The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens

Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., Lewis, K. A., Georghiou, G., Nguyen, H. N., Hamid, M. N., Davis, L., Dogan, T., Atalay, V., Rifaioglu, A. S., Dalkiran, A., Cetin-Atalay, R., Zhang, C., Hurto, R. L., Freddolino, P. L., Zhang, Y., Bhat, P., Supek, F., Fernández, J. M., Gemovic, B., Perovic, V. R., Davidović, R. S., Sumonja, N., Veljkovic, N., Asgari, E., Mofrad, M. R. K., Profiti, G., Savojardo, C., Martelli, P. L., Casadio, R., Boecker, F., Kahanda, I., Thurlby, N., McHardy, A. C., Renaux, A., Saidi, R., Gough, J., Freitas, A. A., Antczak, M., Fabris, F., Wass, M. N., Hou, J., Cheng, J., Hou, J., Wang, Z., Romero, A. E., Paccanaro, A., Yang, H., Goldberg, T., Zhao, C., Holm, L., Törönen, P., Medlar, A. J., Zosa, E., Borukhov, I., Novikov, I., Wilkins, A., Lichtarge, O., Chi, P.-H., Tseng, W.-C., Linial, M., Rose, P. W., Dessimoz, C., Vidulin, V., Dzeroski, S., Sillitoe, I., Das, S., Lees, J. G., Jones, D. T., Wan, C., Cozzetto, D., Fa, R., Torres, M., Vesztrocy, A. W., Rodriguez, J. M., Tress, M. L., Frasca, M., Notaro, M., Grossi, G., Petrini, A., Re, M., Valentini, G., Mesiti, M., Roche, D. B., Reeb, J., Ritchie, D. W., Aridhi, S., Alborzi, S. Z., Devignes, M.-D., Emily Koo, D. C.,

Consortium *Genome Biology* (2019) 20:244
https://doi.org/10.1186/s13059-019-1835-8

Genome Biology

RESEARCH Open Access

The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens

Naihui Zhou[1,2], Yuxiang Jiang[3], Timothy R. Bergquist[4], Alexandra J. Lee[5], Balint Z. Kacsoh[6,7], Alex W. Crocker[8], Kimberley A. Lewis[8], George Georghiou[9], Huy N. Nguyen[1,10], Md Nafiz Hamid[1,2], Larry Davis[2], Tunca Dogan[11,37], Volkan Atalay[12], Ahmet S. Rifaioglu[12,13], Alperen Dalkıran[12], Rengul Cetin Atalay[14], Chengxin Zhang[15], Rebecca L. Hurto[16], Peter L. Freddolino[15,16], Yang Zhang[15,16], Prajwal Bhat[17], Fran Supek[18,19], José M. Fernández[20,21], Branislava Gemovic[22], Vladimir R. Perovic[22], Radoslav S. Davidović[22], Neven Sumonja[22], Nevena Veljkovic[22], Ehsaneddin Asgari[23,24], Mohammad R.K. Mofrad[25], Giuseppe Profiti[26,27], Castrense Savojardo[26], Pier Luigi Martelli[26], Rita Casadio[26], Florian Boecker[28], Heiko Schoof[29], Indika Kahanda[30], Natalie Thurlby[31], Alice C. McHardy[32,33], Alexandre Renaux[34,35,36], Rabie Saidi[37], Julian Gough[38], Alex A. Freitas[39], Magdalena Antczak[40], Fabio Fabris[39], Mark N. Wass[40], Jie Hou[41,42], Jianlin Cheng[42], Zheng Wang[43], Alfonso E. Romero[44], Alberto Paccanaro[44], Haixuan Yang[45], Tatyana Goldberg[129], Chenguang Zhao[49], Liisa Holm[50], Petri Törönen[50], Alan J. Medlar[50], Elaine Zosa[51], Itamar Borukhov[51], Ilya Novikov[53], Angela Wilkins[54], Olivier Lichtarge[54], Po-Han Chi[55], Wei-Cheng T seng[56], Michal Linial[57], Peter W. Rose[58], Christophe Dessimoz[59,60,61], Vedrana Vidulin[62], Saso Dzeroski[63,64], Ian Sillit oe[65], Sayoni Das[66], Jonathan Gill Lees[66,67], David T. Jones[69,70], Cen Wan[68,69], Domenico Cozzetto[68,69], Rui Fa[68,69], Mateo Torres[44], Alex Warwick Vesztrocy[70,71], Jose Manuel Rodriguez[72], Michael L. Tress[73], Marco Frasca[74], Marco Notaro[74], Giuliano Grossi[74], Alessandro Petrini[74], Matteo Re[74], Giorgio Valentini[74], Marco Mesiti[74], Daniel B. Roche[76], Jonas Reeb[76], David W. Ritchie[77], Sabeur Aridhi[77], Seyed Ziaeddin Alborzi[77,79], Marie-Dominique Devignes[77,78,79], Da Chen Emily Koo[80], Richard Bonneau[81,82], Vladimir Gligorijević[83], Meet Barot[84], Hai Fang[85], Stefano Toppo[86], Enrico Lavezzo[86], Marco Falda[87], Michele Berselli[86], Silvio C.E. Tosatto[88,89], Marco Carraro[89], Damiano Piovesan[89], Hafeez Ur Rehman[90], Qizhong Mao[91,92], Shanshan Zhang[91], Slobodan Vucetic[91], Gage S. Black[93,94], Dane Jo[93,94], Erica Suh[93], Jonathan B. Dayton[93,94], Dallas J. Larsen[93,94], Ashton R. Omdahl[93,94], Liam J. McGuffin[95], Danielle A. Brackenridge[95], Patricia C. Babbitt[96,98], Jeffrey M. Yunes[97,98], Paolo Fontana[99], Feng Zhang[100,101], Shanfeng Zhu[102,103,104], Ronghui You[102,103,104], Zihan Zhang[102,104], Suyang Dai[102,104], Shuwei Yao[102,103], Weidong Tian[105,106], Renzhi Cao[107], Caleb Chandler[107], Miguel Amezola[107], Devon Jo hnson[107], Jia-Ming Chang[108], Wen-Hung Liao[108], Yi-Wei Liu[108], Stefano Pascarelli[109], Yotam Frank[110], Robert Hoeh ndorf[111], Maxat Kulmanov[111], Imane Boudellioua[112,113], Gianfranco Politano[114], Stefano Di Carlo[114], Alfredo Benso[114], Kai Hakala[115,116], Filip Ginter[115,117], Farrokh Mehryary[115,116], Suwisa Kaewphan[115,116,118], Jari Björne[119,120], Hans Moen[117], Martti E.E. Tolvanen[121], Tapio Salakoski[119,120], Daisuke Kihara[122,123], Aashish Jain[124], Tomislav Šmuc[125], Adrian Altenhoff[126,127], Asa Ben-Hur[128], Burkhard Rost[129,130], Steven E. Brenner[131], Christine A. Orengo[66], Constance J. Jeffery[132], Giovanni Bosco[133], Deborah A. Hogan[6,8], Maria J. Martin[9], Claire O'Donovan[9], Sean D. Mooney[4], Casey S. Greene[134,135], Predrag Radivojac[136*] and Iddo Friedberg[1*]

*Correspondence: predrag@northeastern.edu; idoerg@iastate.edu
[1]Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA, USA
[136]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
Full list of author information is available at the end of the article

Abstract

Background: The Critical Assessment of Functional Annotation (CAFA) is an ongoing, global, community-driven effort to evaluate and improve the computational annotation of protein function.

Results: Here, we report on the results of the third CAFA challenge, CAFA3, that featured an expanded analysis over the previous CAFA rounds, both in terms of volume of data analyzed and the types of analysis performed. In a novel and major new development, computational predictions and assessment goals drove some of the experimental assays, resulting in new functional annotations for more than 1000 genes. Specifically, we performed experimental whole-genome mutation screening in *Candida albicans* and *aeruginosa* genomes, which provided us with genome-wide experimental data for genes associated with biofilm formation and motility. We further performed targeted assays on selected genes in *Drosophila melanogaster*, which we suspected of being involved in long-term memory.

Conclusion: We conclude that while predictions of the molecular function and biological process annotations have slightly improved over time, those of the cellular component have not. Term-centric prediction of experimental annotations remains equally challenging; although the performance of the top methods is significantly better than the expectations set by baseline methods in *C. albicans* and *D. melanogaster*, it leaves considerable room and need for improvement. Finally, we report that the CAFA community now involves a broad range of participants with expertise in bioinformatics, biological experimentation, biocuration, and bio-ontologies, working together to improve functional annotation, computational function prediction, and our ability to manage big data in the era of large experimental screens.

Keywords: Protein function prediction, Long-term memory, Biofilm, Critical assessment, Community challenge

Introduction

High-throughput nucleic acid sequencing [1] and mass-spectrometry proteomics [2] have provided us with a deluge of data for DNA, RNA, and proteins in diverse species. However, extracting detailed functional information from such data remains one of the recalcitrant challenges in the life sciences and biomedicine. Low-throughput biological experiments often provide highly informative empirical data related to various functional aspects of a gene product, but these experiments are limited by time and cost. At the same time, high-throughput experiments, while providing large amounts of data, often provide information that is not specific enough to be useful [3]. For these reasons, it is important to explore computational strategies for transferring functional information from the group of functionally characterized macromolecules to others that have not been studied for particular activities [4–9].

To address the growing gap between high-throughput data and deep biological insight, a variety of computational methods that predict protein function have been developed over the years [10–24]. This explosion in the number of methods is accompanied by the need to understand how well they perform, and what improvements are needed to satisfy the needs of the life sciences community. The Critical Assessment of Functional Annotation (CAFA) is a community challenge that seeks to bridge the gap between the ever-expanding pool of molecular data and the limited resources available to understand protein function [25–27].

The first two CAFA challenges were carried out in 2010–2011 [25] and 2013–2014 [26]. In CAFA1, we adopted a time-delayed evaluation method, where protein sequences that lacked experimentally verified annotations, or *targets*, were released for prediction. After the submission deadline for predictions, a subset of these targets accumulated experimental annotations over time, either as a consequence of new publications about these proteins or the biocuration work updating the annotation databases. The members of this set of proteins were used as *benchmarks* for evaluating the participating computational methods, as the function was revealed only after the prediction deadline.

CAFA2 expanded the challenge founded in CAFA1. The expansion included the number of ontologies used for predictions, the number of target and benchmark proteins, and the introduction of new assessment metrics that mitigate the problems with functional similarity calculation over concept hierarchies such as Gene Ontology [28]. Importantly, we provided evidence that the top-scoring methods in CAFA2 outperformed the top-scoring methods in CAFA1, highlighting that methods participating in CAFA improved over the 3-year period. Much of this improvement came as a consequence of novel methodologies with some effect of the expanded annotation databases [26]. Both CAFA1 and CAFA2 have shown that computational methods designed to perform function prediction outperform a conventional function transfer through sequence similarity [25, 26].

In CAFA3 (2016–2017), we continued with all types of evaluations from the first 2 challenges and additionally performed experimental screens to identify genes associated with specific functions. This allowed us to provide unbiased evaluation of the term-centric performance based on a unique set of benchmarks obtained by assaying *Candida albicans*, *Pseudomonas aeruginosa*, and *Drosophila melanogaster*. We also held a challenge following CAFA3, dubbed CAFA-π, to provide the participating teams another opportunity to develop or modify prediction models. The genome-wide screens on *C. albicans* identified 240 genes previously not known to be involved in biofilm formation, whereas the screens on *P. aeruginosa* identified 532 new genes involved in biofilm formation and 403 genes involved in motility. Finally, we used CAFA predictions to select genes from *D. melanogaster* and assay them for long-term memory involvement. This experiment allowed us to both evaluate prediction methods and identify 11 new fly genes involved in this biological process [29]. Here, we present the outcomes of the CAFA3 challenge, as well as the accompanying challenge CAFA-π, and discuss further directions for the community interested in the function of biological macromolecules.

Results

Top methods have improved from CAFA2 to CAFA3, but improvement was less dramatic than from CAFA1 to CAFA2

One of CAFA's major goals is to quantify the progress in function prediction over time. We therefore conducted a comparative evaluation of top CAFA1, CAFA2, and CAFA3 methods according to their ability to predict Gene Ontology [28] terms on a set of common benchmark proteins. This benchmark set was created as an intersection of CAFA3 benchmarks (proteins that gained experimental annotation after the CAFA3 prediction submission deadline) and CAFA1 and CAFA2 target proteins. Overall, this set contained 377 protein sequences with annotations in the Molecular Function Ontology (MFO), 717 sequences in the Biological Process Ontology (BPO), and 548 sequences in the Cellular Component Ontology (CCO), which allowed for a direct comparison of all methods that have participated in the challenges so far. The head-to-head comparisons in MFO, BPO, and CCO between the top 5 CAFA3 and CAFA2 methods are shown in Fig. 1. CAFA3 and CAFA1 comparisons are shown in Additional file 1: Figure S1.

We first observe that, in effect, the performance of baseline methods [25, 26] has not improved since CAFA2. The Naïve method, which uses the term frequency in the existing annotation database as a prediction score for every input protein, has the same F_{max} performance using both annotation databases in 2014 (when CAFA2 was held) and in 2017 (when CAFA3 was held), which suggests little change in term frequencies in the annotation database since 2014. In MFO, the BLAST method based on the existing annotations in 2017 is slightly but significantly better than the BLAST method based on 2014 training data. In BPO and CCO, however, the BLAST based on the later database has not outperformed its earlier counterpart, although the changes in effect size (absolute change in F_{max}) in both ontologies are small.

When surveying all 3 CAFA challenges, the performance of both baseline methods has been relatively stable, with some fluctuations of BLAST. Such performance of direct sequence-based function transfer is surprising, given the steady growth of annotations in UniProt-GOA [30]; that is, there were 259,785 experimental annotations in 2011, 341,938 in 2014, and 434,973 in 2017, but there does not seem to be a definitive trend with the BLAST method, as they go up and down in F_{max} across ontologies. We conclude from these observations on the baseline methods that first, the ontologies are in different annotation states and should not be treated as a whole. In fact, the distribution of annotation depth and information content is very different across 3 ontologies, as shown in Additional file 1: Figures S15 and S16. Second, methods that perform direct function transfer based on sequence similarity do not necessarily benefit from a larger training dataset. Although the performance observed in our work is also dependent on the benchmark set, it appears that the annotation databases remain too sparsely populated to effectively exploit function transfer by sequence similarity, thus justifying the need for advanced methodology development for this problem.

Head-to-head comparisons of the top 5 CAFA3 methods against the top 5 CAFA2 methods show mixed results. In MFO, the top CAFA3 method, GOLabeler [23], outperformed all CAFA2 methods by a considerable margin, as shown in Fig. 2. The rest of the 4 CAFA3 top methods did not perform as well as the top 2 methods of CAFA2, although only to a limited extent, with little change in F_{max}. Of the top 12 methods ranked in MFO, 7 are from CAFA3, 5 are from CAFA2, and none from CAFA1. Despite the increase in database size, the majority of function prediction methods do not seem to have improved in predicting protein function in MFO since 2014, except for 1 method that stood out. In BPO, the top 3 methods in CAFA3 outperformed their CAFA2 counterparts, but with very small margins. Out of the top 12 methods in BPO, 8 are from CAFA3, 4 are from CAFA2, and none from CAFA1. Finally, in CCO, although 8 out of the top 12 methods over all CAFA challenges come from CAFA3, the top method is from CAFA2. The differences between the top-performing methods are small, as in the case of BPO.

The performance of the top methods in CAFA2 was significantly better than of those in CAFA1, and it is interesting to note that this trend has not continued in CAFA3.

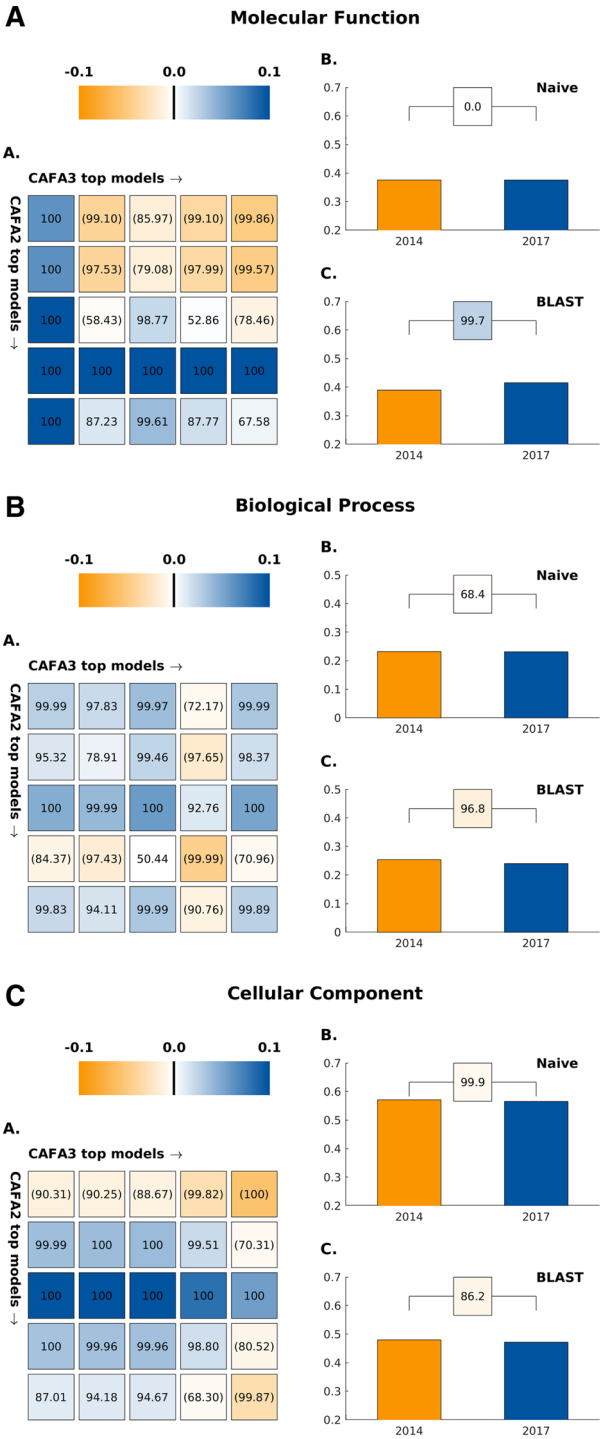

Fig. 1 A comparison in F_{max} between the top 5 CAFA2 models against the top 5 CAFA3 models. Colored boxes encode the results such that (1) the colors indicate margins of a CAFA3 method over a CAFA2 method in F_{max} and (2) the numbers in the box indicate the percentage of wins. **a** CAFA2 top 5 models (rows, from top to bottom) against CAFA3 top 5 models (columns, from left to right). **b** Comparison of the performance (F_{max}) of Naïve baselines trained respectively on SwissProt2014 and SwissProt2017. Colored box between the two bars shows the percentage of wins and margin of wins as in **a**. **c** Comparison of the performance (F_{max}) of BLAST baselines trained on SwissProt2014 and SwissProt2017. Colored box between the two bars shows the percentage of wins and margin of wins as in **a**. Statistical significance was assessed using 10,000 bootstrap samples of benchmark proteins

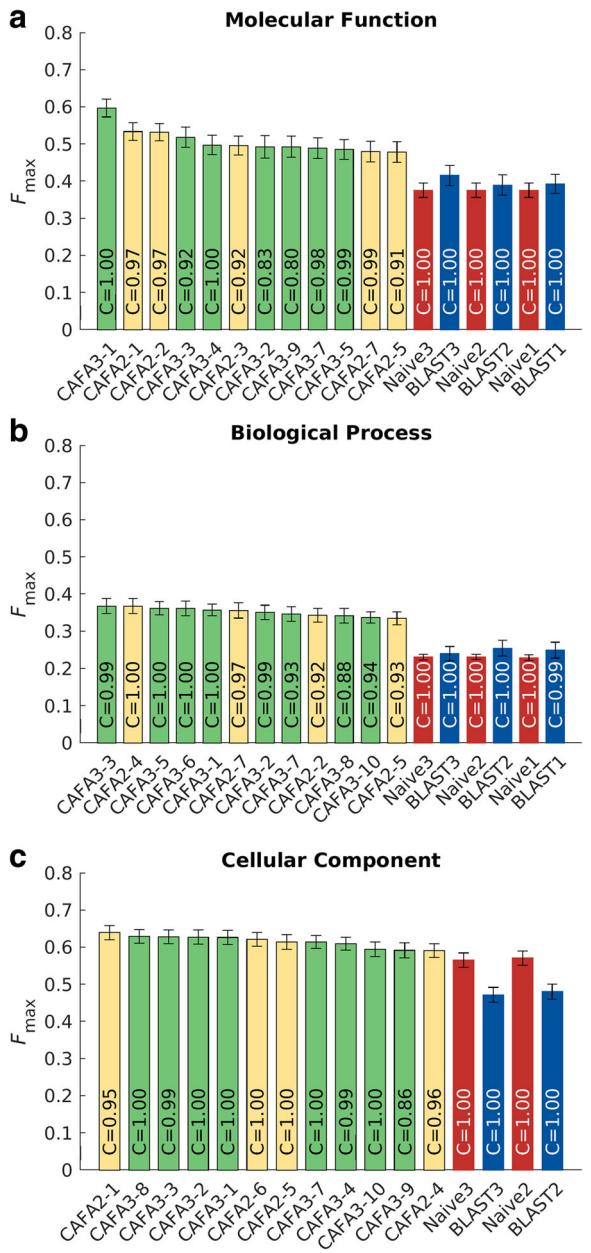

Fig. 2 Performance evaluation based on the F_{max} for the top CAFA1, CAFA2, and CAFA3 methods. The top 12 methods are shown in this barplot ranked in descending order from left to right. The baseline methods are appended to the right; they were trained on training data from 2017, 2014, and 2011, respectively. Coverage of the methods were shown as text inside the bars. Coverage is defined as the percentage of proteins in the benchmark that are predicted by the methods. Color scheme: CAFA2, ivory; CAFA3, green; Naïve, red; BLAST, blue. Note that in MFO and BPO, CAFA1 methods were ranked, but since none made to the top 12 of all 3 CAFA challenges, they were not displayed. The CAFA1 challenge did not collect predictions for CCO. **a**: molecular function; **b**: Biological process; **c**: Cellular Component

This could be due to many reasons, such as the quality of the benchmark sets, the overall quality of the annotation database, the quality of ontologies, or a relatively short period of time between challenges.

Protein-centric evaluation

The *protein-centric* evaluation measures the accuracy of assigning GO terms to a protein. This performance is shown in Figs. 3 and 4.

We observe that all top methods outperform the baselines with the patterns of performance consistent with CAFA1 and CAFA2 findings. Predictions of MFO terms achieved the highest F_{max} compared with predictions in the other two ontologies. BLAST outperforms Naïve in predictions in MFO, but not in BPO or CCO. This is because sequence similarity-based methods such as BLAST tend to perform best when transferring basic biochemical annotations such as enzymatic activity. Functions in biological process, such as pathways, may not be as preserved by sequence similarity, hence the poor BLAST performance in BPO. The reasons behind the difference among the three ontologies include the structure and complexity of the ontology as well as the state of the annotation database, as discussed previously [26, 31]. It is less clear why the performance in CCO is weak, although it might be hypothesized that such performance is related to the structure of the ontology itself [31].

The top-performing method in MFO did not have as high an advantage over others when evaluated using the S_{min} metric. The S_{min} metric weights GO terms by conditional information content, since the prediction of more informative terms is more desirable than less informative, more general, terms. This could potentially explain the smaller gap between the top predictor and the rest of the pack in S_{min}. The weighted F_{max} and normalized S_{min} evaluations can be found in Additional file 1: Figures S4 and S5.

Species-specific categories

The benchmarks in each species were evaluated individually as long as there were at least 15 proteins per species. Here, we present the results from eukaryotic and bacterial species (Fig. 5). We observed that different methods could perform differently on different species. As shown in Fig. 6, bacterial proteins make up a small portion of all benchmark sequences, so their effects on the performances of the methods are often masked. Species-specific analyses are therefore useful to researchers studying certain organisms. Evaluation results on individual species including human (Additional file 1: Figure S6), *Arabidopsis thaliana* (Additional file 1: Figure S7) and *Escherichia coli* (Additional file 1: Figure S10) can be found in Additional file 1: Figure S6-S14.

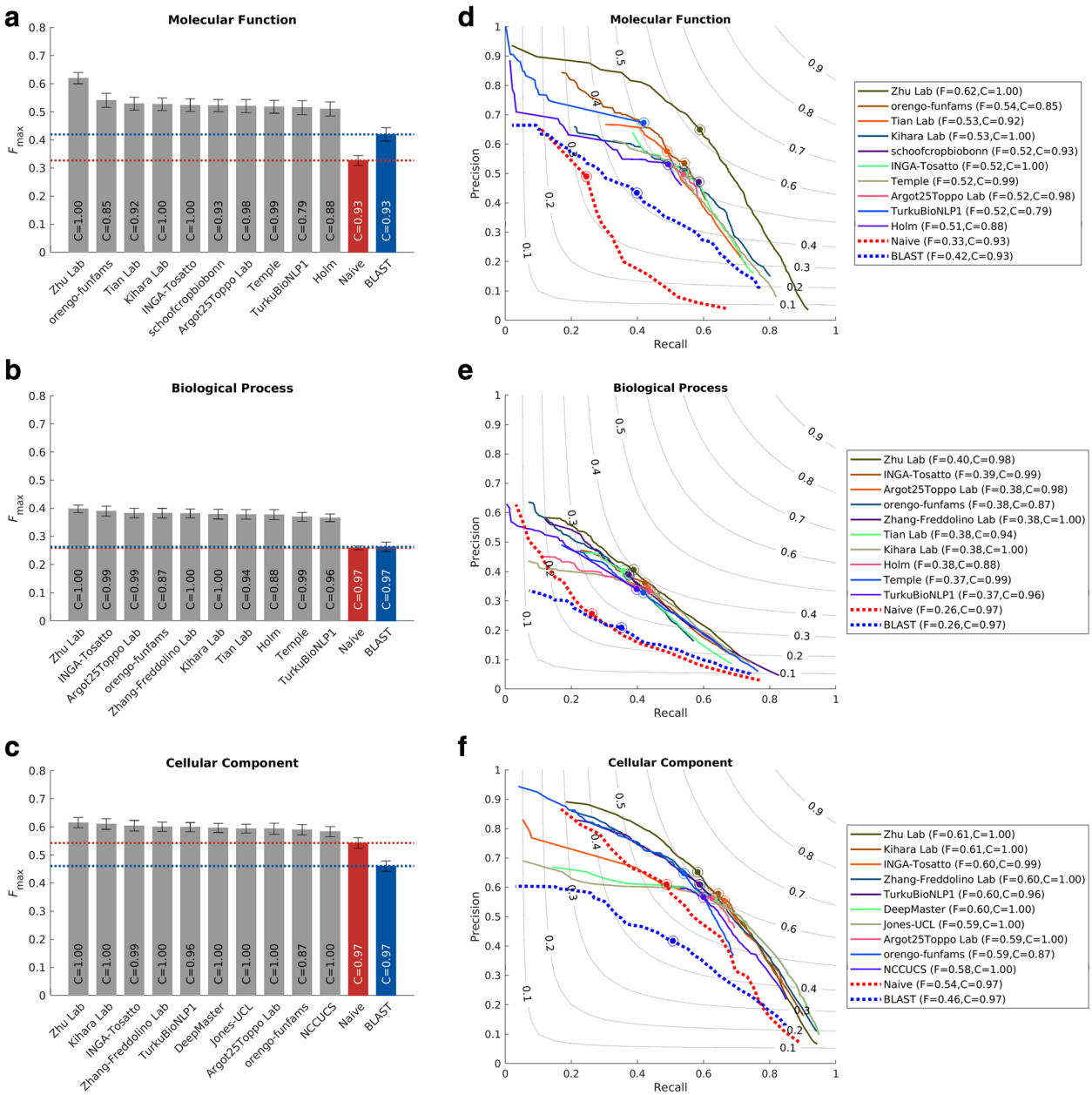

Fig. 3 Performance evaluation based on the F_{max} for the top-performing methods in 3 ontologies. Evaluation was carried out on *No knowledge* benchmarks in the *full* mode. **a**–**c**: bar plots showing the F_{max} of the top 10 methods. The 95% confidence interval was estimated using 10,000 bootstrap iterations on the benchmark set. Coverage of the methods was shown as text inside the bars. Coverage is defined as the percentage of proteins in the benchmark which are predicted by the methods. **d**–**f**: precision-recall curves for the top 10 methods. The perfect prediction should have $F_{max} = 1$, at the top right corner of the plot. The dot on the curve indicates where the maximum F score is achieved

Diversity of methods

It was suggested in the analysis of CAFA2 that ensemble methods that integrate data from different sources have the potential of improving prediction accuracy [32]. Multiple data sources, including sequence, structure, expression profile, genomic context, and molecular interaction data, are all potentially predictive of the function of the protein. Therefore, methods that take advantage of these rich sources as well as existing techniques from other research groups might see improved performance. Indeed, the one method that stood out from the rest in CAFA3 and performed significantly better than all methods across three challenges is a machine learning-based ensemble method [23]. Therefore, it is important to analyze what information sources and prediction algorithms are better at predicting function. Moreover, the similarity

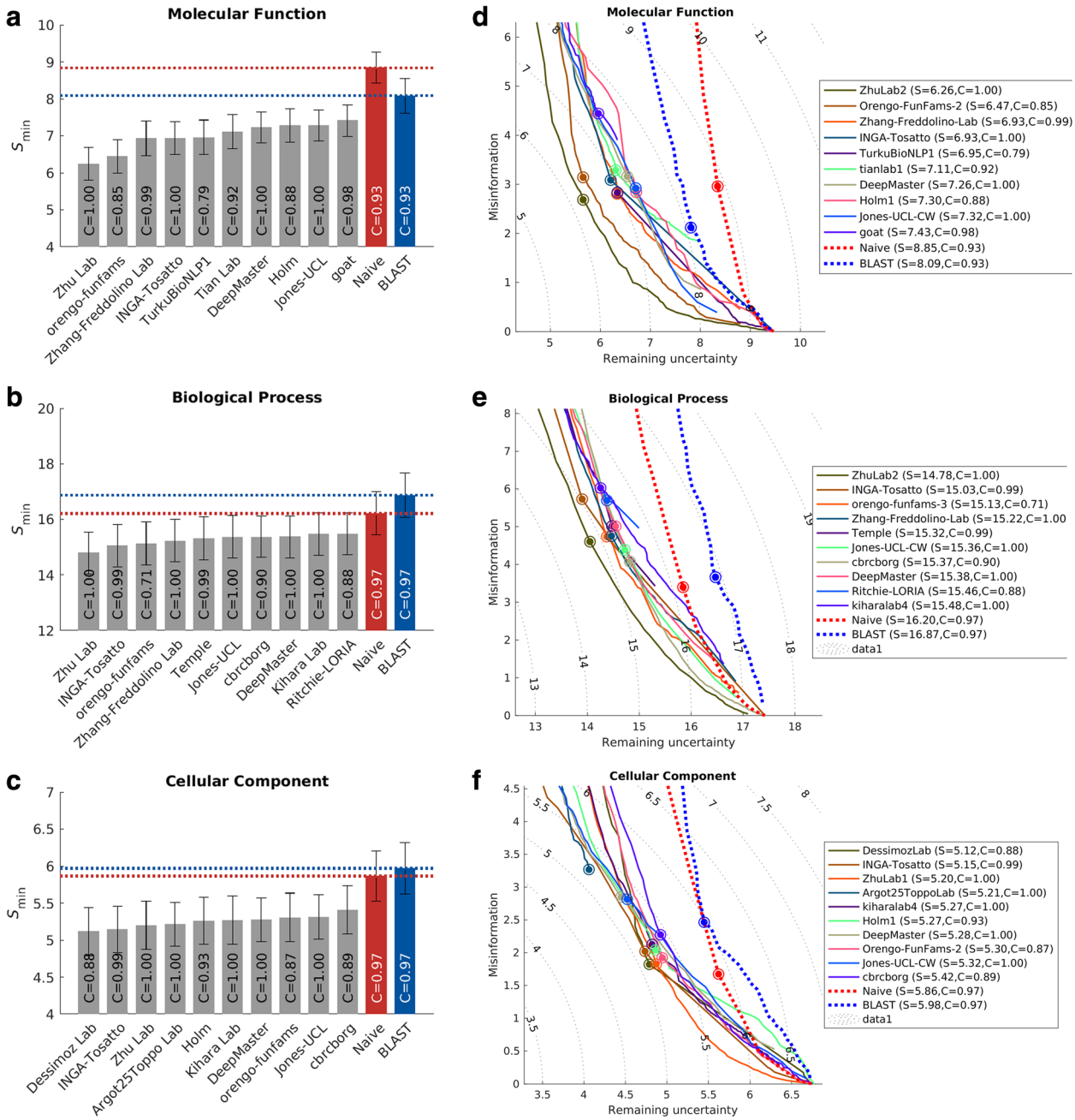

Fig. 4 Performance evaluation based on S_{min} for the top-performing methods in 3 ontologies. Evaluation was carried out on *No knowledge* benchmarks in the *full* mode. **a**–**c**: bar plots showing S_{min} of the top 10 methods. The 95% confidence interval was estimated using 10,000 bootstrap iterations on the benchmark set. Coverage of the methods was shown as text inside the bars. Coverage is defined as the percentage of proteins in the benchmark which are predicted by the methods. **d**–**f**: remaining uncertainty-missing information (RU-MI) curves for the top 10 methods. The perfect prediction should have $S_{min} = 0$, at the bottom left corner of the plot. The dot on the curve indicates where the minimum semantic distance is achieved

of the methods might explain the limited improvement in the rest of the methods in CAFA3.

The top CAFA2 and CAFA3 methods are very similar in performance, but that could be a result of aggregating predictions of different proteins to one metric. When computing the similarity of each pair of methods as the Euclidean distance of prediction scores (Fig. 7), we are not interested whether these predictions are correct according to the benchmarks, but simply whether they are similar to one another. The diagonal blocks in Fig. 7 show that

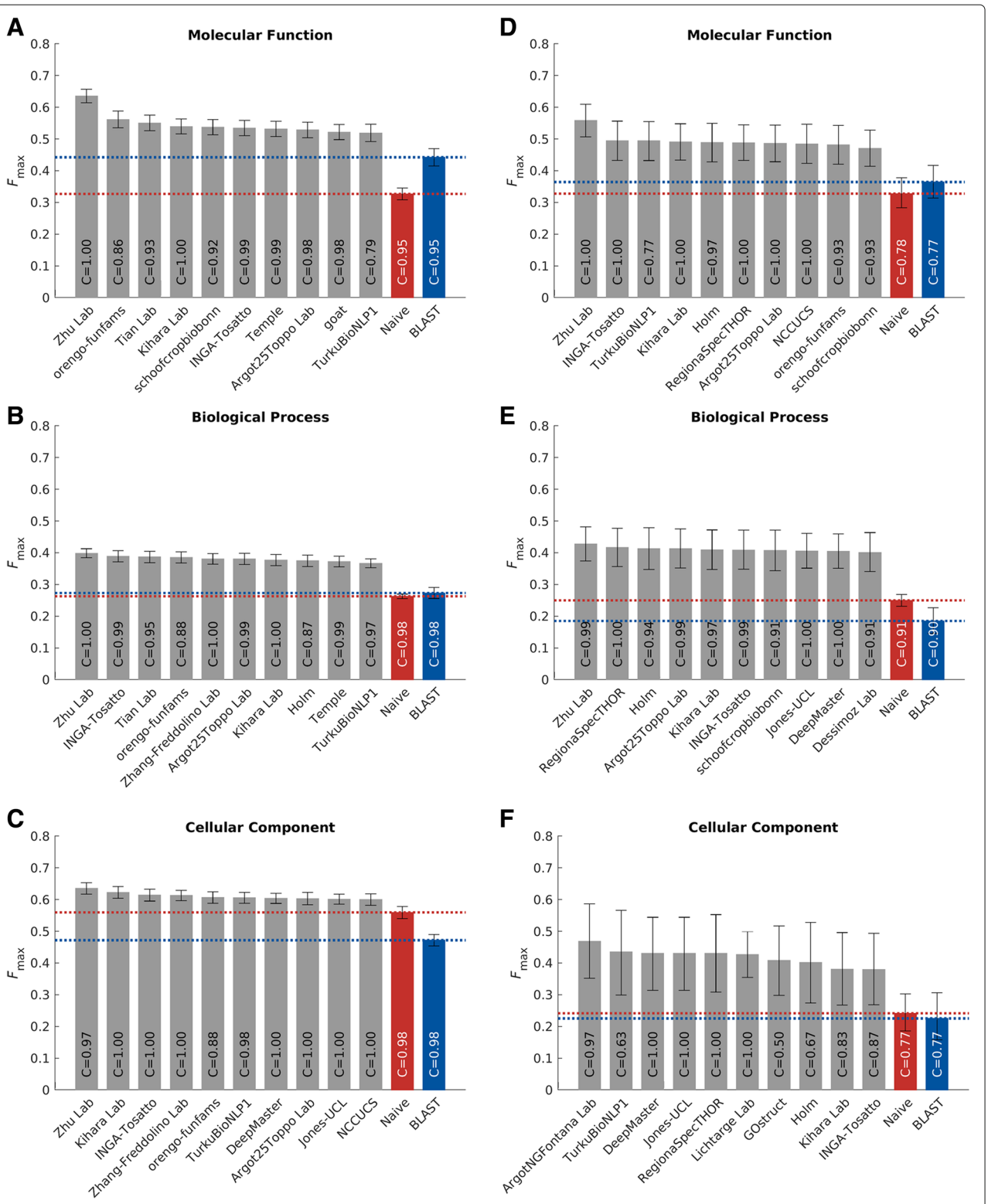

Fig. 5 Evaluation based on the F_{max} for the top-performing methods in eukaryotic and bacterial species

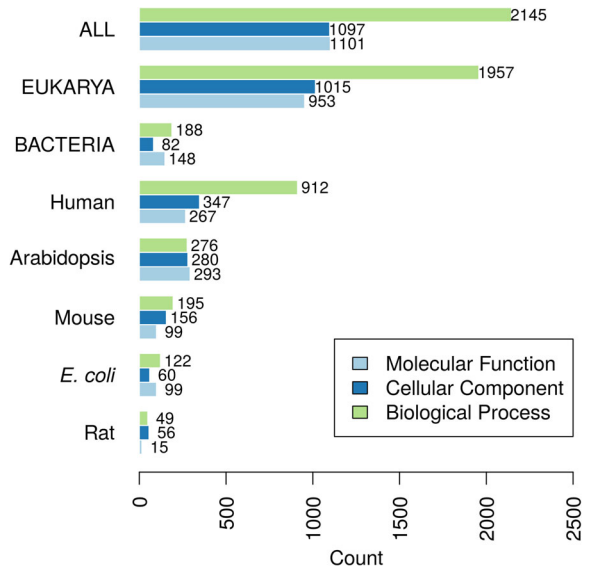

Fig. 6 Number of proteins in each benchmark species and ontology

CAFA1 top methods are more diverse than CAFA2 and CAFA3. The off-diagonal blocks shows that CAFA2 and CAFA3 methods are more similar with each other than with CAFA3 methods. It is clear that some methods are heavily based on the Naïve and BLAST baseline methods.

Participating teams also provided keywords that describe their approach to function prediction with their submissions. A list of keywords was given to the participants, listed in Additional file 1. Figure 8 shows the frequency of each keyword. In addition, we have weighted the frequency of the keywords with the prediction accuracy of the specific method. Machine learning and sequence alignment remain the most used approach by scientists predicting in all three ontologies. By raw count, machine learning is more popular than sequence in all three ontologies, but once adjusted by performance, their difference shrinks. In MFO, sequence alignment even overtakes machine learning as the most popular keyword after adjusting for performance. This indicates that methods that use sequence alignments are more helpful in predicting the correct function than the popularity of their use suggests.

Evaluation via molecular screening

Databases with proteins annotated by biocuration, such as UniProt knowledge base and UniProt Gene Ontology Annotation (GOA) database, have been the primary source of benchmarks in the CAFA challenges. New to CAFA3, we also evaluated the extent to which methods participating in CAFA could predict the results of genetic screens in model organisms done specifically for this project. Predicting GO terms for a protein (protein-centric) and predicting which proteins are associated with a given function (term-centric) are related but different computational problems: the former is a multi-label classification problem with a structured output, while the latter is a binary classification task. Predicting the results of a genome-wide screen for a single or a small number of functions fits the term-centric formulation. To see how well all participating CAFA methods perform term-centric predictions, we mapped the results from the protein-centric CAFA3 methods onto these terms. In addition, we held a separate CAFA challenge, CAFA-π, whose purpose was to attract additional submissions from algorithms that specialize in term-centric tasks.

We performed screens for three functions in three species, which we then used to assess protein function prediction. In the bacterium *Pseudomonas aeruginosa* and the fungus *Candida albicans*, we performed genome-wide screens capable of uncovering genes with two functions, biofilm formation (GO:0042710) and motility (for *P. aeruginosa* only) (GO:0001539), as described in the "Methods" section. In *Drosophila melanogaster*, we performed targeted assays, guided by previous CAFA submissions, of a selected set of genes and assessed whether or not they affected long-term memory (GO:0007616).

We discuss the prediction results for each function below in detail. The performance, as assessed by the genome-wide screens, was generally lower than in the protein-centric evaluations that were curation driven. We hypothesize that it may simply be more difficult to perform term-centric prediction for broad activities such as biofilm formation and motility. For *P. aeruginosa*, an existing compendium of gene expression data was already available [33]. We used the Pearson correlation over this collection of data to provide a complementary baseline to the standard BLAST approach used throughout CAFA. We found that an expression-based method outperformed the CAFA participants, suggesting that success on certain term-centric challenges will require the use of different types of data. On the other hand, the performance of the methods in predicting long-term memory in the Drosophila genome was relatively accurate.

Biofilm formation

In March 2018, there were 3019 annotations to biofilm formation (GO:0042710) and its descendent terms across all species, of which 325 used experimental evidence codes. These experimentally annotated proteins included 131 from the Candida Genome Database [34] for *C. albicans* and 29 for *P. aeruginosa*, the 2 organisms that we screened.

Of the 2746 genes we screened in the *Candida albicans* colony biofilm assay, 245 were required for the formation of wrinkled colony biofilm formation (Table 1). Of these, only 5 were already annotated in UniProt: *MOB*,

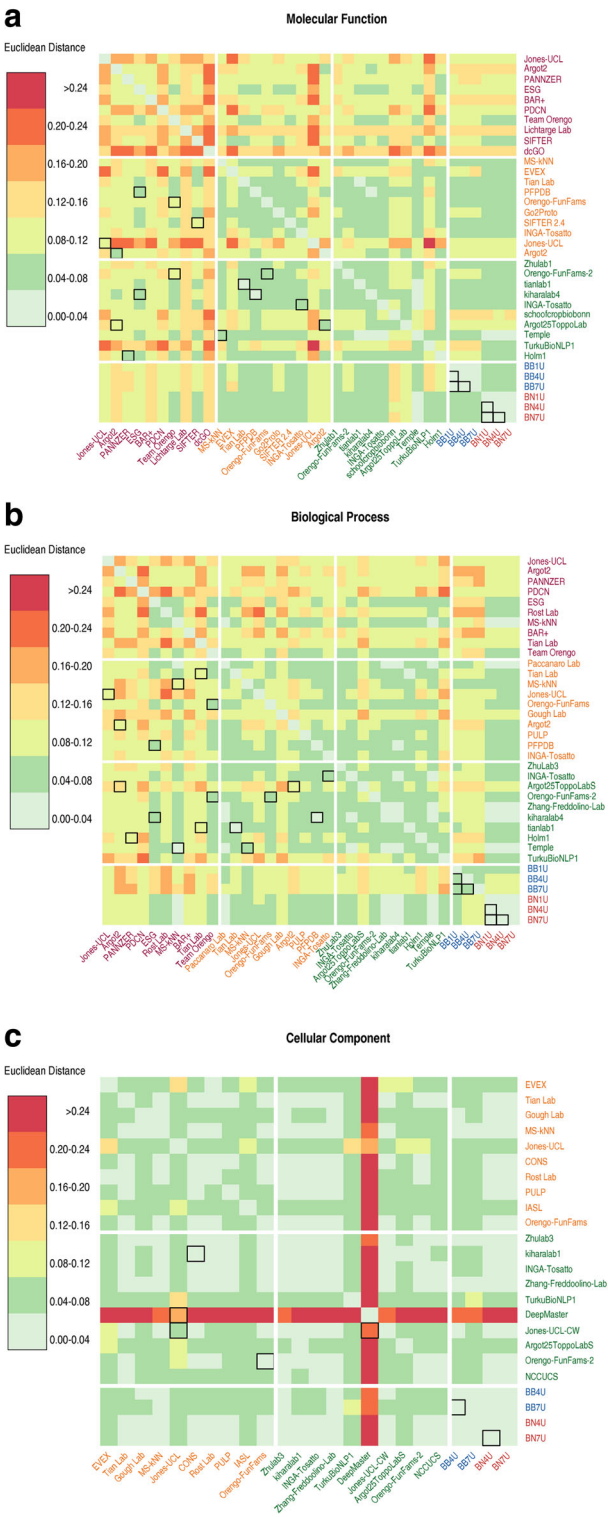

Fig. 7 Heatmap of similarity for the top 10 methods in CAFA1, CAFA2, and CAFA3. Similarity is represented by Euclidean distance of the prediction scores from each pair of methods, using the intersection set of benchmarks in the "Top methods have improved from CAFA2 to CAFA3, but improvement was less dramatic than from CAFA1 to CAFA2" section. The higher (darker red color) the euclidean distance, the less similar the methods are. Top 10 methods from each of the CAFA challenges are displayed and ranked by their performance in F_{max}. Cells highlighted by black borders are between a pair of methods that come from the same PI. **a**: Molecular Function; **b**: Biological Process; **c**: Cellular Component

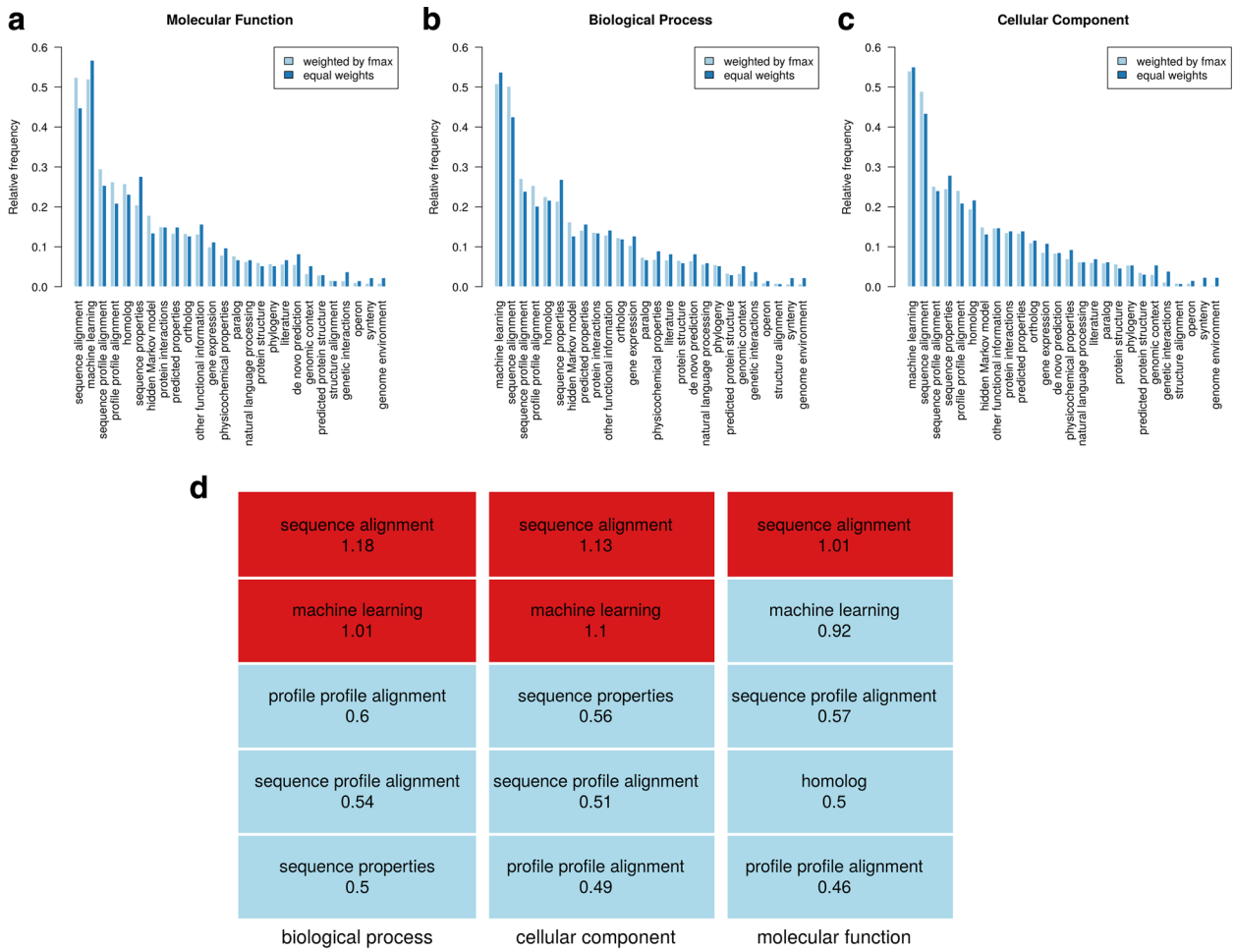

Fig. 8 Keyword analysis of all CAFA3 participating methods. **a**–**c**: both relative frequency of the keywords and weighted frequencies are provided for three respective GO ontologies. The weighted frequencies accounts for the performance of the the particular model using the given keyword. If that model performs well (with high F_{max}), then it gives more weight to the calculation of the total weighted average of that keyword. **d** shows the ratio of relative frequency between the F_{max}-weighted and equal-weighted. Red indicates the ratio is greater than one while blue indicates the ratio is less than one. Only the top five keywords ranked by ratio are shown. The larger the ratio, the more difference there is between the F_{max}-weighted and the equal-weighted

EED1 (*DEF1*), and *YAK1*, which encode proteins involved in hyphal growth, an important trait for biofilm formation [35–38]. Also, *NUP85*, a nuclear pore protein involved in early phase arrest of biofilm formation [39] and *VPS1*, contributes to protease secretion, filamentation, and biofilm formation [40]. Of the 2063 proteins that we did not find to be associated with biofilm formation, 29 were annotated with the term in the GOA database in *C. albicans*. Some of the proteins in this category highlight the need for additional information to GO term annotation. For example, Wor1 and the pheromone receptor are key for biofilm formation in strains under conditions in which the mating pheromone is produced [41], but not required in the monocultures of the commonly studied a/α mating type strain used here.

Table 1 Number of proteins in *Candida albicans* and *Pseudomonas aeruginosa* associated with the GO term "Biofilm formation" (GO:0042710) in the GOA databases versus experimental results

			GOA annotations	
C. albicans	Total, 2308		Unannotated	Annotated
	CAFA experiments	False	2034	29
		True	240	5
P. aeruginosa	Total, 4056		Unannotated	Annotated
	CAFA experiments	False	3491	25
		True	532	9

We used receiver operating characteristic (ROC) curves to measure the prediction accuracy. Area under ROC curves (AUROC) was used to compare the performance. AUROC is a common accuracy measure for classification problems where it evaluates how good a model is at distinguishing between the positive and negative classes. No method in CAFA-π or CAFA3 (not shown) exceeded an AUC of 0.60 on this term-centric challenge (Fig. 9) for either species. Performance for the best methods slightly exceeded a BLAST-based baselines. In the past, we have found that predicting BPO terms, such as biofilm formation, resulted in poorer method performance than predicting MFO terms. Many CAFA methods use sequence alignment as their primary source of information (the "Diversity of methods" section). For *Pseudomonas aeruginosa*, a pre-built expression compendium was available from prior work [33]. Where the compendium was available, simple gene expression-based baselines were the best-performing approaches. This suggests that successful term-centric prediction of biological processes may need to rely more heavily on information that is not sequence-based and, as previously reported, may require methods that use broad collections of gene expression data [42, 43].

Motility

In March 2018, there were 302,121 annotations for proteins with the GO term: cilium or flagellum-dependent cell motility (GO:0001539) and its descendent terms, which included cell motility in all eukaryotic (GO:0060285), bacterial (GO:0071973), and archaeal (GO:0097590) organisms. Of these, 187 had experimental

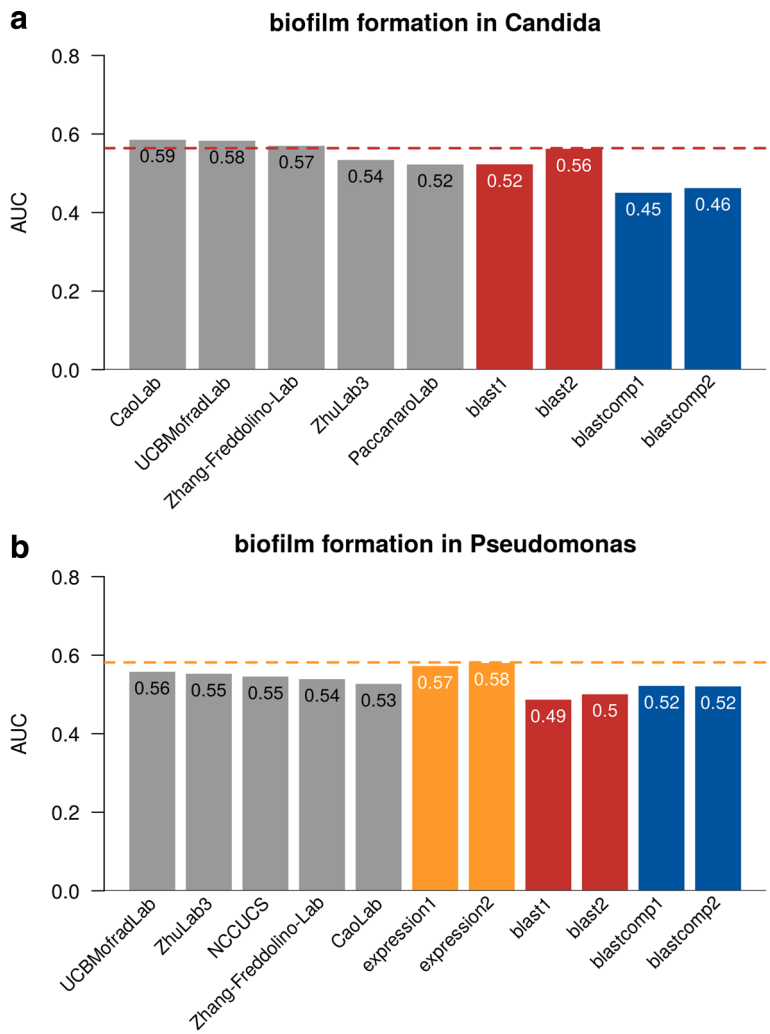

Fig. 9 AUROC of the top five teams in CAFA-π. The best-performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. Four baseline models all based on BLAST were computed for *Candida*, while six baseline models were computed for *Pseudomonas*, including two based on expression profiles. All team methods are in gray while BLAST methods are in red, BLAST computational methods are in blue, and expression are in yellow, see Table 3 for the description of the baselines

evidence codes, and the most common organism to have annotations was *P. aeruginosa*, on which our screen was performed (Additional file 1: Table S2).

As expected, mutants defective in the flagellum or its motor were defective in motility (*fliC* and other *fli* and *flg* genes). For some of the genes that were expected, but not detected, the annotation was based on the experiments performed in a medium different from what was used in these assays. For example, PhoB regulates motility but only when phosphate concentration is low [44]. Among the genes that were scored as defective in motility, some are known to have decreased motility due to over production of carbohydrate matrix material (*bifA*) [45], or the absence of directional swimming due to absence of chemotaxis functions (e.g., *cheW*, *cheA*) and others likely showed this phenotype because of a medium-specific requirement such as biotin (*bioA*, *bioC*, and *bioD*) [46]. Table 2 shows the contingency table for the number of proteins that are detected by our experiment versus GOA annotations.

The results from this evaluation were consistent with what we observed for biofilm formation. Many of the genes annotated as being involved in biofilm formation were identified in the screen. Others that were annotated as being involved in biofilm formation did not show up in the screen because the strain background used here, strain PA14, uses the exopolysaccharide matrix carbohydrate Pel [47] in contrast to the Psl carbohydrate used by another well-characterized strain, strain PAO1 [48, 49]. The *psl* genes were known to be dispensable for biofilm formation in the strain PA14 background, and this nuance highlights the need for more information to be taken into account when making predictions.

The CAFA-π methods outperformed our BLAST-based baselines but failed to outperform the expression-based baselines. Transferred methods from CAFA3 also did not outperform these baselines. It is important to note this consistency across terms, reinforcing the finding that term-centric prediction of biological processes is likely to require non-sequence information to be included.

Long-term memory in D. melanogaster

Prior to our experiments, there were 1901 annotations made in the long-term memory, including 283 experimental annotations. *Drosophila melanogaster* had the most annotated proteins of long-term memory with 217, while human has 7, as shown in Additional file 1: Table S3.

We performed RNAi experiments in *Drosophila melanogaster* to assess whether 29 target genes were associated with long-term memory (GO:0007616). Briefly, flies were exposed to wasps, which triggers a behavior that causes females to lay fewer eggs. The acute response is measured until 24 h post-exposure, and the long-term response is measured at 24 to 48 h post-exposure. RNAi was used to interfere with the expression of the 29 target genes in the mushroom body, a region of the fly brain associated with memory. Using this assay, we identified 3 genes involved in the perception of wasp exposure and 12 genes involved in the long-term memory. For details on the target selection and fly assay, see [29]. None of the 29 genes had an existing annotation in the GOA database. Because no genome-wide screen results were available, we did not release this as part of the CAFA-π and instead relied only on the transfer of methods that predicted the "long-term memory" at least once in *D. melanogaster* from CAFA3. Results from this assessment were more promising than our findings from the genome-wide screens in microbes (Fig. 10). Certain methods performed well, substantially exceeding the baselines.

Table 2 Number of proteins in *Pseudomonas aeruginosa* associated with function motility (GO:0001539) in the GOA databases versus experimental results

		GOA annotations	
Total, 3630		Unannotated	Annotated
CAFA experiments	False	3195	12
	True	403	21

Participation growth

The CAFA challenge has seen growth in participation, as shown in Fig. 11. To cope with the increasingly large data size, CAFA3 utilized the Synapse [50] online platform for submission. Synapse allowed for easier access for participants, as well as easier data collection for the organizers. The results were also released to the individual teams via this online platform. During the submission process, the online platform also allows for customized format checkers to ensure the quality of the submission.

Methods

Benchmark collection

In CAFA3, we adopted the same benchmark generation methods as CAFA1 and CAFA2, with a similar timeline (Fig. 12). The crux of a time-delayed challenge is the annotation growth period between time t_0 and t_1. All target proteins that have gained experimental annotation during this period are taken as benchmarks in all three ontologies. "No knowledge" (NK, no prior experimental annotations) and "Limited knowledge" (LK, partial prior experimental annotations) benchmarks were also distinguished based on whether the newly gained experimental annotation is in an ontology that already have experimental annotations or not. Evaluation results in Figs. 3 and 4 are made using the No knowledge benchmarks. Evaluation results on the Limited knowledge benchmarks are

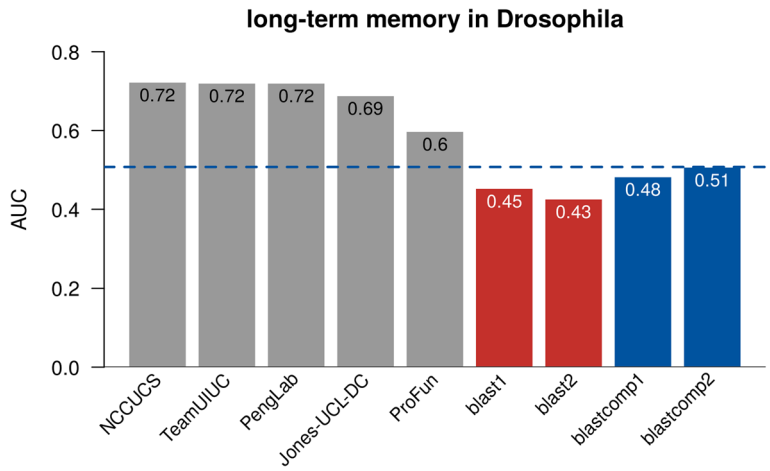

Fig. 10 AUROC of top five teams in CAFA3. The best-performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. All team methods are in gray while BLAST methods are in red and BLAST computational methods are in blue, see Table 3 for the description of the baselines

shown in Additional file 1: Figure S3. For more information regarding NK and LK designations, please refer to the Additional file 1 and the CAFA2 paper [26].

After collecting these benchmarks, we performed two major deletions from the benchmark data. Upon inspecting the taxonomic distribution of the benchmarks, we noticed a large number of new experimental annotations from *Candida albicans*. After consulting with UniProt-GOA, we determined these annotations have already existed in the Candida Genome Database long before 2018 but were only recently migrated to GOA. Since these annotations were already in the public domain before the CAFA3 submission deadline, we have deleted any annotation from *Candida albicans* with an assigned date prior to our CAFA3 submission deadline. Another major change is the deletion of any proteins with only a protein-binding (GO:0005515) annotation. Protein binding is a highly generalized function description, does not provide more specific information about the actual function of a protein, and in many cases may indicate a non-functional, non-specific binding. If it is the only annotation that a protein has gained, then it is hardly an advance in our understanding of that protein; therefore, we deleted these annotations from our benchmark set. Annotations with a depth of 3 make up almost half of all annotations in MFO before the removal (Additional file 1: Figure S15B). After the removal, the most frequent annotations became of depth 5 (Additional file 1: Figure S15A). In BPO, the most frequent annotations are of depth 5 or more, indicating a healthy increase of specific GO terms being added to our annotation database. In CCO, however, most new annotations in our benchmark set are of depths 3, 4, and 5 (Additional file 1: Figure S15). This difference could partially explain why the same computational methods perform very differently in different ontologies and benchmark sets. We have also calculated the total

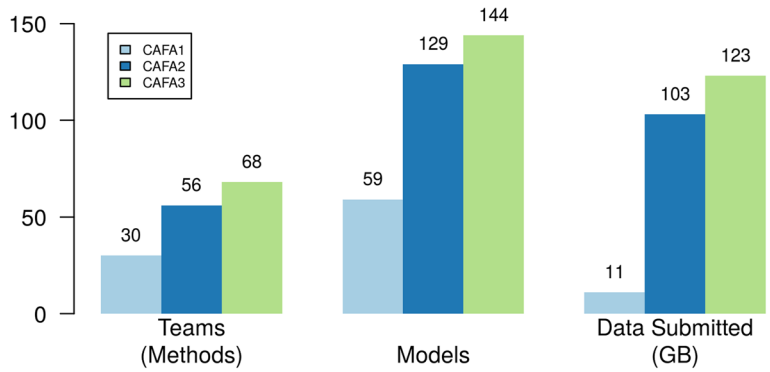

Fig. 11 CAFA participation has been growing. Each principal investigator is allowed to head multiple teams, but each member can only belong to one team. Each team can submit up to three models

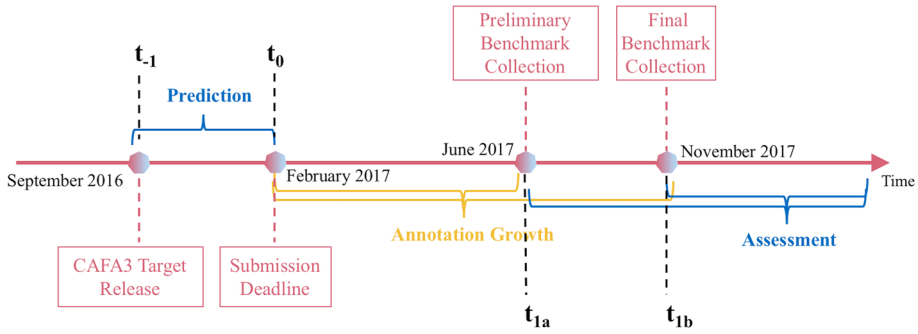

Fig. 12 CAFA3 timeline

information content per protein for the benchmark sets shown in Additional file 1: Figure S16. Taxonomic distributions of the proteins in our final benchmark set are shown in Fig. 6.

Additional analyses were performed to assess the characteristics of the benchmark set, including the overall information content of the terms being annotated.

Protein-centric evaluation

Two main evaluation metrics were used in CAFA3, the F_{max} and the S_{min}. The F_{max} based on the precision-recall curve (Fig. 3), while the S_{min} is based on the remaining uncertainty/missing information (RU-MI) curve as described in [51] (Fig. 4), where S stands for semantic distance. The shortest semantic distance across all thresholds is used as the S_{min} metric. The RU-MI curve takes into account the information content of each GO term in addition to counting the number of true positives, false positives, etc., see Additional file 1 for the precise definition of F_{max} and S_{min}. The information theory-based evaluation metrics counter the high-throughput low-information annotations such as protein binding, but down-weighing these terms according to their information content, as the ability to predict such non-specific functions are not as desirable and useful and the ability to predict more specific functions.

The two assessment modes from CAFA2 were also used in CAFA3. In the partial mode, predictions were evaluated only on those benchmarks for which a model made at least one prediction. The full evaluation mode evaluates all benchmark proteins, and methods were penalized for not making predictions. Evaluation results in Figs. 3 and 4 are made using the full evaluation mode. Evaluation results using the partial mode are shown in Additional file 1: Figure S2.

Two baseline models were also computed for these evaluations. The Naïve method assigns the term frequency as the prediction score for any protein, regardless of any protein-specific properties. BLAST was based on the results using the Basic Local Alignment Search Tool (BLAST) software against the training database [52]. A term will be predicted as the highest local alignment sequence identity among all BLAST hits annotated from the training database. Both of these methods were trained on the experimentally annotated proteins and their sequences in Swiss-Prot [53] at time t_0.

Microbe screens

To assess the matrix production, we used mutants from the PA14 NR collection [54]. Mutants were transferred from the − 80 °C freezer stock using a sterile 48-pin multiprong device into 200 μl LB in a 96-well plate. The cultures were incubated overnight at 37 °C, and their OD600 was measured to assess growth. Mutants were then transferred to tryptone agar with 15 g of tryptone and 15 g of agar in 1L amended with Congo red (Aldrich, 860956) and Coomassie brilliant blue (J.T. Baker Chemical Co., F789-3). Plates were incubated at 37 °C overnight followed by 4-day incubation at room temperature to allow the wrinkly phenotype to develop. Colonies were imaged and scored on day 5. To assess motility, mutants were revived from freezer stocks as described above. After overnight growth, a sterile 48-pin multiprong transfer device with a pin diameter of 1.58 mm was used to stamp the mutants from the overnight plates into the center of swim agar made with M63 medium with 0.2% glucose and casamino acids and 0.3% agar). Care was taken to avoid touching the bottom of the plate. Swim plates were incubated at room temperature (19–22 °C) for approximately 17 h before imaging and scoring. Experimental procedures in *P. aeruginosa* to determine proteins that are associated with the two functions in CAFA-π are shown in Fig. 13.

Biofilm formation in *Candida albicans* was assessed in single gene mutants from the Noble [55] and GRACE [56] collections. In the Noble Collection, mutants of *C. albicans* have had both copies of the candidate gene deleted. Most of the mutants were created in biological duplicate. From this collection, 1274 strains corresponding to 653 unique genes were screened. The GRACE Collection

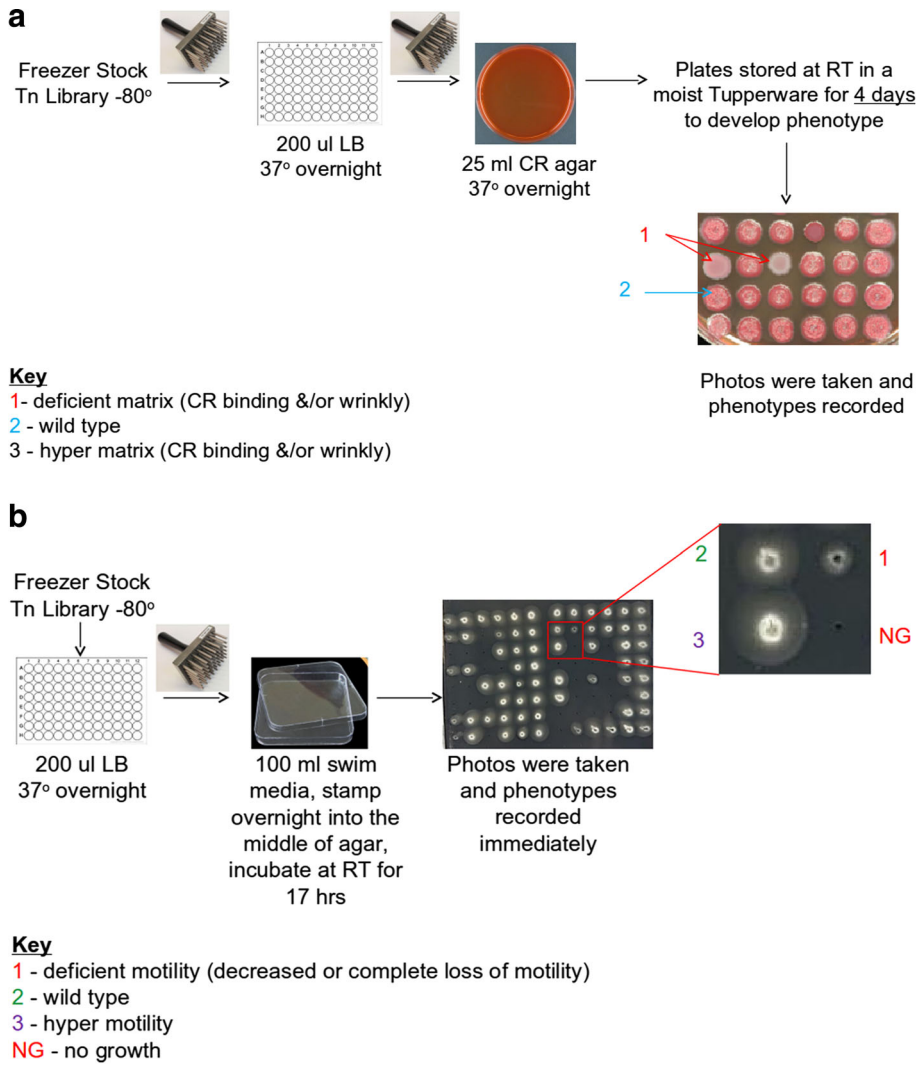

Fig. 13 Experimental procedure of determining genes associated with the functions biofilm formation (**a**) and motility (**b**) in *P. aeruginosa*

provided mutants with one copy of each gene deleted and the other copy placed under the control of a doxycycline-repressible promoter. To assay these strains, we used a medium supplemented with 100 μg/ml doxycycline strains, when rendered them functional null mutants. We screened 2348 mutants from the GRACE Collection, 255 of which overlapped with mutants in the Noble Collection, for 2746 total unique mutants screened in total. To assess the defects in biofilm formation or biofilm-related traits, we performed 2 assays: (1) colony morphology on agar medium and (2) biofilm formation on a plastic surface (Fig. 14). For both of these assays, we used Spider medium, which was designed to induce hyphal growth in *C. albicans* [57] and which promotes biofilm formation [39]. Strains were first replicated from frozen 96-well plates to YPD agar plates. Strains were then replicated from YPD agar to YPD broth and grown overnight at 30 °C. From YPD broth, strains were introduced onto Spider agar plates and into 96-well plates of Spider broth. When strains from the GRACE Collection were assayed, 100 μg/ml doxycycline was included in the agar and broth, and aluminum foil was used to protect the media from light. Spider agar plates inoculated with *C. albicans* mutants were incubated at 37 °C for 2 days before colony morphologies were scored. Strains in Spider broth were shaken at 225 rpm at 37 °C for 3 days and then assayed for biofilm formation at the air-liquid interface as follows. First, broth was removed by slowly tilting the plates and pulling the liquid away by running a gloved hand over the surface. Biofilms were stained by adding 100 μl of 0.1 percent crystal violet dye in water to each well of the plate. After 15 min, plates were gently washed in three

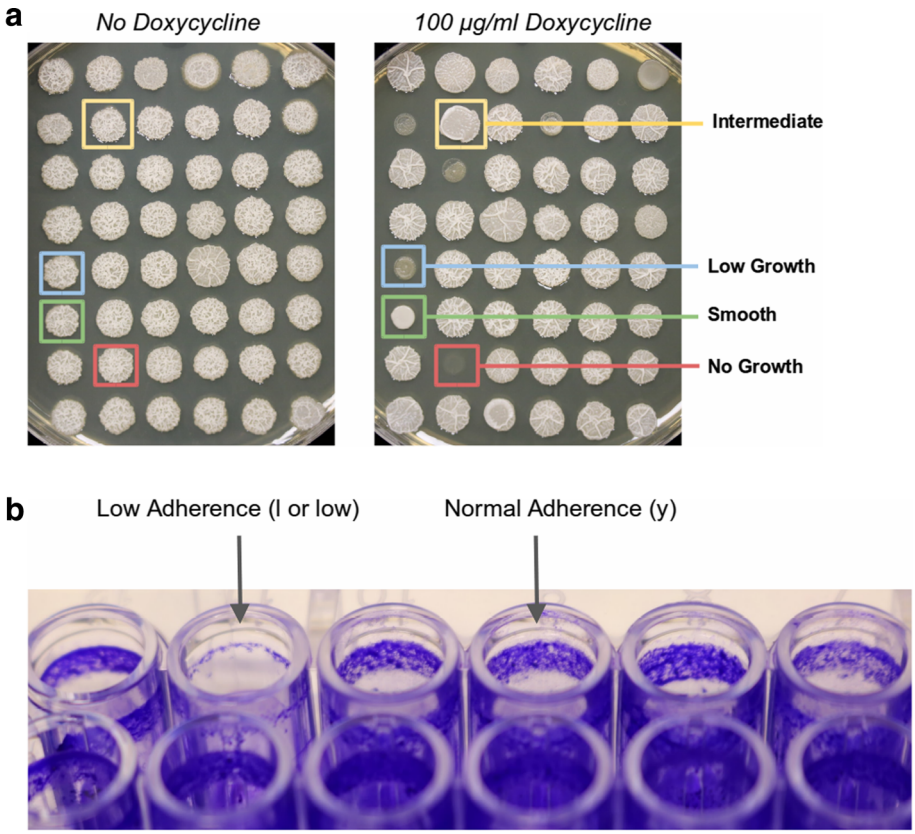

Fig. 14 a: different phenotypes in response to doxycycline treatment: low growth, smooth, no growth and intermediate. **b**: adherence phenotypes. See text for details

baths of water to remove dye without disturbing biofilms. To score biofilm formation for agar plates, colonies were scored by eye as either smooth, intermediate, or wrinkled. A wild-type colony would score wrinkled, and mutants with intermediate or smooth appearance were considered defective in colony biofilm formation. For biofilm formation on a plastic surface, the presence of a ring of cell material in the well indicated normal biofilm formation, while low or no ring formation mutants were considered defective. Genes whose mutations resulted defects in both or either assay were considered true for biofilm function. A complete list of the mutants identified in the screens is available in Additional file 1: Table S1.

A protein is considered true in the biofilm function, if its mutant phenotype is smooth or intermediate under doxycycline.

Term-centric evaluation

The evaluations of the CAFA-π methods were based on the experimental results in the "Microbe screens" section. We adopted F_{max} based on both precision-recall curves and area under ROC curves. There are a total of six baseline methods, as described in Table 3.

Discussion

Since 2010, the CAFA community has been the home to a growing group of scientists across the globe sharing the goal of improving computational function prediction. CAFA has been advancing this goal in three ways. First, through independent evaluation of computational methods against the set of benchmark proteins, thus providing a direct comparison of the methods' reliability and performance at a given time point. Second, the challenge assesses the quality of the current state of the annotations, whether they are made computationally or not, and is set up to reliably track it over time. Finally, as described in this work, CAFA has started to drive the creation of new experimental annotations by facilitating synergies between different groups of researchers interested in function of biological macromolecules. These annotations not only represent new biological discoveries, but simultaneously serve to provide benchmark data for rigorous method evaluation.

CAFA3 and CAFA-π feature the latest advances in the CAFA series to create advanced and accurate methods for protein function prediction. We use the repeated nature of the CAFA project to identify certain trends

Table 3 Baseline methods in term-centric evaluation of protein function prediction

	Model number	Training data	Score assignment
Expression	1	Gene expression compendium for *P. aeruginosa PAO1*	Highest correlation score out of all pairwise correlations
	2		Top 10 average correlation score
BLAST	1	All experimental annotation in UniProt-GOA. Sequences from Swiss-Prot	
	2	All experimental annotation in UniProt-GOA. Sequences from Swiss-Prot and TrEMBL	Highest sequence identity out of all pairwise BLASTp hits
blastcomp	1	All experimental and computational annotations in UniProt-GOA. Sequences from Swiss-Prot	
	2	All experimental and computational annotations in UniProt-GOA. Sequences from Swiss-Prot and TrEMBL	

via historical assessments. The analysis revealed that the performance of CAFA methods improved dramatically between CAFA1 and CAFA2. However, the protein-centric results for CAFA3 are mixed when compared to historical methods. Though the best-performing CAFA3 method outperformed the top CAFA2 methods (Fig. 1), this was not consistently true for other rankings. Among all 3 CAFA challenges, CAFA2 and CAFA3 methods inhabit the top 12 places in MFO and BPO. Between CAFA2 and CAFA3, the performance increase is more subtle. Based on the annotations of methods (Additional file 1), many of the top-ranking methods are improved versions of the methods that have been evaluated in CAFA2. Interestingly, the top-performing CAFA3 method, which consistently outperformed the methods from all past CAFAs in the major categories (GOLabeler [23]), utilized 5 component classifiers trained from different features; those included GO term frequency, sequence alignment, amino acid trigram, domains, motifs, and biophysical properties. It performs best in the Molecular Function Ontology, where sequence features perform best. Another method which did not participate in CAFA3 yet seems to perform well under CAFA parameters is NetGO [58], which utilizes the information from STRING, a network association database [59] in addition to sequence information. Taken together, the strong predictive performance of mRNA co-expression data (Figs. 9 and 15) leads us to hypothesize that including more varied sources of data can lead to additional large improvements in protein function prediction. We are looking forward to testing this hypothesis in future CAFA challenges. It should be noted that CAFA uses both $F_{\max}$ and $S_{\min}$.

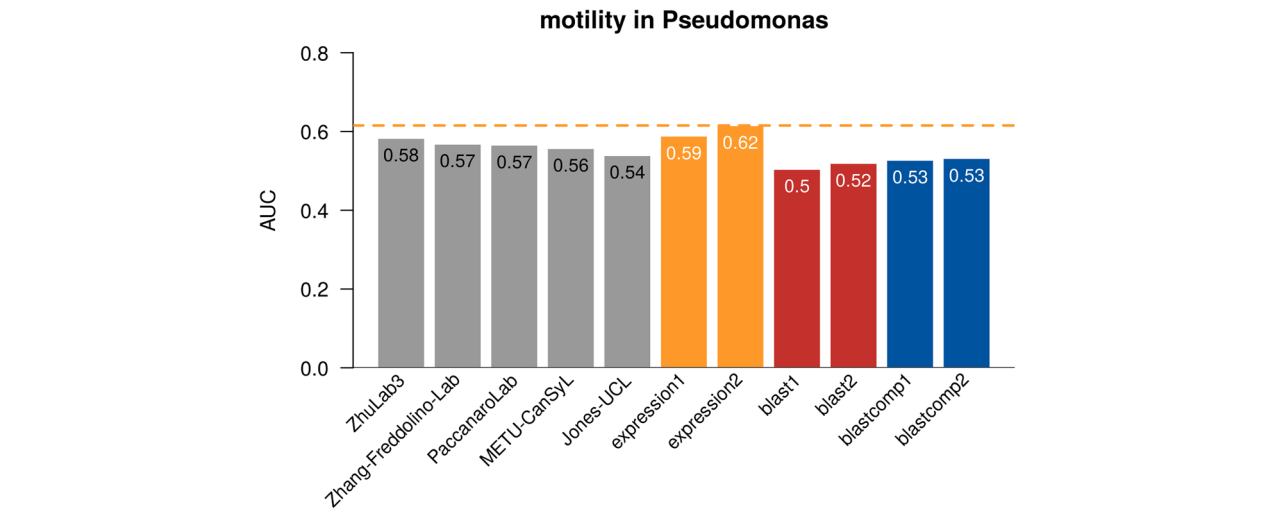

Fig. 15 AUROC of top 5 teams in CAFA-π. The best-performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. All team methods are in gray while BLAST methods are in red, BLAST computational methods are in blue and expression are in yellow. See Table 3 for description of the baselines

F_{max}'s strength lies in its interpretability, as it is simply the maximum F_1 given for each model. At the same time, precision/recall-based assessment does not capture the hierarchical nature of ontologies or the differences in information content between different GO terms. For that reason, we also use the S_{min} score which incorporates information content, but is somewhat less interpretable than F_{max} and less robust to the problems of incomplete annotation [60, 61]. Additionally, since the information content of a GO term is derived from its frequency in the corpus [62], it is somewhat malleable depending on the corpus from which it is derived. We therefore use both measures for scoring, to achieve a more comprehensive picture of the models' performance.

For this iteration of CAFA, we performed genome-wide screens of phenotypes in *P. aeruginosa* and *C. albicans* as well as a targeted screen in *D. melanogaster*. This not only allowed us to assess the accuracy with which methods predict genes associated with select biological processes, but also to use CAFA as an additional driver for new biological discovery. Note that high-throughput screening for a single phenotype should be interpreted with caution as the phenotypic effect may be the result of pleiotropy, and the phenotype in question may be expressed as part of a set of other phenotypes. The results of genome-wide screenings typically lack context for the observed phenotypic effects, and each genotype-phenotype association should be examined individually to ascertain how immediate is the phenotypic effect from the seeming genotypic cause.

In sum, our experimental work identified more than a thousand new functional annotations in three highly divergent species. Though all screens have certain limitations, the genome-wide screens also bypass questions of biases in curation. This evaluation provides key insights: CAFA3 methods did not generalize well to selected terms. Because of that, we ran a second effort, CAFA-π, in which participants focused solely on predicting the results of these targeted assays. This targeted effort led to improved performance, suggesting that when the goal is to identify genes associated with a specific phenotype, tuning methods may be required.

For CAFA evaluations, we have included both Naïve and sequence-based (BLAST) baseline methods. For the evaluation of *P. aeruginosa* screen results, we were also able to include a gene expression baseline from a previously published compendium [33]. Intriguingly, the expression-based predictions outperformed the existing methods for this task. In future CAFA efforts, we will include this type of baseline expression-based method across evaluations to continue to assess the extent to which this data modality informs gene function prediction. The results from the CAFA3 effort suggest that gene expression may be particularly important for successfully predicting term-centric biological process annotations.

The primary takeaways from CAFA3 are as follows: (1) genome-wide screens complement annotation-based efforts to provide a richer picture of protein function prediction; (2) the best-performing method was a new method, instead of a light retooling of an existing approach; (3) gene expression, and more broadly, systems data may provide key information to unlocking biological process predictions, and (4) performance of the best methods has continued to improve. The results of the screens released as part of CAFA3 can lead to a re-examination of approaches which we hope will lead to improved performance in CAFA4.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1835-8.

Additional file 1: Additional figures and tables referenced in the article.
Additional file 2: Review History.

Acknowledgements

NZ and IF acknowledge the invaluable input from Michael C Gerten and Shatabdi Sen and all members of the Friedberg Lab for the ongoing support for stimulating discussions.

Review history

The review history is available as Additional file 2.

Authors' contributions

The experiment was designed by IF, PR, CSG, SDM, COD, MJM, and NZ. NZ, YJ, MNH, HNN, AJL, and LD performed the computational analyses. NZ, SDM, and TM were responsible for managing the participants' submissions to the CAFA challenge. KAL, AWC, and DAH performed the experimental work in *C. albicans* and *P. aeruginosa*. BZK and GB performed the experimental work in *D. melanogaster*. CJJ, MJM, COD, and GG provided the novel biocurated data for the benchmarks and incorporated data into UniprotKB. All other co-authors developed the computational function prediction methods participating in the challenge, performed the computational protein function predictions, and submitted the results for analysis in CAFA3. NZ, IF, CSG, DAH, and PR wrote the manuscript. All authors have read and approved the final manuscript.

Funding

The work of IF was funded, in part, by the National Science Foundation award DBI-1458359. The work of CSG and AJL was funded, in part, by the National Science Foundation award DBI-1458390 and GBMF 4552 from the Gordon and Betty Moore Foundation. The work of DAH and KAL was funded, in part, by the National Science Foundation award DBI-1458390, National Institutes of Health NIGMS P20 GM113132, and the Cystic Fibrosis Foundation CFRDP STANTO19R0. The work of AP, HY, AR, and MT was funded by BBSRC grants BB/K004131/1, BB/F00964X/1 and BB/M025047/1, Consejo Nacional de Ciencia y Tecnología Paraguay (CONACyT) grants 14-INV-088 and PINV15-315, and NSF Advances in BioInformatics grant 1660648. The work of JC was partially supported by an NIH grant (R01GM093123) and two NSF grants (DBI 1759934 and IIS1763246). ACM acknowledges the support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2155 "RESIST" - Project ID 39087428. DK acknowledges the support from the National Institutes of Health (R01GM123055) and the National Science Foundation (DMS1614777, CMMI1825941). PB acknowledges the support from the National Institutes of Health (R01GM60595). GB and BZK acknowledge the support from the National Science Foundation (NSF 1458390) and NIH DP1MH110234. FS was funded by the ERC StG 757700 "HYPER-INSIGHT" and by the Spanish Ministry of Science, Innovation and Universities grant BFU2017-89833-P. FS further acknowledges the funding from the Severo Ochoa award to the IRB Barcelona. TS was funded by the Centre of Excellence project "BioProspecting of Adriatic

Sea", co-financed by the Croatian Government and the European Regional Development Fund (KK.01.1.1.01.0002). The work of SK was funded by ATT Tieto käyttöön grant and Academy of Finland. JB and HM acknowledge the support of the University of Turku, the Academy of Finland and CSC – IT Center for Science Ltd. TB and SM were funded by the NIH awards UL1 TR002319 and U24 TR002306. The work of CZ and ZW was funded by the National Institutes of Health R15GM120650 to ZW and start-up funding from the University of Miami to ZW. The work of PWR was supported by the National Cancer Institute of the National Institutes of Health under Award Number U01CA198942. PR acknowledges NSF grant DBI-1458477. PT acknowledges the support from Helsinki Institute for Life Sciences. The work of AJM was funded by the Academy of Finland (No. 292589). The work of FZ and WT was funded by the National Natural Science Foundation of China (31671367, 31471245, 91631301) and the National Key Research and Development Program of China (2016YFC1000505, 2017YFC0908402]. CS acknowledges the support by the Italian Ministry of Education, University and Research (MIUR) PRIN 2017 project 2017483NH8. SZ is supported by the National Natural Science Foundation of China (No. 61872094 and No. 61572139) and Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01). PLF and RLH were supported by the National Institutes of Health NIH R35-GM128637 and R00-GM097033. JG, DTJ, CW, DC, and RF were supported by the UK Biotechnology and Biological Sciences Research Council (BB/N019431/1, BB/L020505/1, and BB/L002817/1) and Elsevier. The work of YZ and CZ was funded in part by the National Institutes of Health award GM083107, GM116960, and AI134678; the National Science Foundation award DBI1564756; and the Extreme Science and Engineering Discovery Environment (XSEDE) award MCB160101 and MCB160124. The work of BG, VP, RD, NS, and NV was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Project No. 173001. The work of YWL, WHL, and JMC was funded by the Taiwan Ministry of Science and Technology (106-2221-E-004-011-MY2). YWL, WHL, and JMC further acknowledge the support from "the Human Project from Mind, Brain and Learning" of the NCCU Higher Education Sprout Project by the Taiwan Ministry of Education and the National Center for High-performance Computing for computer time and facilities. The work of IK and AB was funded by Montana State University and NSF Advances in Biological Informatics program through grant number 0965768. BR, TG, and JR are supported by the Bavarian Ministry for Education through funding to the TUM. The work of RB, VG, MB, and DCEK was supported by the Simons Foundation, NIH NINDS grant number 1R21NS103831-01 and NSF award number DMR-1420073. CJJ acknowledges the funding from a University of Illinois at Chicago (UIC) Cancer Center award, a UIC College of Liberal Arts and Sciences Faculty Award, and a UIC International Development Award. The work of ML was funded by Yad Hanadiv (grant number 9660 /2019). The work of OL and IN was funded by the National Institute of General Medical Science of the National Institute of Health through GM066099 and GM079656. Research Supporting Plan (PSR) of University of Milan number PSR2018-DIP-010-MFRAS. AWV acknowledges the funding from the BBSRC (CASE studentship BB/M015009/1). CD acknowledges the support from the Swiss National Science Foundation (150654). CO and MJM are supported by the EMBL-European Bioinformatics Institute core funds and the CAFA BBSRC BB/N004876/1. GG is supported by CAFA BBSRC BB/N004876/1. SCET acknowledges funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 778247 (IDPfun) and from COST Action BM1405 (NGP-net). SEB was supported by NIH/NIGMS grant R01 GM071749. The work of MLT, JMR, and JMF was supported by the National Human Genome Research Institute of the National of Health, grant numbers U41 HG007234. The work of JMF and JMR was also supported by INB Grant (PT17/0009/0001 - ISCIII-SGEFI / ERDF). VA acknowledges the funding from TUBITAK EEEAG-116E930. RCA acknowledges the funding from KanSil 2016K121540. GV acknowledges the funding from Università degli Studi di Milano - Project "Discovering Patterns in Multi-Dimensional Data" and Project "Machine Learning and Big Data Analysis for Bioinformatics". SZ is supported by the National Natural Science Foundation of China (No. 61872094 and No. 61572139) and Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01). RY and SY are supported by the 111 Project (NO. B18015), the key project of Shanghai Science & Technology (No. 16JC1420402), Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01), and ZJLab. ST was supported by project Ribes Network POR-FESR 3S4H (No. TOPP-ALFREVE18-01) and PRID/SID of University of Padova (No. TOPP-SID19-01). CZ and ZW were supported by the NIGMS grant R15GM120650 to ZW and start-up funding from the University of Miami to ZW. The work of MK and RH was supported by the funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3454-01-01 and URF/1/3790-01-01. The work of SDM is funded, in part, by NSF award DBI-1458443.

Availability of data and materials

Data repository: A data repository providing all additional data, analyses, and all anonymous prediction results for all methods are available at https://figshare.com/articles/Supplementary_data/8135393/3 [63]. Code: The assessment software used in this paper is available under GNU-GPLv3 license on GitHub in both Matlab [64] (https://doi.org/10.5281/zenodo.3403452) and Python [65](http://doi.org/10.5281/zenodo.3401694).

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA, USA. [2]Program in Bioinformatics and Computational Biology, Ames, IA, USA. [3]Indiana University Bloomington, Bloomington, Indiana, USA. [4]University of Washington, Department of Biomedical Informatics and Medical Education, Seattle, WA, USA. [5]Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA. [6]Geisel School of Medicine at Dartmouth, Hanover, NH, USA. [7]Department of Molecular and Systems Biology, Hanover, NH, USA. [8]Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA. [9]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom. [10]Program in Computer Science, Ames, IA, USA. [11]Department of Computer Engineering, Hacettepe University, Ankara, Turkey. [12]Department of Computer Engineering, Middle East Technical University (METU), Ankara, Turkey. [13]Department of Computer Engineering, Iskenderun Technical University, Hatay, Turkey. [14]CanSyL, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey. [15]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. [16]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA. [17]Achira Labs, Bangalore, India. [18]Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain. [19]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [20]INB Coordination Unit, Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Catalonia, Spain. [21](former) INB GN2, Structural and Computational Biology Programme, Spanish National Cancer Research Centre, Barcelona, Catalonia, Spain. [22]Laboratory for Bioinformatics and Computational Chemistry, Institute of Nuclear Sciences VINCA, University of Belgrade, Belgrade, Serbia. [23]Molecular Cell Biomechanics Laboratory, Departments of Bioengineering, University of California Berkeley, Berkeley, CA, USA. [24]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Berkeley, CA, USA. [25]Departments of Bioengineering and Mechanical Engineering, Berkeley, CA, USA. [26]Bologna Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy. [27]National Research Council, IBIOM, Bologna, Italy. [28]University of Bonn: INRES Crop Bioinformatics, Bonn, North Rhine-Westphalia, Germany. [29]INRES Crop Bioinformatics, University of Bonn, Bonn, Germany. [30]Gianforte School of Computing, Montana State University, Bozeman, Montana, USA. [31]University of Bristol, Computer Science, Bristol, Bristol, United Kingdom. [32]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Brunswick, Germany. [33]RESIST, DFG Cluster of Excellence 2155, Brunswick, Germany. [34]Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles - Vrije Universiteit Brussel, Brussels, Belgium. [35]Machine Learning Group, Université libre de Bruxelles, Brussels, Belgium. [36]Artificial Intelligence lab, Vrije Universiteit Brussel, Brussels, Belgium. [37]European Molecular Biolo gy Labora tory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [38]MRC Laboratory of Molecular Biology, Cambridge, United Kingdom. [39]University of Kent, School of Computing, Canterbury, United Kingdom. [40]School of Biosciences, University of Kent, Canterbury, Kent, United Kingdom. [41]University of Missouri, Computer Science, Columbia, Missouri, USA. [42]Department of Electrical Engineering and Computer Science, University

of Missouri, Columbia, MO, USA. [43]University of Miami, Coral Gables, Florida, USA. [44]Centre for Systems and Synthetic Biology, Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, United Kingdom. [45]School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Galway, Ireland. [46]Technical University of Munich, Garching, Germany. [47]Faculty for Informatics, Garching, Germany. [48]Department for Bioinformatics and Computational Biology, Garching, Germany. [49]School of Computing Sciences and Computer Engineering, Hattiesburg, Mississippi, USA. [50]Institute of Biotechnology, Helsinki Institute of Life Sciences, University of Helsinki, Finland, Helsinki, Finland. [51]Institute of Biotechnology, University of Helsinki, Helsinki, Finland. [52]Compugen Ltd., Holon, Israel. [53]Baylor College of Medicine, Department of Biochemistry and Molecular Biology, Houston, TX, USA. [54]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, USA. [55]National TsingHua University, Hsinchu, Taiwan. [56]Department of Electrical Engineering in National Tsing Hua University, Hsinchu City, Taiwan. [57]The Hebrew University of Jerusalem, Jerusalem, Israel. [58]University of California San Diego, San Diego Supercomputer Center, La Jolla, California, USA. [59]Department of Computational Biology and Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [60]Department of Genetics, Evolution & Environment, and Department of Computer Science, University College London, London, UK. [61]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [62]Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia. [63]Jozef Stefan Institute, Ljubljana, Slovenia. [64]Jozef Stefan International Postgraduate School, Ljubljana, Slovenia. [65]Research Department of Structural and Molecular Biology, University College London, London, England. [66]Research Department of Structural and Molecular Biology, University College London, London, United Kingdom. [67]Oxford Brookes University, Department of Health and Life Sciences, London, UK. [68]University College London, Department of Computer Science, London, United Kingdom. [69]The Francis Crick Institute, Biomedical Data Science Laboratory, London, United Kingdom. [70]Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, United Kingdom. [71]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [72]Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Madrid, Spain. [73]Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [74]Università degli Studi di Milano - Computer Science Department - AnacletoLab, Milan, Milan, Italy. [75]Institut de Biologie Computationnelle, LIRMM, CNRS-UMR 5506, Universite de Montpellier, Montpellier, France. [76]Department of Informatics, Bioinformatics and Computational Biology—i12, Technische Universitat Munchen, Munich, Germany. [77]University of Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France. [78]University of Lorraine, Nancy, Lorraine, France. [79]Inria, Nancy, France. [80]Department of Biology, New York University, New York, NY, USA. [81]NYU Center for Data Science, New York, NY 10010, USA. [82]Flatiron Institute, CCB, 10010 New York, NY, USA. [83]Center for Computational Biology (CCB), Flatiron Institute, Simons Foundation, New York, New York, USA. [84]Center for Data Science, New York University, New York 10011, NY, USA. [85]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [86]Department of Molecular Medicine, University of Padova, Padova, Italy. [87]Department of Biology - University of Padova, Padova, Italy. [88]CNR Institute of Neuroscience, Padova, Italy. [89]Department of Biomedical Sciences, University of Padua, Padova, Italy. [90]Department of Computer Science, National University of Computer and Emerging Sciences, Peshawar, Khyber Pakhtoonkhwa, Pakistan. [91]Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. [92]University of California, Riverside, Philadelphia, PA, USA. [93]Department of Biology, Brigham Young University, Provo, UT, USA. [94]Bioinformatics Research Group, Provo, UT, USA. [95]School of Biological Sciences, University of Reading, Reading, England, United Kingdom. [96]Department of Pharmaceutical Chemistry, San Francisco, CA, USA. [97]UC Berkeley - UCSF Graduate Program in Bioengineering, University of California, San Francisco 94158, CA, USA. [98]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco 94158, CA, USA. [99]Research and Innovation Center, Edmund Mach Foundation, 38010 San Michele all'Adige, Italy. [100]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai, Shanghai, China. [101]Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai, Shanghai, China. [102]School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China. [103]Institute of Science and Technology for Brain-Inspired Intelligence and Shanghai Institute of Artificial Intelligence Algorithms, Fudan University, Shanghai, China. [104]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, Shanghai, China. [105]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai, Shanghai, China. [106]Department of Pediatrics, Brain Tumor Center, Division of Experimental Hematology and Cancer Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. [107]Pacific Lutheran University, Department of Computer Science, Tacoma, WA, USA. [108]Department of Computer Science, National Chengchi University, Taipei, Taiwan. [109]Okinawa Institute of Science and Technology, Tancha, Okinawa, Japan. [110]Tel Aviv University, Tel Aviv, Israel. [111]Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Jeddah, Saudi Arabia. [112]Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [113]Computer, Electrical and Mathematical Sciences Engineering Division (CEMSE), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [114]Politecnico di Torino, Control and Computer Engineering Department, Torino, TO, Italy. [115]University of Turku, Department of Future Technologies, Turku NLP Group, Turku, Finland. [116]University of Turku Graduate School (UTUGS), Turku, Finland. [117]University of Turku, Turku, Finland. [118]Turku Centre for Computer Science (TUCS), Turku, Finland. [119]Department of Future Technologies, Faculty of Science and Engineering, University of Turku, FI-20014 Turku, Finland. [120]Turku Centre for Computer Science (TUCS), Agora, Vesilinnantie 3, FI-20500, Turku, Finland. [121]University of Turku, Department of Future Technologies, Turku, Finland. [122]Department of Biological Sciences, Department of Computer Science, Purdue University, West Lafayette, IN, 47907, USA. [123]Department of Pediatrics, University of Cincinnati, Cincinnati 45229, OH, USA. [124]Department of Computer Science, Purdue University, West Lafayette, IN, USA. [125]Division of Electronics, Rudjer Boskovic Institute, Zagreb, Croatia. [126]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [127]SIB Swiss Institute of Bioinformatics, Zurich, Switzerland. [128]Department of Computer Science, Colorado State University, Fort Collins, CO, USA. [129]Department of Informatics, Bioinformatics & Computational Biology—i12, Technische Universitat Munchen, Munich, Germany. [130]Institute for Food and Plant Sciences WZW, Technische Universität München, Freising, Germany. [131]University of California, Berkeley, CA, USA. [132]Biological Sciences, University of Illinois at Chicago, Chicago, Illinois, USA. [133]Geisel School of Medicine at Dartmouth, Department of Molecular and Systems Biology, Hanover, NH, USA. [134]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [135]Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, Pennsylvania, USA. [136]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA.

Received: 16 May 2019 Accepted: 24 September 2019
Published online: 19 November 2019

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
2. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198–207.
3. Schnoes AM, Ream DC, Thorman AW, Babbitt PC, Friedberg I. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput Biol. 2013;9(5):1003063.
4. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci. 2003;60(12):2637–50.
5. Friedberg I. Automated protein function prediction–the genomic challenge. Brief Bioinform. 2006;7(3):225–42.
6. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007;3:88.
7. Rentzsch R, Orengo CA. Protein function prediction–the power of multiplicity. Trends Biotechnol. 2009;27(4):210–9.
8. Shehu A, Barbara D, Molloy K. A survey of computational methods for protein function predictions. Cham: Springer; 2016, pp. 225–98.
9. Cozzetto D, Jones DT. Computational methods for annotation transfers from sequence. Methods Mol Biol. 2017;1446:55–67.

10. 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(8):4285–8.
11. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S. Prediction of human protein function from post-translational modifications and localization features. J Mol Biol. 2002;319(5):1257–65.
12. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. J Comput Biol. 2003;10(6):947–60.
13. Pazos F, Sternberg MJ. Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci USA. 2004;101(41):14754–9.
14. 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(21 Suppl 1):302–10.
15. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol. 2005;1(5):45.
16. Enault F, Suhre K, Claverie JM. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics. 2005;6:247.
17. Hawkins T, Luban S, Kihara D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci. 2006;15(6):1550–6.
18. Wass MN, Sternberg MJ. Confunc–functional annotation in the twilight zone. Bioinformatics. 2008;24(6):798–806.
19. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. 2008;9(Suppl 1):4.
20. Sokolov A, Ben-Hur A. Hierarchical classification of gene ontology terms using the GOstruct method. J Bioinform Comput Biol. 2010;8(2):357–76.
21. Clark WT, Radivojac P. Analysis of protein function and its prediction from amino acid sequence. Proteins. 2011;79(7):2086–96.
22. Piovesan D, Tosatto SCE. INGA 2.0: improving protein function prediction for the dark proteome. Nucleic Acids Res. 2019;47(W1):373–8. https://doi.org/10.1093/nar/gkz375.
23. You R, Zhang Z, Xiong Y, Sun F, Mamitsuka H, Zhu S. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics. 2018;34(14):2465–73.
24. Fa R, Cozzetto D, Wan C, Jones DT. Predicting human protein function with multi-task deep neural networks. PLoS One. 2018;13(6):0198216.
25. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Toronen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DW, Bryson K, Jones DT, Limaye B, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013;10(3):221–7.
26. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, Koo da CE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraeian SM, Martelli PL, Profiti G, Casadio R, Cao R, Zhong Z, Cheng J, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016;17(1):184.
27. Friedberg I, Radivojac P. Community-wide evaluation of computational function prediction. Methods Mol Biol. 2017;1446:133–46.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
29. Kacsoh BZ, Barton S, Jiang Y, Zhou N, Mooney SD, Friedberg I, Radivojac P, Greene CS, Bosco G. New Drosophila long-term memory genes revealed by assessing computational function prediction methods. G3. 2019;9(1):251–67.
30. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C. The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res. 2015;43(Database issue):1057–63.
31. Peng Y, Jiang Y, Radivojac P. Enumerating consistent sub-graphs of directed acyclic graphs: an insight into biomedical ontologies. Bioinformatics. 2018;34(13):313–22.
32. Wang L, Law J, Kale SD, Murali TM, Pandey G. Large-scale protein function prediction using heterogeneous ensembles. F1000Res. 2018;7.
33. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, Perchuk B, Laub MT, Hogan DA, Greene CS. Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. Cell Syst. 2017;5(1):63–71.
34. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. 2017;45(Database issue):592–6.
35. Goyard S, Knechtle P, Chauvel M, Mallet A, Prevost MC, Proux C, Coppee JY, Schwarz P, Dromer F, Park H, Filler SG, Janbon G, d'Enfert C. The Yak1 kinase is involved in the initiation and maintenance of hyphal growth in Candida albicans. Mol Biol Cell. 2008;19(5):2251–66.
36. Gutierrez-Escribano P, Gonzalez-Novo A, Suarez MB, Li CR, Wang Y, de Aldana CR, Correa-Bordes J. CDK-dependent phosphorylation of Mob2 is essential for hyphal development in Candida albicans. Mol Biol Cell. 2011;22(14):2458–69.
37. Lassak T, Schneider E, Bussmann M, Kurtz D, Manak JR, Srikantha T, Soll DR, Ernst JF. Target specificity of the Candida albicans Efg1 regulator. Mol Microbiol. 2011;82(3):602–18.
38. Martin R, Moran GP, Jacobsen ID, Heyken A, Domey J, Sullivan DJ, Kurzai O, Hube B. The Candida albicans-specific gene EED1 encodes a key regulator of hyphal extension. PLoS One. 2011;6(4):18394.
39. Richard ML, Nobile CJ, Bruno VM, Mitchell AP. Candida albicans biofilm-defective mutants. Eukaryot Cell. 2005;4(8):1493–502.
40. Bernardo SM, Khalique Z, Kot J, Jones JK, Lee SA. Candida albicans VPS1 contributes to protease secretion, filamentation, and biofilm formation. Fungal Genet Biol. 2008;45(6):861–77.
41. Yi S, Sahni N, Daniels KJ, Lu KL, Huang G, Srikantha T, Soll DR. Self-induction of a/a or α/α biofilms in Candida albicans is a pheromone-based paracrine system requiring switching. Eukaryot Cell. 2011;10(6):753–60.
42. Hess DC, Myers CL, Huttenhower C, Hibbs MA, Hayes AP, Paw J, Clore JJ, Mendoza RM, Luis BS, Nislow C, Giaever G, Costanzo M, Troyanskaya OG, Caudy AA. Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. PLOS Genetics. 2009;5(3):1–16. https://doi.org/10.1371/journal.pgen.1000407.
43. Hibbs MA, Myers CL, Huttenhower C, Hess DC, Li K, Caudy AA, Troyanskaya OG. Directing experimental biology: a case study in mitochondrial biogenesis. PLOS Comput Biol. 2009;5(3):1–12. https://doi.org/10.1371/journal.pcbi.1000322.
44. Blus-Kadosh I, Zilka A, Yerushalmi G, Banin E. The effect of pstS and phoB on quorum sensing and swarming motility in Pseudomonas aeruginosa. PLoS One. 2013;8(9):74444.
45. Kuchma SL, Brothers KM, Merritt JH, Liberati NT, Ausubel FM, O'Toole GA. BifA, a cyclic-Di-GMP phosphodiesterase, inversely regulates biofilm formation and swarming motility by Pseudomonas aeruginosa PA14. J Bacteriol. 2007;189(22):8165–78.
46. Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, Brinkman FS. Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas Genome Database. Nucleic Acids Res. 2016;44(D1):646–53.
47. Friedman L, Kolter R. Genes involved in matrix formation in Pseudomonas aeruginosa PA14 biofilms. Mol Microbiol. 2004;51(3):675–90.
48. Friedman L, Kolter R. Two genetic loci produce distinct carbohydrate-rich structural components of the Pseudomonas aeruginosa biofilm matrix. J Bacteriol. 2004;186(14):4457–65.
49. Jackson KD, Starkey M, Kremer S, Parsek MR, Wozniak DJ. Identification of psl, a locus encoding a potential exopolysaccharide that is essential for Pseudomonas aeruginosa PAO1 biofilm formation. J Bacteriol. 2004;186(14):4466–75.
50. Synapse. https://www.synapse.org/. Accessed 1 Jan 2016.
51. Clark WT, Radivojac P. Information-theoretic evaluation of predicted ontological annotations. Bioinformatics. 2013;29(13):53–61.
52. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

53. Consortium TU. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45(D1):158–69.
54. Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, Wu G, Villanueva J, Wei T, Ausubel FM. An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. Proc Natl Acad Sci USA. 2006;103(8):2833–8.
55. Noble SM, French S, Kohn LA, Chen V, Johnson AD. Systematic screens of a Candida albicans homozygous deletion library decouple morphogenetic switching and pathogenicity. Nat Genet. 2010;42(7): 590–8.
56. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C, Martel N, Veronneau S, Lemieux S, Kauffman S, Becker J, Storms R, Boone C, Bussey H. Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. Mol Microbiol. 2003;50(1):167–81.
57. Liu H, Kohler J, Fink GR. Suppression of hyphal formation in Candida albicans by mutation of a STE12 homolog. Science. 1994;266(5191): 1723–6.
58. You R, Yao S, Xiong Y, Huang X, Sun F, Mamitsuka H, Zhu S. NetGO: improving large-scale protein function prediction with massive network information. Nucleic Acids Res. 2019;47(W1):379–87. https://doi.org/10.1093/nar/gkz388.
59. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):447–52. https://doi.org/10.1093/nar/gku1003.
60. Dessimoz C, Skunca N, Thomas PD. CAFA and the open world of protein function predictions. Trends Genet. 2013;29(11):609–10.
61. Jiang Y, Clark WT, Friedberg I, Radivojac P. The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. Bioinformatics. 2014;30(17): 609–16.
62. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics. 2003;19(10):1275–83. https://doi.org/10.1093/bioinformatics/btg153. Accessed 1 Aug 2019.
63. Zhou N. Supplementary data. figshare. 2019. https://doi.org/10.6084/m9.figshare.8135393.v3. https://figshare.com/articles/Supplementary_data/8135393/3.
64. Jiang Y. CAFA2. Zenodo. 2019. https://doi.org/10.5281/zenodo.3403452.
65. Zhou N, Gerten M, Friedberg I. CAFA_assessment_tool. Zenodo. 2019. https://doi.org/10.5281/zenodo.3401694.

Publisher's Note