methods predicted poly(U/G) RNA binding[43] or micro RNA binding. However, most methods that did predict function correctly predicted 3’-5’ exoribonuclease activity and polyribonucletide nucelotidyltransferase activity. It should be noted that poly(U/G) binding and micro RNA binding are not common throughout the PNPase lineage. This may be the reason why none of the programs predicted these terms.

In the Biological Processes category, the most prominent function of hPNPase in the literature is the import of nuclear 5S rRNA into the mitochondrion;[39] indeed, it is hypothesized that this is the reason for hPNPase’s location in the inter-membrane matrix.

However, this function, along with other important terms, such as cellular senescence, was not predicted by any of the top-performing methods at the optimal threshold levels. Generally, the biological process predictions were highly non-specific for most models. In sum, the multi-domain architecture of hPNPase, its pleiotropy, and the different functions it assumes in different taxa all contribute to the challenge of correctly predicting its function.

Discussion

Protein function is difficult to predict for several reasons. First, function is studied from various aspects and at multiple levels; e.g. it describes the biochemical events involving the protein, and also how each protein affects pathways, cells, tissues, and the entire organism. Second, protein function and its experimental characterization are context-dependent: a particular experiment is unlikely to determine a protein's entire functional repertoire under all conditions (e.g. temperature, pH, presence of interacting partners). Third, proteins are often multifunctional[44] and promiscuous;[45] in fact, 30% of the experimentally annotated proteins in Swiss-Prot have more than one leaf term in the Molecular Function ontology and 60% in the Biological Process ontology.[16] Fourth, in addition to being incomplete, available functional annotations are error prone due to experiment interpretation or curation issues.[37, 46] Finally, current efforts largely map protein function to gene names, thus confounding the functions of potentially diverse isoforms. Despite these challenges, the CAFA experiment revealed progress in automated function annotation over the past decade.

Top algorithms are useful and significantly outperform BLAST

The first generation of function prediction methods performed a simple function transfer via pairwise sequence similarity; i.e. the most-similar annotated hit was used as the basis of function prediction.[47] Several studies have been aimed at characterizing performance of these methods.[3, 16, 48] The CAFA experiment provides evidence that the best algorithms universally outperform simple functional transfer. The experiment also showed that BLAST is largely ineffective at predicting functional terms related to the Biological Process ontology. This is possibly due to homologs assuming different biological roles in different tissues and organisms.[49]

Principles underlying best methods

The methods evaluated in CAFA exploited a variety of biological and computational concepts. Most methods exploited sequence alignments with an underlying hypothesis that sequence similarity is correlated with functional similarity. Recent studies have shown that this correlation is weak when applied to pairs of proteins[16] and that domain assignments are not sufficient to resolve function.[50] Therefore, the main challenge for the alignment-based methods was to devise ways of combining multiple hits or identified domains into a single prediction score. A number of methods exploited data beyond sequence similarity, e.g. types of evolutionary relationships, protein structure, protein-protein interactions, or gene expression data. The challenge for these methods was finding ways to integrate disparate data sources and properly handle incomplete and noisy data. For example, the protein-protein interaction network for yeast is nearly complete (although noisy), while the sets of available interactions for *A. thaliana* and *X. laevis* are rather sparse (but less noisy, given a

smaller fraction of high-throughput data). Finally, some methods used literature mining, which could also be related to the task of retrieving the correct function rather than predicting it from the set of textual descriptions about a protein. As information retrieval is still a challenging research problem, it was useful to evaluate performance accuracy of the methods that exploited literature search.

On the computational side, most methods exploited machine learning principles, i.e. they typically found combinations of sequence-based or other features that correlated with a specific function in a training set of experimentally annotated proteins. While these methods automate the task of learning and inference, they also require experience in selecting classification models (e.g. a support vector machine), learning parameters, features, or the training data that would result in good performance. In addition, the sets of rules according to which these methods score new proteins may be difficult to interpret. Despite the added layer of complexity, machine learning generally played a positive role in increasing prediction accuracy. Thus, it may be expected that top-performing methods in the future will be based on well-founded principles of statistical learning and inference.

With few exceptions the same methods that performed well for the Molecular Function category also performed well in the Biological Process category; however, their overall performance in the latter category was inferior. We believe that this is because homologs may perform their biochemical roles in different pathways, and prediction methods are less able to discern those differences at this time. Because sequence similarity is less predictive of the biological roles of proteins, a key to improving the prediction of a protein's biological function will depend on our ability to generate better-quality systems data and to develop computational tools that exploit them.

Evaluation metrics

The choice of evaluation metrics was another interesting aspect of the experiment. We decided to use simple and easily interpretable metrics (Online Methods), although simple measures based on precision and recall have limitations in this domain. First, such metrics are sensitive to problems related to the non-uniform distribution of proteins over GO terms due to giving all terms equal weight. Second, proteins are weighted equally regardless of the depth of their experimental annotation, i.e. a correct prediction on a protein annotated with a shallow term (and its ancestors) is considered as good as a correct prediction on a protein annotated with a deep term. Third, a method that only reports high confidence deep annotations for a small number of proteins will be penalized (in terms of recall) compared to a method that annotates all proteins with frequently occurring general terms. Finally, in some cases, it is not clear whether to consider a prediction correct or erroneous; with our current approach, we consider only the experimental annotation and more general predictions to be correct. As such, correct and highly specific predictions will be penalized if the protein has been experimentally annotated only in a more generic way. For those reasons, we encourage the development of a diverse set of metrics to understand better the strengths and weaknesses of function prediction in different application contexts.

Summary

The CAFA experiment was designed to enable the community to periodically reassess the performance of computational methods as experimental evidence further accumulates. In addition, the large set of targets released to the community provided us with prediction scores for most proteins across multiple methods. If the experiment is repeated, we expect to be able to evaluate future methods against those that deposited predictions in the first CAFA experiment and therefore monitor progress in the field over time.

While the CAFA experiment has certainly seen positive outcomes, it is also clear that there is significant room for the improvement of protein function prediction. In the Molecular Function category, performance may be considered accurate. However, in the Biological Process category, the overall performance of the top scoring methods was below our expectations. This was true for any subset of targets. Another area in need of improvement is the availability of tools that can easily be used by experimental scientists and that could be maintained and upgraded on a regular basis. As the community moves beyond the initial algorithm development stage, there is a need to provide standalone tools (similar to the BLAST package) capable of predicting protein function at several different levels.

Given its significance, intellectual challenge, and the growing need for accurate functional annotations, protein function prediction is likely to remain an active and growing research field. As the quality of data improves and the number of experimentally annotated proteins grows, we expect that computational prediction will become more accurate. Based on the CAFA experiment, it seems that the most powerful methods will be those that will devise principled ways to integrate a variety of experimental evidence and weight different data appropriately and separately for each functional term. Novel ideas and approaches are necessary as well.

Online Methods

Experiment design

The CAFA experiment was conceived in the fall of 2009. The Organizing, Steering and Assessment Committees were designated by March 2010. During the same period a feasibility study was conducted to determine the rate at which experimental annotations accumulated in Swiss-Prot between 2007 and 2010. We concluded that a period of six months or more would result in annotations of at least 300-500 proteins, which would be sufficient for statistically reliable comparisons between algorithms. The experiment was announced in July 2010 and subsequently heavily advertised. The set of targets was announced on September 15, 2010 with a prediction submission deadline of January 18, 2011 (Fig. 1).

Predictors were asked to submit predictions for each target along with scores ranging between 0 and 1, which would indicate the strength of the prediction (ideally, posterior probabilities). To reduce the amount of data to be submitted, no more than 1,000 term annotations were allowed for each target. Prediction algorithms were also associated with keywords from a pre-determined set, which were used to provide insight into the types of

approaches that performed well. A list of all participating teams, principal investigators, and methods is provided in Supplementary Table 3.

Initial comparative evaluation of models was conducted in July 2011 during the Automated Function Prediction (AFP) Special Interest Group (SIG) meeting associated with ISMB 2011 conference. This study provides the analysis on a set of targets from the Swiss-Prot database from December 14, 2011.

Target proteins

A set of 48,298 target amino acid sequences was announced in September 2010. Because our feasibility study showed that only a handful of species were steadily accumulating experimental annotations, target proteins were predominantly selected from those species. The targets contained all the sequences in Swiss-Prot from seven eukaryotic and eleven prokaryotic species that were not associated with any experimental GO terms. A protein was considered experimentally annotated if it was associated with GO terms having EXP, IDA, IMP, IGI, IEP, TAS, or IC evidence codes. An additional set of targets was announced consisting of 1,301 enzymes from multiple species and metagenomic studies that were the focus of the Enzyme Function Initiative project.[51]

January 18, 2011 was set as the deadline for the submission of function predictions. To exclude targets that had accumulated annotations prior to the submission deadline, annotated proteins were obtained from the January version of Swiss-Prot, GO,[35] and UniProt-GOA[52] databases. We refer to those sets of proteins as Swiss-Prot(t_0), GO(t_0) and GOA(t_0), respectively.

The evaluation set of target proteins was later determined by downloading a newer version of the Swiss-Prot database, denoted as Swiss-Prot(t). The set of target proteins for the CAFA experiment was then selected using the following scheme:

$$Targets(t) = SwissProt(t) - SwissProt(t_0) - GO(t_0) - GOA(t_0)$$

Note that this experiment was designed to allow for reassessing algorithm performance at some later point in time.

Evaluation metrics

Algorithms were evaluated in two scenarios: (i) protein-centric and (ii) term-centric. These two types of evaluations were chosen to address the following related questions: (i) what is the function of a particular protein and (ii) what are the proteins associated with a particular functional term.

1) Protein-centric metrics—The main evaluation metric in CAFA was the precision/recall curve. For a given target protein i and some decision threshold $t \in [0,1]$, the precision and recall were calculated as

$$pr_i(t) = \frac{\sum_f I(f \in P_i(t) \hat{} f \in T_i)}{\sum_f I(f \in P_i(t))}, \; and$$
$$rc_i(t) = \frac{\sum_f I(f \in P_i(t) \hat{} f \in T_i)}{\sum_f I(f \in T_i)}$$

where f is a functional term in the ontology, T_i is a set of experimentally determined (true) nodes for protein i, and $P_i(t)$ is a set of predicted terms for protein i with score greater than or equal to t. Note that f ranges over the entire ontology (separately for Molecular Function and Biological Process), excluding the root. Function $I(\cdot)$ is the standard indicator function. For a fixed threshold t, a point in the precision/recall space is then created by averaging precision and recall across targets. Precision at threshold t is calculated as

$$pr(t) \frac{1}{m(t)} \cdot \sum_{i=1}^{m(t)} pr_i(t)$$

where $m(t)$ is the number of proteins on which at least one prediction was made above threshold t. On the other hand, recall is calculated over all n proteins in a target set, i.e.

$$rc(t) = \frac{1}{n} \cdot \sum_{i=1}^{n} rc_i(t)$$

regardless of the prediction threshold. The maximum ratio between $m(t)$ and n (over all thresholds t) is referred to as the prediction coverage. If a particular algorithm only outputs a fixed score (e.g. 1), its performance will be described by a single point in the precision/recall space, instead of by a curve.

For submissions with unpropagated functional annotations, the organizers recursively propagated all scores towards the root of the ontology such that each parent term received the highest score among its children. The annotations were propagated regardless of the type of relationship between terms. We note that it may be useful to associate different weights with different ontological terms and therefore reward algorithms that are better at predicting more difficult or less frequent terms. However, for simplicity, in our main evaluation, each term was associated with an equal weight of 1 (weighted precision/recall curves are shown in Supplementary Fig. 8).

The main appeal of the precision/recall evaluation stems from its interpretability; i.e. if for a particular threshold a method has precision of 0.7 at recall of 0.5, this indicates that on average 70% of the predicted terms will be correct and that about 50% of the true annotations will be revealed for a previously unseen protein. On the other hand, a limitation of this evaluation method is that the terms are not independent due to ontological relationships, and that the unequal level of specificity of functional terms at the same depth in the ontology was not taken into account.

To provide a single number for comparisons between methods, we calculated the F-measure (a harmonic mean between precision and recall) for each threshold and calculated its maximum value over all thresholds. More specifically, we used

$$F_{max} = \max_t \left\{ \frac{2 \cdot pr(t) \cdot rc(t)}{pr(t) + rc(t)} \right\}.$$

2) Term-centric metrics—For each functional term *f*, we calculated the area under the ROC curve (AUC) using a sliding threshold approach. The ROC curve is a plot of sensitivity (or recall) for a given false positive rate (or 1 – specificity). The sensitivity and specificity for a particular functional term *f* and threshold *t* were calculated as

$$sn_f(t) = \frac{\sum_i I(f \in P_i(t) \wedge f \in T_i)}{\sum_i I(f \in T_i)}, \text{ and}$$
$$sp_f(t) = \frac{\sum_i I(f \notin P_i(t) \wedge f \notin T_i)}{\sum_i I(f \notin T_i)}$$

where $P_i(t)$ is the set of predicted terms for protein *i* with score greater than or equal to threshold *t*, and T_i is the set of true terms for protein *i*. Once the sensitivity and specificity for a particular functional term were determined over all proteins for different values of the prediction threshold, the AUC was calculated using the trapezoid rule. The AUC has a useful probabilistic interpretation: given a randomly selected protein associated with functional term *f* and a randomly selected protein not associated with *f*, the AUC is the probability that the former protein will receive a higher score than the latter protein.[53]

Baseline methods

In addition to the methods implemented by the community, we used two additional methods as baselines. The first such method is based on BLAST[11] hits to the database of proteins with experimentally annotated functions (roughly 37,000 proteins). The score for a particular term was calculated as the maximum sequence identity between the target protein and any protein experimentally annotated with that term. More specifically, if a particular protein was hit with the local sequence identity 75%, all its functional terms were transferred to the target sequence with the score of 0.75. If a term was hit with multiple sequence identity scores, the highest one was retained. BLAST was selected as a baseline method because of its ubiquitous use. We note that the same method was tested using the BLAST bit scores, which resulted in slightly better performance. In addition to BLAST, we also tested PSI-BLAST[11] where the profiles were created using the most recent nr database and −j 3 −h 0.0001 parameters. These profiles were then searched against a database of experimentally annotated proteins with E-values used to rank the hits. The second baseline method, referred to as Naïve, used the prior probability of each term in the database of experimentally annotated proteins as the prediction score for that term. If a term “protein binding” occurs with relative frequency 0.25, each target protein was associated with score 0.25 for that term. Thus, the Naïve method assigned the same predictions to all targets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Predrag Radivojac[1], Wyatt T Clark[1], Tal Ronnen Oron[2], Alexandra M Schnoes[3], Tobias Wittkop[2], Artem Sokolov[4,5], Kiley Graim[4], Christopher Funk[6], Karin Verspoor[6,7], Asa Ben-Hur[4], Gaurav Pandey[8,9], Jeffrey M Yunes[10], Ameet S Talwalkar[11], Susanna Repo[8,12], Michael L Souza[13], Damiano Piovesan[14], Rita Casadio[14], Zheng Wang[15], Jianlin Cheng[15], Hai Fang[16], Julian Gough[16], Patrik Koskinen[17], Petri Törönen[17], Jussi Nokso-Koivisto[17], Liisa Holm[17], Domenico Cozzetto[18], Daniel W A Buchan[18], Kevin Bryson[18], David T Jones[18], Bhakti Limaye[19], Harshal Inamdar[19], Avik Datta[19], Sunitha K Manjari[19], Rajendra Joshi[19], Meghana Chitale[20], Daisuke Kihara[20,21], Andreas M Lisewski[22], Serkan Erdin[22], Eric Venner[22], Olivier Lichtarge[22], Robert Rentzsch[23], Haixuan Yang[24], Alfonso E Romero[24], Prajwal Bhat[24], Alberto Paccanaro[24], Tobias Hamp[25], Rebecca Kassner[25], Stefan Seemayer[25], Esmeralda Vicedo[25], Christian Schaefer[25], Dominik Achten[25], Florian Auer[25], Ariane Böhm[25], Tatjana Braun[25], Maximilian Hecht[25], Mark Heron[25], Peter Hönigschmid[25], Thomas Hopf[25], Stefanie Kaufmann[25], Michael Kiening[25], Denis Krompass[25], Cedric Landerer[25], Yannick Mahlich[25], Manfred Roos[25], Jari Björne[26], Tapio Salakoski[26], Andrew Wong[27], Hagit Shatkay[27,28], Fanny Gatzmann[29], Ingolf Sommer[29], Mark N Wass[30,31], Michael J E Sternberg[30], Nives Škunca[32], Fran Supek[32], Matko Bošnjak[32], Panče Panov[33], Sašo Džeroski[33], Tomislav Šmuc[32], Yiannis A I Kourmpetis[34,35], Aalt D J van Dijk[36], Cajo J F ter Braak[34], Yuanpeng Zhou[37], Qingtian Gong[37], Xinran Dong[37], Weidong Tian[37], Marco Falda[38], Paolo Fontana[39], Enrico Lavezzo[38], Barbara Di Camillo[40], Stefano Toppo[38], Liang Lan[41], Nemanja Djuric[41], Yuhong Guo[41], Slobodan Vucetic[41], Amos Bairoch[42], Michal Linial[43], Patricia C Babbitt[3], Steven E Brenner[8], Christine Orengo[23], Burkhard Rost[25], Sean D Mooney[2], and Iddo Friedberg[44,45]

Affiliations

[1]School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA [2]Buck Institute for Research on Aging, Novato, California, USA [3]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California, USA [4]Department of Computer Science, Colorado State University, Fort Collins, Colorado, USA [5]Department of Biomolecular Engineering, University of California, Santa Cruz, California, USA [6]Computational Bioscience Program, University of Colorado School of Medicine, Aurora, Colorado, USA [7]National ICT Australia, Victoria Research Lab, Melbourne, Australia [8]Department of Plant and Microbial Biology, University of California, Berkeley, California, USA [9]Mount Sinai School of Medicine, New York, New York, USA [10]Joint Graduate Group in Bioengineering, University of California, Berkeley, California, USA [11]Department of Electrical Engineering and Computer Science, University of California, Berkeley, California, USA [12]European Molecular Biology Laboratory, European Bioinformatics

Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK [13]Biophysics Graduate Program, University of California, Berkeley, California, USA [14]Department of Biology, University of Bologna, Bologna, Italy [15]Department of Computer Science, University of Missouri, Columbia, Missouri, USA [16]Department of Computer Science, University of Bristol, Bristol, UK [17]Department of Biological and Environmental Sciences, Viikki Biocentre, University of Helsinki, Helsinki, Finland [18]Department of Computer Science, University College London, London, UK [19]Bioinformatics Group, Centre for Development of Advanced Computing, Pune University Campus, Pune, India [20]Department of Computer Science, Purdue University, West Lafayette, Indiana, USA [21]Department of Computer Science, Purdue University, West Lafayette, Indiana, USA [22]Department of Molecular and Human Genetics, Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, Texas, USA [23]University College London, Institute for Structural and Molecular Biology, London, UK [24]Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway, University of London, Egham, UK [25]Technische Universität München, Bioinformatik - i12, Informatik, Garching, Germany [26]Department of Information Technology, University of Turku, Turku Centre for Computer Science, Turku, Finland [27]School of Computing, Queen's University, Kingston, Ontario, Canada [28]Department of Computer and Information Sciences, University of Delaware, Newark, Delaware, USA [29]Max-Planck-Institute for Informatics, Saarbrücken, Germany [30]Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College, London, UK [31]Structural Computational Biology Group, Spanish National Cancer Research Centre, Madrid, Spain [32]Division of Electronics, Ruder Boškovic Institute, Zagreb, Croatia [33]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia [34]Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands [35]Bioinformatics Systems, Nestlé Institute of Health Sciences, Lausanne, Switzerland [36]Applied Bioinformatics, Plant Research International, Wageningen, The Netherlands [37]Institute of Biostatistics, School of Life Sciences, Fudan University, Shanghai, China [38]Department of Molecular Medicine, University of Padova, Padova, Italy [39]Istituto Agrario San Michele all'Adige Research and Innovation Centre, Trento, Italy [40]Department of Information Engineering, University of Padova, Padova, Italy [41]Department of Computer and Information Sciences, Temple University, Philadelphia, Pennsylvania, USA [42]Swiss Institute of Bioinformatics, Geneva, Switzerland and Department of Human Protein Sciences, University of Geneva, Geneva, Switzerland [43]Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Israel [44]Department of Microbiology, Miami University, Oxford, Ohio, USA [45]Department of Computer Science and Software Engineering, Miami University, Oxford, Ohio, USA

Acknowledgments

We gratefully acknowledge Inbal Landsberg-Halperin for coining the term "CAFA", Tracey Theriault for the initial graphical design of Figure 1, Gadi Schuster for illuminating discussions on hPNPase, as well as Andrea Facchinetti, Riccardo Velasco, Elisa Cilia, David A. Lee, Pankaj Vats, Ruma Banerjee, and Avinash Bayaskar for participation

in various individual projects. Finally, we thank the three anonymous reviewers for their constructive comments and criticisms that improved the presentation and quality of this study.

Funding

The Automated Function Prediction Special Interest Group meeting at the ISMB 2011 conference was supported by the National Institutes of Health grant R13 HG006079-01A1 (PR) and Office of Science (BER), U.S. Department of Energy grant DE-SC0006807TDD (IF). Individual projects were partially supported by the following awards: NSF DBI-0644017 (PR), Marie Curie International Outgoing Fellowship PIOF-QA-2009-237751 (SR), NSF ABI-0965768 (AB-H), PRIN 2009 project 009WXT45Y Italian Ministry for University and Research MIUR (RC), NIH GM093123 (JC), FP7 "Infrastructures" project TransPLANT Award 283496 (A-J vD), BBSRC grant BB/G022771/1 (JG), UK Biotechnology and Biological Sciences Research Council (DTJ), Marie Curie Intra European Fellowship Award PIEF-GA-2009-237292 (DTJ), Bioinformatics Division, Department of Information Technology, MCIT, Govt. of India (RJ), NIH GM075004 (DK), NIH GM097528 (DK), NSF DMS0800568 (DK), NIH GM079656 (OL), NIH GM066099 (OL), NSF CCF 0905536 (OL), NSF DBI 1062455 (OL), EU, BBSRC and NIH Awards (CO), NSERC Discovery Award #298292-2009 (HS), Discovery Accelerator Award #380478-2009 (HS), CFI New Opportunities Award 10437 (HS), and Ontario's Early Researcher Award #ER07-04-085 (HS), Biotechnology and Biological Sciences Research Council Award BB/F020481/1 (MW and MJS), Netherlands Genomics Initiative (YK, CtB), NIH LM00945102, NSF DBI-0965616, NSF DBI-0965768, NICTA (KV), NIH R01 GM071749 (SEB), DOE BER KP110201 (SEB), Alexander von Humboldt-Foundation (BR), NIH LM009722 (SDM), NIH HG004028 (SDM), NSF ABI-1146960 (IF).

References

1. Liolios K, et al. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 2010; 38:D346–354. [PubMed: 19914934]
2. Bork P, et al. Predicting function: from genes to genomes and back. J Mol Biol. 1998; 283:707–725. [PubMed: 9790834]
3. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci. 2003; 60:2637–2650. [PubMed: 14685688]
4. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. Curr Opin Struct Biol. 2005; 15:275–284. [PubMed: 15963890]
5. Friedberg I. Automated protein function prediction--the genomic challenge. Brief Bioinform. 2006; 7:225–242. [PubMed: 16772267]
6. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007; 3:88. [PubMed: 17353930]
7. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol. 2007; 8:995–1005. [PubMed: 18037900]
8. Punta M, Ofran Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput Biol. 2008; 4:e1000160. [PubMed: 18974821]
9. Rentzsch R, Orengo CA. Protein function prediction--the power of multiplicity. Trends Biotechnol. 2009; 27:210–219. [PubMed: 19251332]
10. Xin F, Radivojac P. Computational methods for identification of functional residues in protein structures. Curr Protein Pept Sci. 2011; 12:456–469. [PubMed: 21787297]
11. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]
12. Jensen LJ, et al. Prediction of human protein function from post-translational modifications and localization features. J Mol Biol. 2002; 319:1257–1265. [PubMed: 12079362]
13. Wass MN, Sternberg MJ. ConFunc--functional annotation in the twilight zone. Bioinformatics. 2008; 24:798–806. [PubMed: 18263643]
14. 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. [PubMed: 15550167]

15. 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. [PubMed: 16672240]
16. Clark WT, Radivojac P. Analysis of protein function and its prediction from amino acid sequence. Proteins. 2011; 79:2086–2096. [PubMed: 21671271]
17. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA. 1999; 96:4285–4288. [PubMed: 10200254]
18. Marcotte EM, et al. Detecting protein function and protein-protein interactions from genome sequences. Science. 1999; 285:751–753. [PubMed: 10427000]
19. Enault F, Suhre K, Claverie JM. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics. 2005; 6:247. [PubMed: 16221304]
20. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol. 2005; 1:e45. [PubMed: 16217548]
21. Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief Bioinform. 2011; 12:449–462. [PubMed: 21873635]
22. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. J Comput Biol. 2003; 10:947–960. [PubMed: 14980019]
23. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics. 2003; 19(Suppl 1):i197–204. [PubMed: 12855458]
24. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol. 2003; 21:697–700. [PubMed: 12740586]
25. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics. 2005; 21(Suppl 1):i302–310. [PubMed: 15961472]
26. Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci USA. 2004; 101:14754–14759. [PubMed: 15456910]
27. Pal D, Eisenberg D. Inference of protein function from protein structure. Structure. 2005; 13:121–130. [PubMed: 15642267]
28. Laskowski RA, Watson JD, Thornton JM. Protein function prediction using local 3D templates. J Mol Biol. 2005; 351:614–626. [PubMed: 16019027]
29. Huttenhower C, Hibbs M, Myers C, Troyanskaya OG. A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics. 2006; 22:2890–2897. [PubMed: 17005538]
30. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci USA. 2003; 100:8348–8353. [PubMed: 12826619]
31. Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. Science. 2004; 306:1555–1558. [PubMed: 15567862]
32. Costello JC, et al. Gene networks in Drosophila melanogaster: integrating experimental data to predict gene function. Genome Biol. 2009; 10:R97. [PubMed: 19758432]
33. Kourmpetis YA, van Dijk AD, Bink MC, van Ham RC, ter Braak CJ. Bayesian Markov Random Field analysis for protein function prediction based on network data. PLoS One. 2010; 5:e9293. [PubMed: 20195360]
34. Sokolov A, Ben-Hur A. Hierarchical classification of gene ontology terms using the GOstruct method. J Bioinform Comput Biol. 2010; 8:357–376. [PubMed: 20401950]
35. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]
36. Bairoch A, et al. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2005; 33(Database Issue):D154–159. [PubMed: 15608167]

37. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009; 5:e1000605. [PubMed: 20011109]
38. Punta M, et al. The Pfam protein families database. Nucleic Acids Res. 2012; 40:D290–301. [PubMed: 22127870]
39. Wang G, et al. PNPASE regulates RNA import into mitochondria. Cell. 2010; 142:456–467. [PubMed: 20691904]
40. Sarkar D, et al. Down-regulation of Myc as a potential target for growth arrest induced by human polynucleotide phosphorylase (hPNPaseold-35) in human melanoma cells. J Biol Chem. 2003; 278:24542–24551. [PubMed: 12721301]
41. Wu J, Li Z. Human polynucleotide phosphorylase reduces oxidative RNA damage and protects HeLa cell against oxidative stress. Biochem Biophys Res Commun. 2008; 372:288–292. [PubMed: 18501193]
42. Wang DD, Shu Z, Lieser SA, Chen PL, Lee WH. Human mitochondrial SUV3 and polynucleotide phosphorylase form a 330-kDa heteropentamer to cooperatively degrade double-stranded RNA with a 3'-to-5' directionality. J Biol Chem. 2009; 284:20812–20821. [PubMed: 19509288]
43. Portnoy V, Palnizky G, Yehudai-Resheff S, Glaser F, Schuster G. Analysis of the human polynucleotide phosphorylase (PNPase) reveals differences in RNA binding and response to phosphate compared to its bacterial and chloroplast counterparts. RNA. 2008; 14:297–309. [PubMed: 18083836]
44. Jeffery CJ. Moonlighting proteins. Trends Biochem Sci. 1999; 24:8–11. [PubMed: 10087914]
45. Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem. 2010; 79:471–505. [PubMed: 20235827]
46. Brenner SE. Errors in genome annotation. Trends Genet. 1999; 15:132–133. [PubMed: 10203816]
47. Doolittle RF. Of URFS and ORFS: A Primer on How to Analyze Derived Amino Acid Sequences. University Science Books. 1986
48. Addou S, Rentzsch R, Lee D, Orengo CA. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. J Mol Biol. 2009; 387:416–430. [PubMed: 19135455]
49. Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol. 2011; 7:e1002073. [PubMed: 21695233]
50. Brown SD, Gerlt JA, Seffernick JL, Babbitt PC. A gold standard set of mechanistically diverse enzyme superfamilies. Genome Biol. 2006; 7:R8. [PubMed: 16507141]
51. Gerlt JA, et al. The Enzyme Function Initiative. Biochemistry. 2011; 50:9950–9962. [PubMed: 21999478]
52. Barrell D, et al. The GOA database in 2009--an integrated Gene Ontology Annotation resource. Nucleic Acids Res. 2009; 37:D396–403. [PubMed: 18957448]
53. Hanley J, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]
54. Cozzeto D, Buchan DWA, Bryson K, Jones DT. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. BMC Bioinformatics. 2013
55. Falda M, et al. Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. BMC Bioinformatics. 2012; 13:S14. [PubMed: 22536960]
56. Chitale M, Hawkins T, Park C, Kihara D. ESG: extended similarity group method for automated protein function prediction. Bioinformatics. 2009; 25:1739–1745. [PubMed: 19435743]
57. Bartoli L, et al. The Bologna Annotation Resource: a non hierarchical method for the functional and structural annotation of protein sequences relying on a comparative large-scale genome analysis. J Proteome Res. 2009; 8:4362–4371. [PubMed: 19552451]
58. Piovesan D, et al. BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences. Nucleic Acids Res. 2011; 39:W197–202. [PubMed: 21622657]

59. Wang Z, Cao R, Cheng J. Three-level prediction of protein function by combining profile-sequence search, profile-profile search, and domain co-occurrence networks. BMC Bioinformatics. 2013
60. Rentzsch R, Orengo CA. Protein function prediction using domain families. BMC Bioinformatics. 2013
61. Erdin S, Venner E, Lisevski MA, Lichtarge O. Function prediction from networks of local evolutionary similarity in protein structure. BMC Bioinformatics. 2013
62. Engelhardt BE, Jordan MI, Srouji JR, Brenner SE. Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res. 2011; 21:1969–1980. [PubMed: 21784873]
63. Fang H, Gough J. dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res. 2012
64. Fang H, Gough J. dcGO: a domain-centric gene ontology predictor for functional genomics. BMC Bioinformatics. 2013
65. Lan L, Djuric N, Guo Y, Vucetic S. MS-kNN: Protein function prediction by integrating multiple data sources. BMC Bioinformatics. 2013
66. Hamp T, et al. Homology-based inference sets the bar high for protein function prediction. BMC Bioinformatics. 2013
67. Kourmpetis YA, van Dijk AD, van Ham RC, ter Braak CJ. Genome-wide computational function prediction of Arabidopsis proteins by integration of multiple data sources. Plant Physiol. 2011; 155:271–281. [PubMed: 21098674]
68. Wong A, Shatkay H. Protein function prediction using text-based features extracted from the biomedical literature: the CAFA challenge. BMC Bioinformatics. 2013
69. Lobley A, Swindells MB, Orengo CA, Jones DT. Inferring function using patterns of native disorder in proteins. PLoS Comput Biol. 2007; 3:e162. [PubMed: 17722973]
70. Lobley AE, Nugent T, Orengo CA, Jones DT. FFPred: an integrated feature-based function prediction server for vertebrate proteomes. Nucleic Acids Res. 2008; 36:W297–302. [PubMed: 18463141]
71. Hawkins T, Chitale M, Luban S, Kihara D. PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins. 2009; 74:566–582. [PubMed: 18655063]
72. de Lima Morais DA, et al. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res. 2011; 39:D427–434. [PubMed: 21062816]
73. Weinhold N, Sander O, Domingues FS, Lengauer T, Sommer I. Local function conservation in sequence and structure space. PLoS Comput Biol. 2008; 4:e1000105. [PubMed: 18604264]
74. Fontana P, Cestaro A, Velasco R, Formentin E, Toppo S. Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology. PLoS One. 2009; 4:e4619. [PubMed: 19247487]
75. Schietgat L, et al. Predicting gene function using hierarchical multi-label decision tree ensembles. BMC Bioinformatics. 2010; 11:2. [PubMed: 20044933]
76. Sokolov A, Funk C, Graim K, Verspoor K, Ben-Hur A. Combining heterogeneous data sources for accurate functional annotation of proteins. BMC Bioinformatics. 2013
77. Toronen P, Ojala PJ, Marttinen P, Holm L. Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function. BMC Bioinformatics. 2009; 10:307. [PubMed: 19775443]
78. Conesa A, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005; 21:3674–3676. [PubMed: 16081474]
79. Arakaki AK, Huang Y, Skolnick J. EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. BMC Bioinformatics. 2009; 10:107. [PubMed: 19361344]
80. Camon EB, et al. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics. 2005; 6(Suppl 1):S17. [PubMed: 15960829]
81. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol. 1996; 257:342–358. [PubMed: 8609628]

82. Erdin S, Lisewski AM, Lichtarge O. Protein function prediction: towards integration of similarity metrics. Curr Opin Struct Biol. 2011; 21:180–188. [PubMed: 21353529]
83. Kristensen DM, et al. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. BMC Bioinformatics. 2008; 9:17. [PubMed: 18190718]
84. Ward RM, et al. De-orphaning the structural proteome through reciprocal comparison of evolutionarily important structural features. PLoS One. 2008; 3:e2136. [PubMed: 18461181]
85. Erdin S, Ward RM, Venner E, Lichtarge O. Evolutionary trace annotation of protein function in the structural proteome. J Mol Biol. 2010; 396:1451–1473. [PubMed: 20036248]
86. Venner E, et al. Accurate protein structure annotation through competitive diffusion of enzymatic functions over a network of local evolutionary similarities. PLoS One. 2010; 5:e14286. [PubMed: 21179190]
87. von Mering C, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005; 33:D433–437. [PubMed: 15608232]
88. Cuff AL, et al. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. Nucleic Acids Res. 2011; 39:D420–426. [PubMed: 21097779]
89. Lee DA, Rentzsch R, Orengo C. GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. Nucleic Acids Res. 2010; 38:720–737. [PubMed: 19923231]
90. Lees J, Yeats C, Redfern O, Clegg A, Orengo C. Gene3D: merging structure and function for a Thousand genomes. Nucleic Acids Res. 2010; 38:D296–300. [PubMed: 19906693]
91. Zhou D, Bousquet O, Navin Lal T, Weston J, Schölkopf B. Learning with local and global consistency. Advances in Neural Information Processing Systems. 2004:321–328.
92. Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research. 2005; 8:1453–1484.
93. Bjorne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T. Complex event extraction at PubMed scale. Bioinformatics. 2010; 26:i382–390. [PubMed: 20529932]
94. Brady S, Shatkay H. EpiLoc: a (working) text-based system for predicting protein subcellular location. Pac Symp Biocomput. 2008:604–615. [PubMed: 18229719]
95. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009; 23:205–211. [PubMed: 20180275]
96. Kuzniar A, et al. ProGMap: an integrated annotation resource for protein orthology. Nucleic Acids Res. 2009; 37:W428–434. [PubMed: 19494185]
97. Harris MA, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004; 32:D258–261. [PubMed: 14681407]
98. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007; 23:1282–1288. [PubMed: 17379688]
99. Hubbard T, et al. The Ensembl genome database project. Nucleic Acids Res. 2002; 30:38–41. [PubMed: 11752248]
100. Wu CH, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 2006; 34:D187–191. [PubMed: 16381842]
101. Apweiler R, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 2001; 29:37–40. [PubMed: 11125043]
102. Bateman A, et al. The Pfam protein families database. Nucleic Acids Res. 2004; 32:138–141.
103. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003; 31:371–373. [PubMed: 12520025]
104. Pandit SB, et al. SUPFAM--a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. Nucleic Acids Res. 2002; 30:289–293. [PubMed: 11752317]
105. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. 2000; 28:231–234. [PubMed: 10592234]
106. Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. Nucleic Acids Res. 1999; 27:215–219. [PubMed: 9847184]

107. Attwood TK. The PRINTS database: a resource for identification of protein families. Brief Bioinform. 2002; 3:252–263. [PubMed: 12230034]

108. Thomas PD, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003; 13:2129–2141. [PubMed: 12952881]

109. Lima T, et al. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Res. 2009; 37:D471–478. [PubMed: 18849571]

110. Buchan DW, et al. Gene3D: structural assignments for the biologist and bioinformaticist alike. Nucleic Acids Res. 2003; 31:469–473. [PubMed: 12520054]

111. Wishart DS, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006; 34:D668–672. [PubMed: 16381955]

112. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics. 2006; 7:474. [PubMed: 17064419]

113. John B, et al. Human MicroRNA targets. PLoS Biol. 2004; 2:e363. [PubMed: 15502875]

114. Karolchik D, et al. The UCSC Genome Browser Database. Nucleic Acids Res. 2003; 31:51–54. [PubMed: 12519945]

115. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004; 306:636–640. [PubMed: 15499007]

116. Matys V, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006; 34:D108–110. [PubMed: 16381825]

117. Tian W, et al. Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. Genome Biol. 2008; 9(Suppl 1):S7. [PubMed: 18613951]

118. Stark C, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006; 34:D535–539. [PubMed: 16381927]

119. Zanzoni A, et al. MINT: a Molecular INTeraction database. FEBS Lett. 2002; 513:135–140. [PubMed: 11911893]

120. Hermjakob H, et al. IntAct: an open source molecular interaction database. Nucleic Acids Res. 2004; 32:D452–455. [PubMed: 14681455]

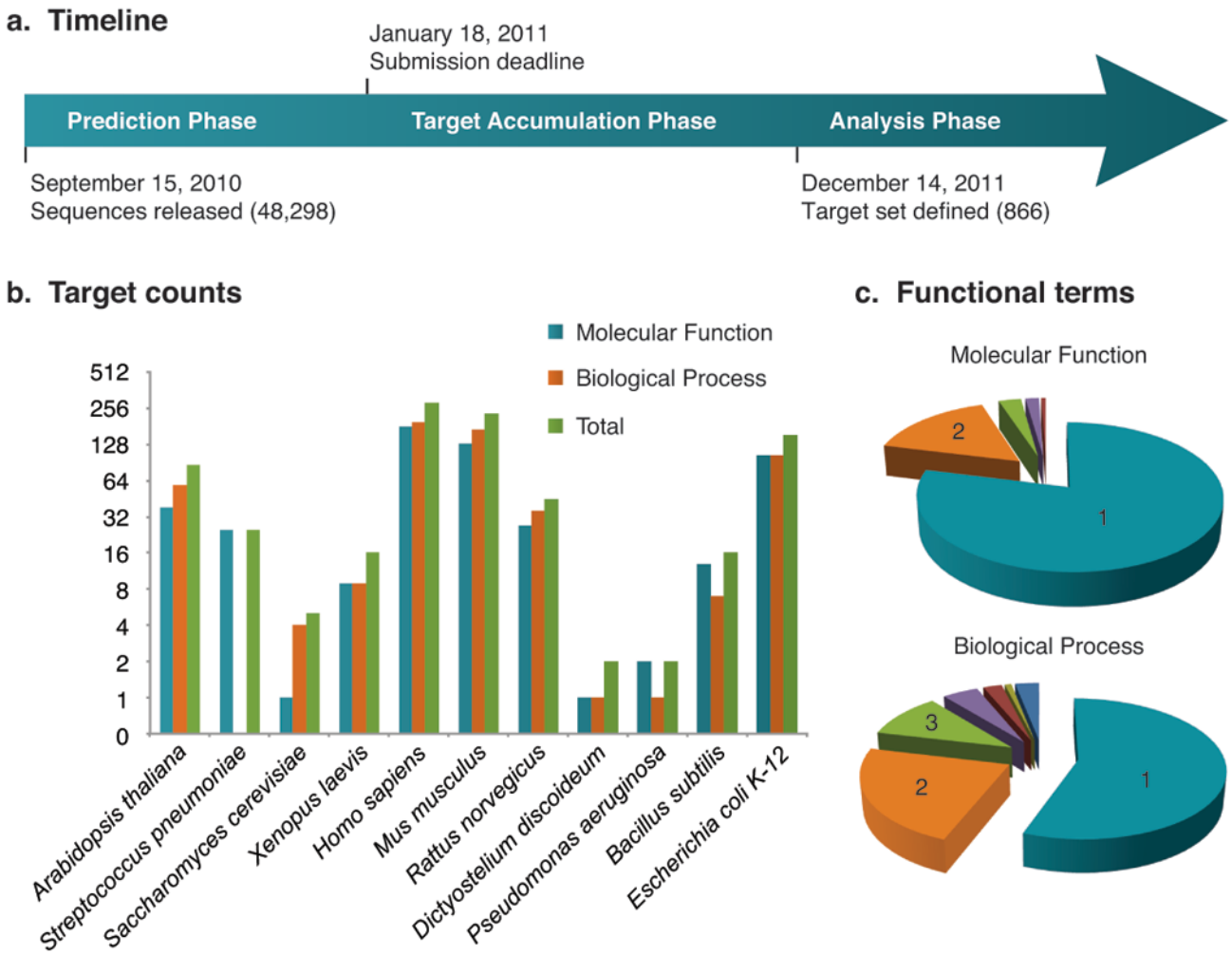

Figure 1.

Experiment timeline and target analysis. (A) Timeline for the CAFA experiment. (B) The number of target sequences per organism. The graphs show the number of target sequences for each of the ontologies (Molecular Function and Biological Process), as well as the total number of targets, obtained as a union between sequences in the two ontologies. Out of 866 proteins, 531 had Molecular Function annotations, and 587 had Biological Process annotations. (C) The distribution of target sequences in each ontology according to the number of leaf terms available for each protein sequence. A term is considered a leaf term for a particular target, if no other GO term associated with that sequence is its descendant.

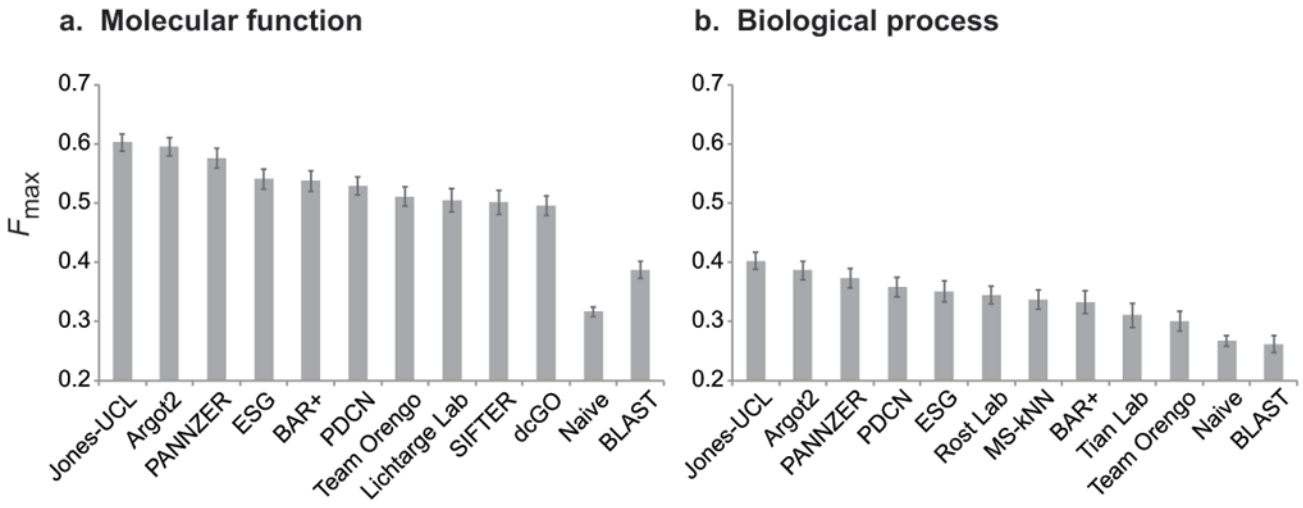

Figure 2.

Overall performance evaluation. The maximum F-measure for the top-performing methods for (A) Molecular Function ontology and (B) Biological Process ontology. All panels show the top ten participating methods in each category, as well as the BLAST and Naïve baseline methods. Note that 33 models outperformed BLAST in the Molecular Function category, whereas 26 models outperformed BLAST in the Biological Process category (cut-off scores below which methods were excluded from the panels were 0.468 and 0.300, for the Molecular Function and Biological Process categories, respectively). In the Molecular Function category, proteins with a single leaf term "protein binding" were excluded from the analysis because the protein binding term was not considered informative (results that include those proteins are presented in Supplementary Fig. 3). A perfect predictor would be characterized with $F_{max} = 1$. Confidence intervals (95%) were determined using bootstrapping with $n = 10{,}000$ iterations on the set of target sequences. In cases where a Principal Investigator (PI) participated with multiple teams, only the results of the best-scoring method are presented.

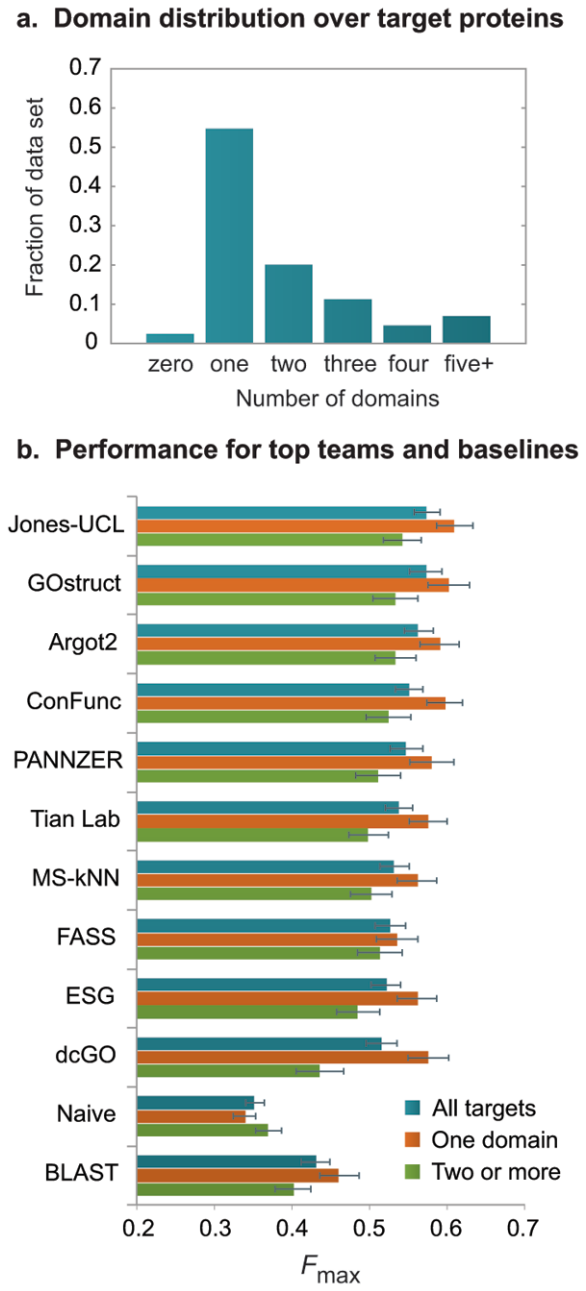

Figure 3.

Domain analysis and performance evaluation for single- vs. multi-domain eukaryotic targets. (A) Distribution of target proteins with respect to the number of Pfam domains they contain. (B) Performance evaluation in the Molecular Function category. Each of the top ten performing methods showed higher accuracy on single-domain proteins. Confidence intervals (95%) were determined using bootstrapping with n = 10,000 iterations on the set of target sequences.

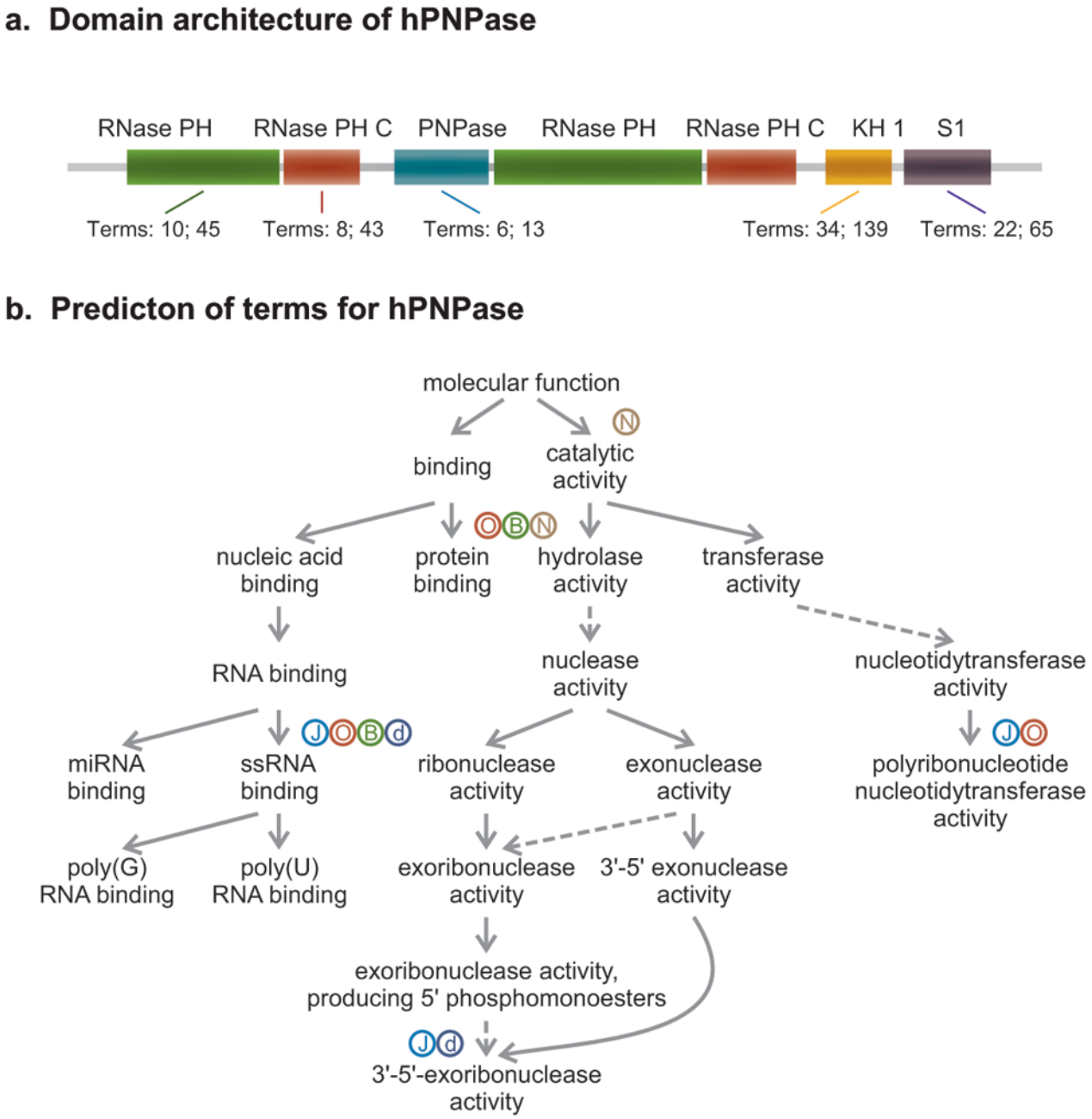

Figure 4.

Case study on the human *PNPT1* gene. (A) Domain architecture of human *PNPT1* gene according to the Pfam classification. For each domain, shown are the numbers of different leaf terms (for the Molecular Function and Biological Process categories) associated with any protein in Swiss-Prot containing this domain. (B) Molecular Function terms (six of which are leaves) associated with the human *PNPT1* gene in Swiss-Prot, as of December 2011. Colored circles represent the predicted terms for three representative methods as well as two baseline methods. The prediction threshold for each method was selected to correspond to the point in the precision/recall space that provides the maximum F-measure.

J (cyan) = Jones-UCL; O (magenta) = Team Orengo; d (navy blue) = dcGO; B (green) = BLAST; N (brown) = Naïve. Dashed lines in the ontology indicate presence of other terms between the source and destination nodes.