RESEARCH

An expanded evaluation of protein function prediction methods shows an improvement in accuracy

The CAFA Consortium

Full list of author information is available at the end of the article
*Corresponding Authors

Abstract

Background: The increasing volume and variety of genotypic and phenotypic data is a major defining characteristic of modern biomedical sciences. At the same time, the limitations in technology for generating data and the inherently stochastic nature of biomolecular events have led to the discrepancy between the volume of data and the amount of knowledge gleaned from it. A major bottleneck in our ability to understand the molecular underpinnings of life is the assignment of function to biological macromolecules, especially proteins. While molecular experiments provide the most reliable annotation of proteins, their relatively low throughput and restricted purview have led to an increasing role for computational function prediction. However, accurately assessing methods for protein function prediction and tracking progress in the field remain challenging.

Methodology: We have conducted the second Critical Assessment of Functional Annotation (CAFA), a timed challenge to assess computational methods that automatically assign protein function. One hundred twenty-six methods from 56 research groups were evaluated for their ability to predict biological functions using the Gene Ontology and gene-disease associations using the Human Phenotype Ontology on a set of 3,681 proteins from 18 species. CAFA2 featured significantly expanded analysis compared with CAFA1, with regards to data set size, variety, and assessment metrics. To review progress in the field, the analysis also compared the best methods participating in CAFA1 to those of CAFA2.

Conclusions: The top performing methods in CAFA2 outperformed the best methods from CAFA1, demonstrating that computational function prediction is improving. This increased accuracy can be attributed to the combined effect of the growing number of experimental annotations and improved methods for function prediction. The assessment also revealed that the definition of top performing algorithms is ontology specific, that different performance metrics can be used to probe the nature of accurate predictions, and the relative diversity of predictions in the biological process and human phenotype ontologies. While we have observed methodological improvement between CAFA1 and CAFA2, the interpretation of results and usefulness of individual methods remain context-dependent.

Keywords: Protein function prediction; disease gene prioritization

Introduction

Computational challenges in the life sciences have a successful history of driving the development of new methods by independently assessing performance and providing discussion forums for the researchers [1]. In 2010-2011, we organized the first Critical

Assessment of Functional Annotation (CAFA) challenge to evaluate methods for the automated annotation of protein function and to assess the progress in method development in the first decade of the 2000s [2]. The challenge used a time-delayed evaluation of predictions for a large set of target proteins without any experimental functional annotation. A subset of these target proteins accumulated experimental annotations after the predictions were submitted and was used to estimate the performance accuracy. The estimated performance was subsequently used to draw conclusions about the status of the field.

The CAFA1 experiment showed that advanced methods for the prediction of Gene Ontology (GO) terms [3] significantly outperformed a straightforward application of function transfer by local sequence similarity. In addition to validating investment in the development of new methods, CAFA1 also showed that using machine learning to integrate multiple sequence hits and multiple data types tends to perform well. However, CAFA1 also identified nontrivial challenges for experimentalists, biocurators and computational biologists. These challenges include the choice of experimental techniques and proteins in functional studies and curation, the structure and status of biomedical ontologies, the lack of comprehensive systems data that is necessary for accurate prediction of complex biological concepts, as well as limitations of evaluation metrics [2, 4, 5, 6, 7]. Overall, by establishing the state-of-the-art in the field and identifying challenges, CAFA1 set the stage for quantifying progress in the field of protein function prediction over time.

In this study, we report on the major outcomes of the second CAFA experiment (CAFA2) that was organized and conducted in 2013-2014, exactly three years after the original experiment. We were motivated to evaluate the progress in method development for function prediction as well as to expand the experiment to new ontologies. The CAFA2 experiment also greatly expanded the performance analysis to new types of evaluation and included new performance metrics.

Methods

Experiment overview

The timeline for the second CAFA experiment followed that of the first experiment and is illustrated in Figure 1. Briefly, CAFA2 was announced in July 2013 and officially started in September 2013, when 100,816 *target sequences* from 27 organisms were made available to the community. Teams were required to submit prediction scores within the $(0, 1]$ range for each protein-term pair they chose to predict on. The submission deadline for depositing these predictions was set for January 2014 (time point t_0). We then waited until September 2014 (time point t_1) for new experimental annotations to accumulate on the target proteins and assessed the performance of the prediction methods. We will refer to the set of all experimentally annotated proteins available at t_0 as the *training set* and to a subset of target proteins that accumulated experimental annotations during $(t_0, t_1]$ and used for evaluation as the *benchmark set*. It is important to note that the benchmark proteins and the resulting analysis vary based on the selection of time point t_1. For example, a preliminary analysis of the CAFA2 experiment was provided during the Automated Function Prediction Special Interest Group (AFP-SIG) meeting at the Intelligent Systems for Molecular Biology (ISMB) conference in July 2014.

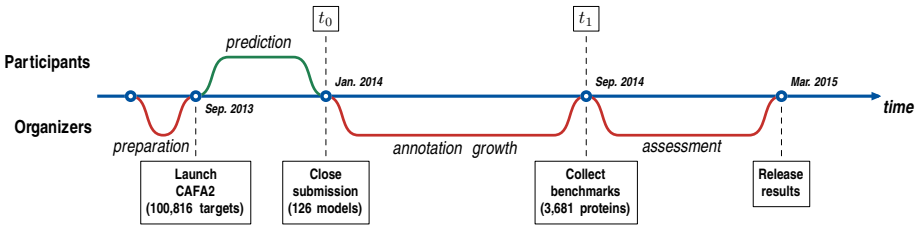

Figure 1 Timeline for the CAFA2 experiment.

The participating methods were evaluated according to their ability to predict terms in Gene Ontology (GO) [3] and Human Phenotype Ontology (HPO) [8]. In contrast with CAFA1, where the evaluation was carried out only for the Molecular Function Ontology (MFO) and Biological Process Ontology (BPO), in CAFA2 we also assessed the performance for the prediction of Cellular Component Ontology (CCO) terms in GO. The set of human proteins was further used to evaluate methods according to their ability to associate these proteins with disease terms from HPO, which included all sub-classes of the term HP:0000118, "Phenotypic abnormality".

In total, 56 groups submitting 126 methods participated in CAFA2. From those, 125 methods made valid predictions on a sufficient number of sequences. One-hundred and twenty-one methods submitted predictions for at least one of the GO benchmarks, while 30 methods participated in the disease-gene prediction tasks using HPO.

Evaluation

The CAFA2 experiment expanded the assessment of computational function prediction compared with CAFA1. This includes the increased number of targets, benchmarks, ontologies, and method comparison metrics.

We distinguish between two major types of method evaluation. The first, *protein-centric evaluation*, assesses performance accuracy of methods that predict all ontological terms associated with a given protein sequence. The second type, *term-centric evaluation*, assesses performance accuracy of methods that predict if a single ontology term of interest is associated with a given protein sequence [2]. The protein-centric evaluation can be viewed as a multi-label or structured-output learning problem of predicting a set of terms or a directed acyclic graph (a subgraph of the ontology) for a given protein. Because the ontologies contain many terms, the output space in this setting is extremely large and the evaluation metrics must incorporate similarity functions between groups of mutually interdependent terms (directed acyclic graphs). In contrast, the term-centric evaluation is an example of binary classification, where a given ontology term is assigned (or not) to an input protein sequence. These methods are particularly common in disease gene prioritization [9]. Put otherwise, a protein-centric evaluation considers a ranking of ontology terms for a given protein, whereas the term-centric evaluation considers a ranking of protein sequences for a given ontology term.

Both types of evaluation have merits in assessing performance. This is partly due to the statistical dependency between ontology terms, the statistical dependency

among protein sequences and also the incomplete and biased nature of the experimental annotation of protein function [6]. In CAFA2, we provide both types of evaluation, but we emphasize the protein-centric scenario for easier comparisons with CAFA1. We also draw important conclusions regarding method assessment in these two scenarios.

No-knowledge and limited-knowledge benchmark sets

In CAFA1, a protein was eligible to be in the benchmark set if it had not had any experimentally-verified annotations in any of the GO ontologies at time t_0 but accumulated at least one functional term with an experimental evidence code between t_0 and t_1. In CAFA2, we refer to such benchmark proteins as *no-knowledge* benchmarks. On the other hand, proteins with *limited knowledge* are those that had been experimentally annotated in one or two GO ontologies, but not in all three, at time t_0. For example, for the performance evaluation in MFO, a protein without any annotation in MFO prior to the submission deadline was allowed to have experimental annotations in BPO and CCO.

During the growth phase, the no-knowledge targets that have acquired experimental annotations in one or more ontologies became benchmarks in those ontologies. The limited-knowledge targets that have acquired additional annotations became benchmarks only for those ontologies for which there were no prior experimental annotations. The reason for using limited-knowledge targets was to identify whether the correlations between experimental annotations across ontologies can be exploited to improve function prediction.

The selection of benchmark proteins for evaluating HPO-term predictors was separated from the GO analyses. There exists only a no-knowledge benchmark set in the HPO category.

Partial and full evaluation modes

Many function prediction methods apply only to certain types of proteins, such as proteins for which 3D structure data are available, proteins from certain taxa, or specific subcellular localizations. To accommodate these methods, CAFA2 provided predictors with an option of choosing a subset of the targets to predict on as long as they computationally annotated at least 5,000 targets, of which at least 10 accumulated experimental terms. We refer to the assessment mode in which the predictions were evaluated only on those benchmarks for which a model made at least one prediction at any threshold as *partial evaluation mode*. In contrast, the *full evaluation mode* corresponds to the same type of assessment performed in CAFA1 where all benchmark proteins were used for the evaluation and methods were penalized for not making predictions.

In most cases, for each benchmark category, we have two types of benchmarks, no-knowledge (NK) and limited-knowledge (LK), and two modes of evaluation, full-mode (FM) and partial-mode (PM). Exceptions are all HPO categories that only have no-knowledge benchmarks. The full mode is appropriate for comparisons of general-purpose methods designed to make predictions on any protein, while the partial mode gives an idea of how well each method performs on a self-selected subset of targets.

Evaluation metrics

Precision-recall (*pr-rc*) curves and remaining uncertainty-misinformation (*ru-mi*) curves were used as the two chief metrics in the protein-centric mode. We also provide a single measure evaluation in both types of curves as a real-valued scalar to compare methods; however, we note that any choice of a single point on those curves is somewhat arbitrary and may not match the intended application objectives for a given algorithm. Thus, a careful understanding of the evaluation metrics used in CAFA is necessary to properly interpret the results.

Precision (pr), recall (rc) and the resulting $F_{\max}$ are defined as

$$\begin{aligned} pr(\tau) &= \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \frac{\sum_f \mathbb{1}\left(f \in P_i(\tau) \wedge f \in T_i\right)}{\sum_f \mathbb{1}\left(f \in P_i(\tau)\right)}, \\ rc(\tau) &= \frac{1}{n_e} \sum_{i=1}^{n_e} \frac{\sum_f \mathbb{1}\left(f \in P_i(\tau) \wedge f \in T_i\right)}{\sum_f \mathbb{1}\left(f \in T_i\right)}, \\ F_{\max} &= \max_\tau \left\{ \frac{2 \cdot pr(\tau) \cdot rc(\tau)}{pr(\tau) + rc(\tau)} \right\}, \end{aligned}$$

where $P_i(\tau)$ denotes the set of terms that have predicted scores greater than or equal to τ for a protein sequence i, T_i denotes the corresponding ground-truth set of terms for that sequence, $m(\tau)$ is the number of sequences with at least one predicted score greater than or equal to τ, $\mathbb{1}(\cdot)$ is an indicator function and n_e is the number of targets used in a particular mode of evaluation. In the full evaluation mode $n_e = n$, the number of benchmark proteins, whereas in the partial evaluation mode $n_e = m(0)$, i.e. the number of proteins which were chosen to be predicted using the particular method. For each method, we refer to $m(0)/n$ as the *coverage* because it provides the fraction of benchmark proteins on which the method made any predictions.

The remaining uncertainty (ru), misinformation (mi) and the resulting minimum semantic distance ($S_{\min}$) are defined as

$$\begin{aligned} ru(\tau) &= \frac{1}{n_e} \sum_{i=1}^{n_e} \sum_f ic(f) \cdot \mathbb{1}\left(f \notin P_i(\tau) \wedge f \in T_i\right), \\ mi(\tau) &= \frac{1}{n_e} \sum_{i=1}^{n_e} \sum_f ic(f) \cdot \mathbb{1}\left(f \in P_i(\tau) \wedge f \notin T_i\right), \\ S_{\min} &= \min_\tau \left\{ \sqrt{ru(\tau)^2 + mi(\tau)^2} \right\}, \end{aligned}$$

where $ic(f)$ is the information content of the ontology term f [10]. It is estimated in a maximum likelihood manner as the negative binary logarithm of the conditional probability that the term f is present in a protein's annotation given that all its parent terms are also present. Note that here, $n_e = n$ in the full evaluation mode and $n_e = m(0)$ in the partial evaluation mode applies to both ru and mi.

In addition to the main metrics, we used two secondary metrics. Those were the weighted version of the precision-recall curves and the version of the *ru-mi* curves normalized to the $[0, 1]$ interval. These metrics and the corresponding evaluation results are shown in Supplementary Materials.

For the term-centric evaluation we used the area under the Receiver Operating Characteristic (ROC) curve (AUC). The AUCs were calculated for all terms that have acquired at least 10 positively annotated sequences, whereas the remaining benchmarks were used as negatives. The term-centric evaluation was used both for ranking models and to differentiate well and poorly predictable terms. The performance of each model on each term is provided in Supplementary Materials.

As we required all methods to keep two significant figures for prediction scores, the threshold τ in all metrics used in this study exhaustively runs from 0.01 to 1.00 with the step size of 0.01.

Data sets

Protein function annotations for the Gene Ontology assessment were extracted, as a union, from three major protein databases that are available in the public domain: Swiss-Prot [11], UniProt-GOA [12] and the data from the GO consortium web site [3]. We used evidence codes EXP, IDA, IMP, IGI, IEP, TAS and IC to build benchmark and ground-truth sets. Annotations for the HPO assessment were downloaded from the Human Phenotype Ontology database [8].

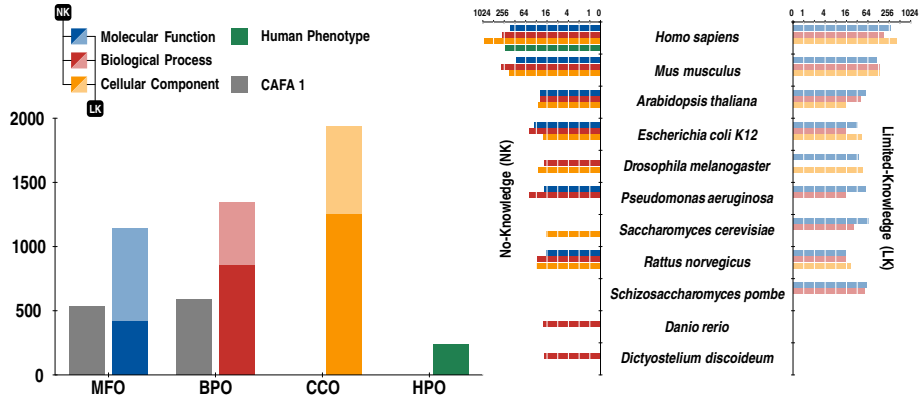

Figure 2 CAFA2 benchmark breakdown. The left panel shows the benchmark size for each of the four ontologies. The right panel lists the breakdown of benchmarks for both types over 11 species (with no less than 15 proteins) sorted according to the total number of benchmark proteins. For both panels, dark colors (blue, red, yellow) correspond to no-knowledge (NK) types, while their light color counterparts correspond to limited-knowledge (LK) types. The size of CAFA 1 benchmarks are shown in gray.

Figure 2 summarizes the benchmarks we used in this study. The left panel shows the benchmark sizes for each of the ontologies and compares these numbers to CAFA1. All species that have at least 15 proteins in any of the benchmark categories are listed in the right panel.

Baseline models

We built two baseline methods, Naïve and BLAST, and compared them with all participating methods. The Naïve method simply predicts the frequency of a term being annotated in a database [13]. BLAST was based on search results using the Basic Local Alignment Search Tool (BLAST) software against the training database [14]. A term will be predicted as the highest local alignment sequence identity among

all BLAST hits annotated with the term. Both of these two methods were "trained" on the experimentally annotated proteins available in Swiss-Prot at time t_0, except for HPO where the two baseline models were trained using the annotations from the t_0 release of the Human Phenotype Ontology.

Results

Overall performance

The performance accuracies of the top 10 methods are shown in Figures 3 and 4. The 95% confidence intervals were estimated using bootstrapping on the benchmark set with $B = 10,000$ iterations [15]. The results provide a broad insight into the state of the art.

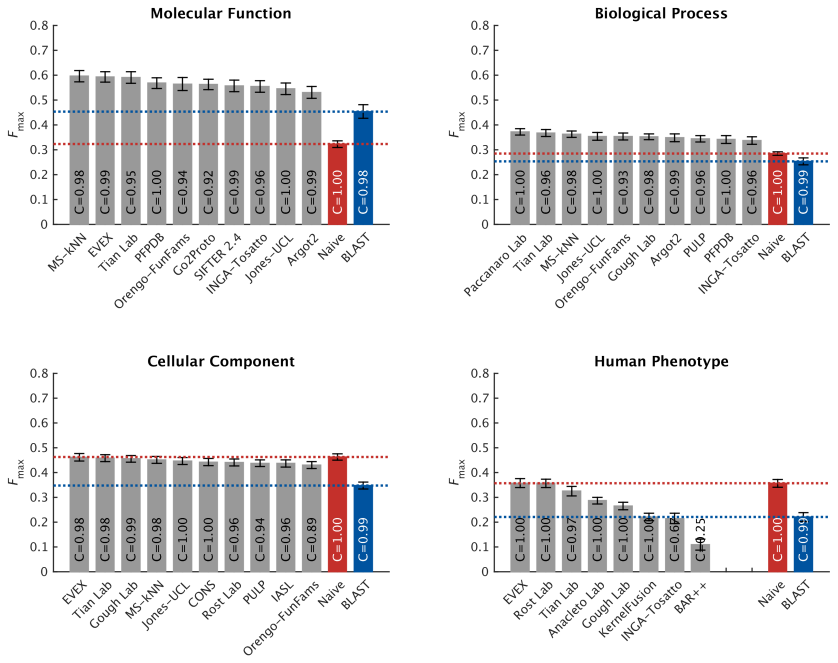

Figure 3 Overall evaluation using the maximum F-measure, F_{max}. Evaluation was carried out on no-knowledge benchmark sequences in the full mode. The coverage of each method is shown within its performance bar. A perfect predictor would be characterized with $F_{max} = 1$. Confidence intervals (95%) were determined using bootstrapping with 10,000 iterations on the set of benchmark sequences. For cases in which a principal investigator participated in multiple teams, only the results of the best-scoring method are presented. Details for all methods are provided in Supplementary Materials.

Predictors performed very differently across the four ontologies. Various reasons contribute to this effect including: (1) the topological properties of the ontology such as the size, depth, and branching factor; (2) term predictability; for example, the BPO terms are considered to be more abstract in nature than the MFO and CCO terms; (3) the annotation status, such as the size of the training set at t_0 as well as various annotation biases [6].

In general, CAFA2 methods perform better in predicting MFO terms than any other ontology. Top methods achieved the F_{max} scores around 0.6 and considerably

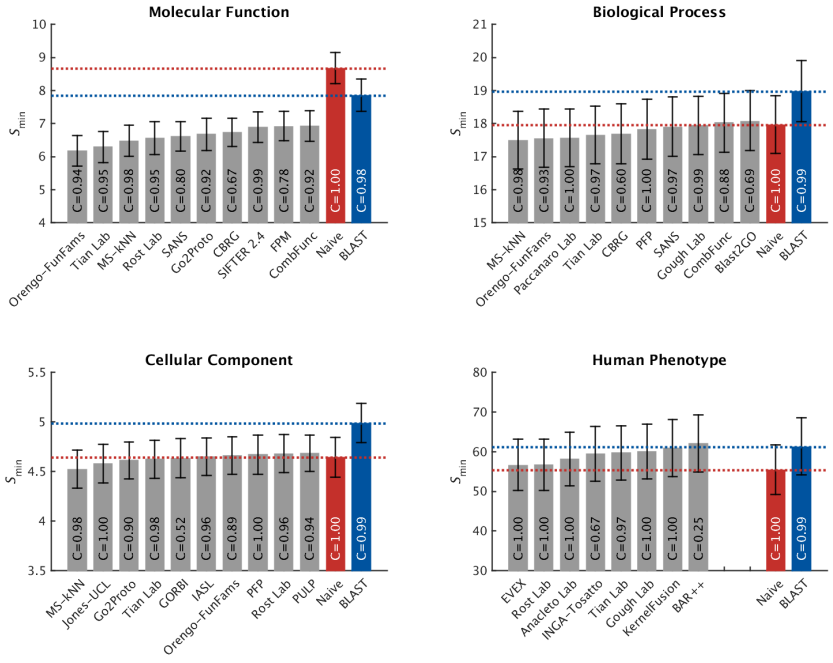

Figure 4 Overall evaluation using the minimum semantic distance, $S_{\min}$. Evaluation was carried out on no-knowledge benchmark sequences in the full mode. The coverage of each method is shown within its performance bar. A perfect predictor would be characterized with $S_{\min} = 0$. Confidence intervals (95%) were determined using bootstrapping with 10,000 iterations on the set of benchmark sequences. For cases in which a principal investigator participated in multiple teams, only the results of the best-scoring method are presented. Details for all methods are provided in Supplementary Materials.

surpassed the two baseline models. Maintaining the pattern from CAFA1, the performance accuracies in the BPO category were not as good as in the MFO category. The best-performing method scored slightly below 0.4.

For the two newly-added ontologies in CAFA2, we observed that the top predictors performed no better than the Naïve method under $F_{\max}$, whereas they slightly outperformed the Naïve method under $S_{\min}$ in CCO. One possible reason for the competitive performance of the Naïve method in the CCO category is the fact that a small number of relatively general terms are frequently used, and those relative frequencies do not diffuse quickly enough with the depth of the graph. For instance, the annotation frequency of "organelle" (GO:0043226, level 2), "intracellular part" (GO:0044424, level 3) and "cytoplasm" (GO:0005737, level 4) are all above the best threshold for the Naïve method ($\tau_{\text{optimal}} = 0.32$). Correctly predicting these terms increases the number of "true positives" and thus boosts the performance of the Naïve method under the $F_{\max}$ evaluation. However, once the less informative terms are down-weighted (using the $S_{\min}$ measure), the Naïve method becomes significantly penalized and degraded. The weighted $F_{\max}$ and normalized $S_{\min}$ evaluations can be found in Supplementary Materials.

However, high frequency of general terms does not seem to be the major reason for the observed performance in the HPO category. One possible explanation for this effect would be that the average number of HPO terms associated with a human protein is much larger than in GO. The mean number of annotations per protein in HPO is 84, while the MFO, BPO and CCO the mean number of annotations per protein are 10, 39, and 14 respectively. The high number of annotations per protein makes prediction using HPO terms significantly more difficult. In addition, unlike for GO terms, the HPO annotations cannot be transferred from other species based on homology and other available data. Successfully predicting the HPO terms in the protein-centric mode is a difficult problem.

Term-centric evaluation

Protein-centric view, despite its power in showing the strengths of a predictor, does not gauge a predictor's performance for a specific function. We therefore also assessed predictors in the term-centric manner by calculating AUCs for individual terms. Averaging those AUCs over terms provides a metric for ranking predictors, whereas averaging performances over terms provides insights into how well this term can be predicted computationally by the community.

Figure 5 shows the performance evaluation where the AUCs for each method were averaged over all terms for which at least ten positive sequences were available. Proteins without predictions were counted as predictions with a score of 0. As shown in Figures 3-4, correctly predicting CCO and HPO terms for a protein might not be an easy task according to the protein-centric results. However, the overall poor performances could also result from the dominance of poorly predictable terms. Therefore, a term-centric view can help differentiate prediction quality across terms. As shown in Figure 6, most of the terms in HPO obtain AUC greater than the Naïve model, with some terms on average achieving reasonably well AUCs around 0.7. Depending on the training data available for participating methods, well predicted phenotype terms range from mildly specific such as "Lymphadenopathy" and "Thrombophlebitis" to general ones such as "Abnormality of the Skin Physiology".

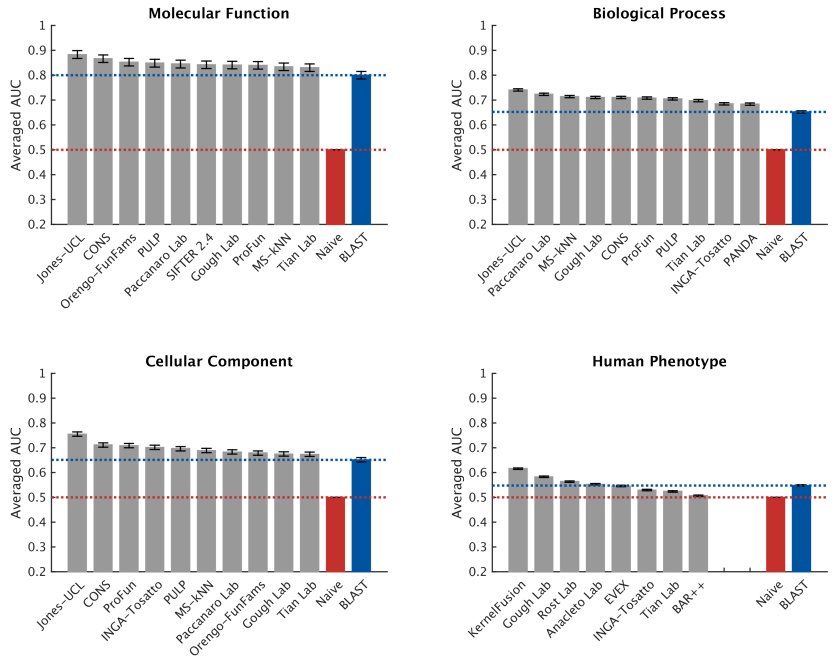

Figure 5 Overall evaluation using the averaged AUC over terms with no less than 10 positive annotations. Evaluation was carried out on no-knowledge benchmark sequences in the full mode. Error bars indicate the standard error in averaging AUC over terms for each method. For cases in which a principal investigator participated in multiple teams, only the results of the best-scoring method are presented. Details for all methods are provided in Supplementary Materials.

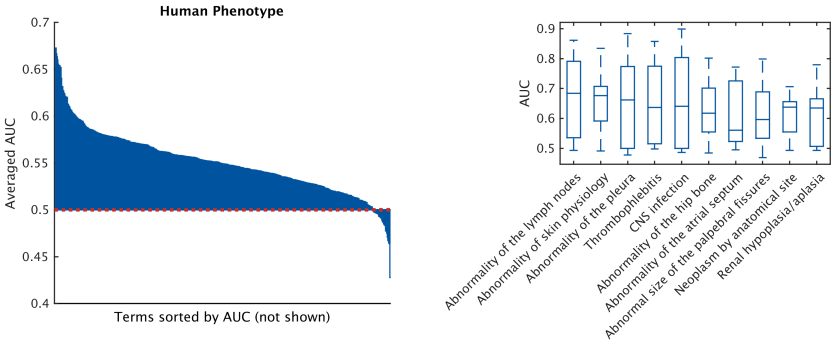

Figure 6 Averaged AUC per term for Human Phenotype ontology. Left panel: Terms are sorted based on AUC, dashed red line indicates the performance of the Naïve method. Right panel: The top 10 accurately predicted terms without overlapping ancestors (except for the root).

Performance on various categories of benchmarks

Easy vs. difficult benchmarks

As in CAFA1, the no-knowledge GO benchmarks were divided into "easy" versus "difficult" categories based on their maximal global sequence identity with proteins in the training set. Since the distribution of sequence identities roughly forms a bimodal shape (Supplementary Materials), a cutoff of 60% was manually chosen to define the two categories. The same cutoff was used in CAFA1. Unsurprisingly,

across all three ontologies, the performance of the BLAST model was substantially impacted for the difficult category because of the lack of high sequence identity homologs and as a result, transferring annotations was relatively unreliable. However, we also observed that most top methods were insensitive to the types of benchmarks, which provides us with encouraging evidence that state-of-the-art protein function predictors can successfully combine multiple potentially unreliable hits, as well as multiple types of data, into a reliable prediction.

Species-specific categories

The benchmark proteins were split into even smaller categories for each species as long as the resulting category contained at least 15 sequences. However, because of space limitations, we only show the breakdown results on eukarya and prokarya benchmarks in Figure 7 (the species-specific results are provided in Supplementary Materials). It is worth noting that the performance accuracies on the entire benchmark sets were dominated by the targets from eukarya due to their larger proportion in the benchmark set and annotation preferences. The eukarya benchmark rankings therefore coincide with the overall rankings, but the smaller categories typically showed different rankings and may be informative to more specialized research groups.

For all three GO ontologies, no-knowledge prokarya benchmark sequences collected over the annotation growth phase mostly (over 80%) came from two species: *E. coli* and *P. aeruginosa* (for CCO, 21 out of 22 proteins were from *E. coli*). Thus, one should keep in mind that the prokarya benchmarks essentially reflect the performance on proteins from these two species. Methods predicting the MFO terms for prokaryotes are slightly worse than those for eukaryotes. In addition, direct function transfer by homology for prokaryotes did not work well using this ontology. However, the performance was better using the other two ontologies, especially CCO. It is not very surprising that top methods achieved good performance for *E. coli* as it is a well-studied model organism.

Top methods have improved since CAFA1

The second CAFA experiment was conducted three years after the first one. As our knowledge of protein function has increased since then, it was worthwhile to assess whether computational methods have also been improved and if so, to what extent. Therefore, to monitor the progress of the community over time, we revisit some of the top methods in CAFA1 and compare them with their successors.

The comparison was done on an overlapping benchmark set created from CAFA1 targets and CAFA2 targets. More precisely, we used the stored predictions on the target proteins from CAFA1 and compared them with the new predictions from CAFA2 on the overlapping set of CAFA2 benchmarks and CAFA1 targets (a sequence had to be a no-knowledge target in both experiments to be eligible in this evaluation). For this purpose, we used a hypothetical ontology by taking the intersection of the two Gene Ontology snapshots (versions from January 2011 and June 2013) so as to mitigate the influence of ontology changes. We thus collected 356 benchmark proteins for MFO comparisons and 698 for BPO comparisons. The two baseline methods were trained on respective Swiss-Prot annotations for both

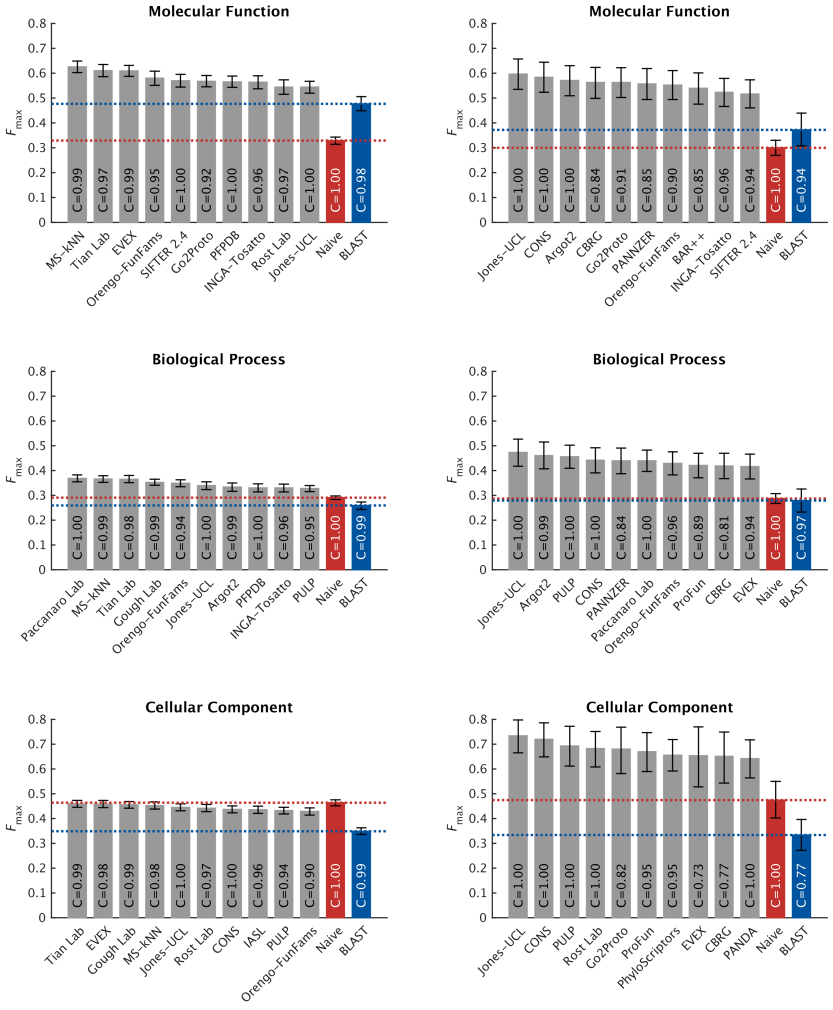

Figure 7 Performance evaluation using the maximum F-measure, F_{max}, on eukaryotic (left) versus prokaryotic (right) benchmark sequences. Evaluation was carried out on no-knowledge benchmark sequences in the full mode. The coverage of each method is shown within its performance bar. Confidence intervals (95%) were determined using bootstrapping with 10,000 iterations on the set of benchmark sequences. For cases in which a principal investigator participated in multiple teams, only the results of the best-scoring method are presented. Details for all methods are provided in Supplementary Materials.

ontologies so that they serve as controls for database change. In particular, SwissProt2011 (for CAFA1) contained 29,330 and 31,282 proteins for MFO and BPO, while SwissProt2014 (for CAFA2) contained 26,907 and 41,959 proteins for the two ontologies.

To conduct a "head-to-head" analysis between any two methods, we generated $B = 10,000$ bootstrap samples and let methods compete on each such benchmark set. The average performance metric as well as the number of wins were recorded. Figure 8 summarizes the results of this analysis. We use a color code from green to

red to indicate the performance improvement δ from CAFA1 to CAFA2,

$$\delta(m_2, m_1) = \frac{1}{n}\sum_{i=1}^{n} F_{\max}^{(i)}(m_2) - \frac{1}{n}\sum_{i=1}^{n} F_{\max}^{(i)}(m_1)$$

where m_1 and m_2 stand for methods from CAFA1 and CAFA2, respectively, and $F_{\max}^{(i)}(\cdot)$ represents the $F_{\max}$ of a method evaluated on the i-th bootstrapped benchmark set. The selection of top methods for this study was based on their performance in each ontology on the entire benchmark sets. Panels B and C in Figure 8 show the comparison between baseline methods trained on different data sets. We see no improvements of these baselines except for BLAST on BPO where it is slightly better to use the newer version of Swiss-Prot as the reference database for the search. On the other hand, all top methods in CAFA2 outperformed their counterparts in CAFA1. For predicting molecular functions, even though transferring functions from BLAST hits does not give better results, the top models still managed to perform better. It is possible that the newly acquired annotations since CAFA1 enhanced BLAST, which involves direct function transfer, and perhaps lead to better performances of those "downstream" methods that rely on sequence alignments. However, this effect does not completely explain the extent of performance improvement achieved by those methods. This is promising evidence that top methods from the community have improved since CAFA1 and that the improvement was not simply due to updates of curated databases.

Diversity of methodology

We analyzed the extent to which methods generated similar predictions within each ontology. We calculated the pairwise Pearson correlation between methods on a common set of gene-concept pairs and then visualized these similarities as networks (Supplementary Materials).

In the molecular function ontology, where we observed the highest overall performance of prediction methods, eight of ten top methods were in the largest connected component. In addition, we observed a high connectivity between methods, suggesting that the participating methods are leveraging similar sources of data in similar ways. Predictions for the biological process ontology showed a contrasting pattern. In this ontology, the largest connected component contained only two of the top ten methods. The other top methods were contained in components made up of other methods produced by the same lab. This suggests that the approaches that participating groups have taken generate more diverse predictions for this ontology and that there are many different paths to a top performing biological process prediction method. Results for the human phenotype ontology were more similar to the biological process ontology, while results for cellular component were more similar in structure to molecular function.

Taken together, these results suggest that ensemble approaches that aim to include independent sources of high quality predictions may benefit from leveraging the data and techniques used by different research groups and that such approaches that effectively weigh and integrate disparate methods may demonstrate more substantial improvements over existing methods in the process and phenotype ontologies where current prediction approaches share less similarity.

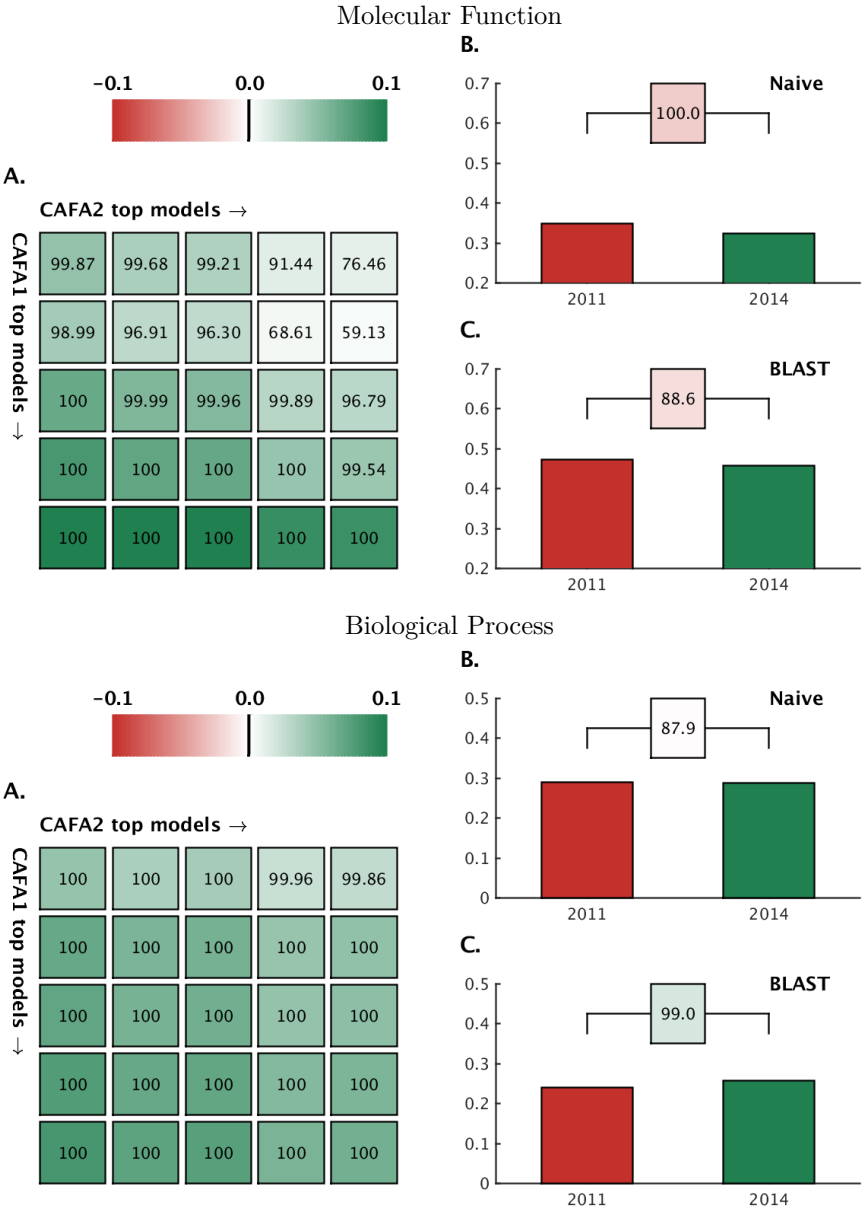

Figure 8 CAFA1 versus CAFA2 (top methods) A comparison in $F_{\max}$ between top 5 CAFA1 models against top 5 CAFA2 models. Colored boxes encode the results such that (1) colors indicate margins of a CAFA2 method over a CAFA1 method in $F_{\max}$ and (2) numbers in the box indicate the percentage of wins. For both MFO and BPO results, **A**. CAFA1 top 5 models (rows, from top to bottom) against CAFA2 top 5 models (columns, from left to right) **B**. Comparison of Naïve baselines trained respectively on SwissProt2011 and SwissProt2014. **C**. Comparison of BLAST baselines trained on SwissProt2011 and SwissProt2014.

Conclusions

Accurately annotating the function of biological macromolecules is difficult, and requires the concerted effort of experimental scientists, biocurators, and computational biologists. We conducted the second CAFA challenge to assess the status of computational function prediction of proteins and to quantify the progress in the field. Following the success of CAFA1 three years ago, we decided to significantly

expand the number of protein targets, the number of biomedical ontologies used for annotation, the number of analysis scenarios, as well as the metrics used for evaluation. We believe the results of the CAFA2 experiment provide useful information on the status of the state-of-the-art in protein function prediction, can guide the development of new concept annotation methods, and help experimental studies through prioritization. Understanding the function of biological macromolecules brings us closer to understanding life at the molecular level and improving human health.

The field has moved forward

Three years ago, in CAFA1, we concluded that the top methods for function prediction outperform straightforward function transfer by homology. In CAFA2, we observe that the methods for function prediction have improved compared to those from CAFA1. As part of the CAFA1 experiment, we stored all predictions from all methods on 48,298 target proteins from 18 species. We used those stored predictions and compared them to the newly deposited predictions from CAFA2 on the overlapping set of benchmark proteins and CAFA1 targets. The head-to-head comparisons among top five CAFA1 methods against top five CAFA2 methods reveal that the top CAFA2 methods outperformed all top CAFA1 methods.

Although it is difficult to disentangle the contributions of larger training sets from those of methodological novelties, the fact that the BLAST algorithm using the data from 2011 and data from 2014 showed little difference, led us to conclude that a larger share of the contribution likely belongs to the new methods. The experiences from CAFA1 and continuous AFP-SIG meetings every year during the ISMB conference where many new developments are readily shared may have contributed to this outcome [16].

Evaluation metrics

A fair performance assessment in protein function prediction is far from straightforward. Although various evaluation metrics have been proposed under the framework of multi-label and structured-output learning, the evaluation in this subfield also needs to be interpretable to a broad community of researchers as well as the public. To address this, we used several metrics in this study as each provides useful insights and complements the others. Understanding the strengths and weaknesses of current metrics and developing better metrics remains important.

One important observation with respect to metrics is that the protein-centric and term-centric views may give different perspectives to the same problem. For example, while in the MFO and BPO we generally observe positive correlation between the two, in CCO and HPO these different metrics might lead to entirely different interpretations of the experiment. Regardless of the underlying cause, as discussed in Results, it is clear that some ontological terms are predictable with high accuracy and can be reliably used in practice even in these ontologies. In the meantime, more effort will be needed to understand the problems associated with statistical and computational aspects of method development.

In CAFA2 we introduced minimum semantic distance as another protein-centric metric [10]. The investigation of the BLAST baseline reveals that the best local sequence identity cutoff for transferring experimental annotations from sequence hits

occurs around 0.5 for all three GO ontologies and just under (0.35) for HPO, if $F_{\max}$ is used as the evaluation metric. However, for $S_{\min}$, these cutoffs are substantially higher to over 0.6 for MFO, 0.7 for HPO and surprisingly over 0.9 for both BPO and CCO. We believe these higher thresholds provide biologically interesting results and have thus decided to use both *pr-rc* curves and *ru-mi* curves in protein-centric performance assessments.

Well-performing methods

We observe that participating methods usually specialize in one or a few categories of protein function prediction and have been developed with their own application objectives in mind. Therefore, performance rankings of methods often change from one benchmark set to another. There are complex factors that influence the final ranking including the selection of the ontology, types of benchmark sets and evaluation, as well as evaluation metrics, as discussed earlier. Most of our assessment results show that the performances of top-performing methods are generally comparable to each other. Thus, although a small group of methods could be considered as generally good, there is no single method that dominates over all benchmarks.

We also observed that when provided a chance to select a reliable set of predictions, the methods generally perform better (partial evaluation mode vs. full evaluation mode). Although most methods seem not to have been actively developed for the partial evaluation mode, this outcome is very encouraging. On the other hand, the limited-knowledge category of assessment seems to have not provided any boost in terms of performance accuracy. However, this was a new prediction category in CAFA and so few methods may have been optimized for prediction in the limited-knowledge scenario. Many important comparisons can be found in Supplementary Materials.

Final notes

The automated functional annotation remains an exciting yet challenging task with implications relevant to the entirety of biomedical sciences. Three years after CAFA1, the top methods from the community have shown encouraging progress in both MFO and BPO categories. However, in terms of raw scores, there is still significant room for improvement in all ontologies, and particularly in BPO, CCO and HPO. There is also a need to develop an experiment-driven, as opposed to curation driven, component of the evaluation to address limitations for term-centric evaluation. In the future CAFA experiments, we will continue to monitor the performance over time and invite a broad range of computational biologists, computer scientists, statisticians and others to address these engaging problems of concept annotation for biological macromolecules through CAFA.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PR and IF conceived of the CAFA experiment, supervised the project and significantly contributed to writing of the manuscript. YJ performed most analyses and significantly contributed to writing. IF, PR, CSG, WTC, ARB, DD and RL contributed to the analyses. SDM managed data acquisition. TRO developed the web interface, including the portal for submission and the storage of predictions. RPH, MJM and CO'D directed the biocuration efforts. EC-U, PD, REF, RH, DL, RCL, MM, ANM, PM-M, KP and AS performed biocuration. YM and PNR co-organized the human phenotype challenge. ML, AT, PCB, SEB, CO and BR steered the CAFA experiment and provided critical guidance. The remaining authors participated in the experiment, provided writing and data for their methods and contributed comments on the manuscript.

Acknowledgements
We acknowledge the contributions by Maximilian Hecht, Alexander Grün, Julia Krumhoff, My Nguyen Ly, Jonathan Boidol, Rene Schoeffel, Yann Spöri, Jessika Binder, Christoph Hamm and Karolina Worf. This work was partially supported by the following grants: National Science Foundation (NSF) grants DBI-1458477 (PR), DBI-1458443 (SDM), DBI-1458390 (CSG), DBI-1458359 (IF), IIS-1319551 (DK), DBI-1262189 (DK) and DBI-1149224 (JC); National Institutes of Health (NIH) grants R01GM093123 (JC), R01GM097528 (DK), R01GM076990 (PP), R01GM071749 (SEB), R01LM009722 (SDM) and UL1TR000423 (SDM); the National Natural Science Foundation of China 3147124 (WT) and 91231116 (WT); the National Basic Research Program of China 2012CB316505 (WT); NSERC RGPIN 371348-11 (PP); FP7 "infrastructures" project TransPLANT Award 283496 (ADJvD); Microsoft Research/FAPESP grant 2009/53161-6 and FAPESP fellowship 2010/50491-1 (DCAeS); Biotechnology and Biological Sciences Research Council (BBSRC) grants BB/L020505/1 (DTJ), BB/F020481/1 (MJES), BB/K004131/1 (AP), BB/F00964X/1 (AP) and BB/L018241/1 (CD); Spanish Ministry of Economics and Competiveness, grant number BIO2012-40205 (MT); KU Leuven CoE PFV/10/016 SymBioSys (YM); the Newton International Fellowship Scheme of the Royal Society grant NF080750 (TN). CSG was supported in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4552. Computational resources were provided by CSC — IT Center for Science Ltd, Espoo, Finland (TS). This work was supported by the Academy of Finland (TS). RCL, ANM were supported by British Heart Foundation grant RG/13/5/30112. PD, RCL, REF were supported by Parkinson's UK grant G-1307. Alexander von Humboldt Foundation through German Federal Ministry for Education and Research and Ernst Ludwig Ehrlich Studienwerk (ELES). Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant no. 173001). This work was a Technology Development effort for ENIGMA - Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research (DE-AC02-05CH11231). ENIGMA only covers the application of this work to microbial proteins.

References

1. Costello, J.C., Stolovitzky, G.: Seeking the wisdom of crowds through challenge-based competitions in biomedical research. Clin Pharmacol Ther **93**(5), 396–398 (2013)
2. Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J.M., Talwalkar, A.S., Repo, S., Souza, M.L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Toronen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D.W., Bryson, K., Jones, D.T., Limaye, B., Inamdar, H., Datta, A., Manjari, S.K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A.M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R., Yang, H., Romero, A.E., Bhat, P., Paccanaro, A., Hamp, T., Kassner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Honigschmid, P., Hopf, T.A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Bjorne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F., Sommer, I., Wass, M.N., Sternberg, M.J., Skunca, N., Supek, F., Bosnjak, M., Panov, P., Dzeroski, S., Smuc, T., Kourmpetis, Y.A., van Dijk, A.D., ter Braak, C.J., Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Di Camillo, B., Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P.C., Brenner, S.E., Orengo, C., Rost, B., Mooney, S.D., Friedberg, I.: A large-scale evaluation of computational protein function prediction. Nat Methods **10**(3), 221–227 (2013)
3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25**(1), 25–29 (2000)
4. Dessimoz, C., Skunca, N., Thomas, P.D.: CAFA and the open world of protein function predictions. Trends Genet **29**(11), 609–610 (2013)
5. Gillis, J., Pavlidis, P.: Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA). BMC Bioinformatics **14**(Suppl 3), 15 (2013)
6. Schnoes, A.M., Ream, D.C., Thorman, A.W., Babbitt, P.C., Friedberg, I.: Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput Biol **9**(5), 1003063 (2013)
7. Jiang, Y., Clark, W.T., Friedberg, I., Radivojac, P.: The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. Bioinformatics **30**(17), 609–616 (2014)
8. Robinson, P.N., Mundlos, S.: The human phenotype ontology. Clin Genet **77**(6), 525–534 (2010)
9. Moreau, Y., Tranchevent, L.C.: Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet **13**(8), 523–536 (2012)
10. Clark, W.T., Radivojac, P.: Information-theoretic evaluation of predicted ontological annotations. Bioinformatics **29**(13), 53–61 (2013)
11. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S.: The Universal Protein Resource (UniProt). Nucleic Acids Res **33**(Database issue), 154–159 (2005)
12. Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J., O'Donovan, C.: The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res **43**(Database issue), 1057–1063 (2015)
13. Clark, W.T., Radivojac, P.: Analysis of protein function and its prediction from amino acid sequence. Proteins **79**(7), 2086–2096 (2011)
14. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**(17), 3389–3402 (1997)
15. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)

16. Wass, M.N., Mooney, S.D., Linial, M., Radivojac, P., Friedberg, I.: The automated function prediction SIG looks back at 2013 and prepares for 2014. Bioinformatics **30**(14), 2091–2092 (2014)

Additional Files

This submission contains two supplementary documents. The first document is provides a subset of CAFA2 analyses that are equivalent to those provided about the CAFA1 experiment in the CAFA1 supplement. The large data repository provides all additional data, analyses as well as full prediction results for every method. The entire library of code used in CAFA2 is available at `https://github.com/yuxjiang/CAFA2`.