Methods 93 (2016) 3–14

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

GoFDR: A sequence alignment based method for predicting protein functions

Qingtian Gong [a,1], Wei Ning [a,1], Weidong Tian [a,b,*]

[a] State Key Laboratory of Genetic Engineering, Collaborative Innovation Center of Genetics and Development, Department of Biostatistics and Computational Biology, School of Life Science, Fudan University, Shanghai 200433, PR China
[b] Children's Hospital of Fudan University, Shanghai 200433, PR China

ARTICLE INFO

Article history:
Received 1 June 2015
Received in revised form 27 July 2015
Accepted 11 August 2015
Available online 12 August 2015

Keywords:
Function prediction
GO annotation
Functional discriminating residues (FDRs)
Raw score adjustment
PSI-BLAST alignment

ABSTRACT

In this study, we developed a method named GoFDR for predicting Gene Ontology (GO)-based protein functions. The input for GoFDR is simply a query sequence-based multiple sequence alignment (MSA) produced by PSI-BLAST. For each GO term annotated to the sequences in the MSA, GoFDR identifies a number of functionally discriminating residues (FDRs) specific to the GO term, and scores the query sequence using a position specific scoring matrix (PSSM) constructed for the FDRs. The raw score is then converted into a probability score according to a score-to-probability table prepared from training sequences. GoFDR outperformed three sequence-based methods for predicting GO functions in a benchmark of 18,520 sequences. In addition, GoFDR was ranked one of the top methods according to the preliminary evaluation report released by the 2nd Critical Assessment of Function Annotation (CAFA2) project. Finally, we applied GoFDR to the complete human proteome sequences, and showed that the predictions made by GoFDR with high confidence significantly expanded current annotations of human proteome. As such, GoFDR is of great value not only for annotating protein functions in newly sequenced genomes, but also for characterizing the function of proteins of interest.

1. Introduction

Understanding the function of every protein in the genome is one of the central goals in biology. However, currently the speed to generate sequence data has far exceeded the speed to characterize protein functions. Computational algorithms for predicting protein functions from protein sequences can greatly accelerate the process to characterize protein functions. On the other hand, although high-throughput functional genomics tools are now available to generate diverse types of functional data at the genome level, new tools are in dire need to mine and integrate the functional genomics data in order to infer the functions of individual genes or proteins. Hence, development of algorithms for accurate prediction of protein function has become one of the most important goals in computational biology.

Algorithms for predicting protein functions can be generally classified into two categories according to the type of data they use, which are sequence-based algorithms that require only the sequence of a protein [1–12], and omics data-based algorithms that take input various types of functional genomics data, such as gene expression, protein–protein interaction, transcription factor binding, phenotypes, etc [13–24]. Sequence-based algorithms are usually based on a simple assumption, i.e., homologous sequences tend to have similar functions. They explore different aspects of sequence-function relationships for transferring the function of known homologous proteins to unknown proteins, such as sequence similarity [3], protein domains [11], sequence patterns/motifs [25], functionally conserved residues [7], and so on. Omics data-based algorithms typically employ the "guilt-by-association" principle for inferring protein functions from gene–gene relationships obtained through mining functional genomics data [26–28], and often involve the implementation of machine learning algorithms [29,30]. Omics data-based algorithms are currently under fast development because of the accumulation of

Abbreviations: GO, Gene Ontology; AA, amino acid; IEA, inferred from electronic evidence; RCA, inferred from reviewed computational analysis; BP, biological process; CC, cellular component; MF, molecular function; PSSM, position-specific score matrix; FDRs, functionally discriminating residues; CAFA, Critical Assessment of Function Annotation.

* Corresponding author.
E-mail address: weidong.tian@fudan.edu.cn (W. Tian).

[1] Qingtian Gong and Wei Ning contributed equally to this article.

http://dx.doi.org/10.1016/j.ymeth.2015.08.009
1046-2023/

enormous functional genomics data and the emergence of new types of omics data. Though using only sequence information, sequence-based algorithms have proven to be generally more accurate than omics data-based algorithms for predicting protein functions, especially for predicting proteins' molecular functions [13]. Therefore, in this study we focused on the development of a sequence-based algorithm.

Given the sequence of a query protein, sequence-based algorithms usually start from the detection of functionally known homologous sequences, which is typically done by BLAST [31] or PSI-BLAST [32]. In the easiest case, if the sequence identity of a known sequence is above 60% to the query sequence, then the function of the known sequence can be transferred to the query sequence with above 90% accuracy [33,34]. However, this condition is often not met, and more sophisticated approaches have to be developed. PFP [6], GOtcha [5], PDCN [35], BAR+ [8], etc., explore the *E*-values produced by BLAST or PSI-BLAST for inferring protein functions. There are also methods that infer protein functions based on the detection of functional domain or motifs, such as SUPERFAMILY [36], FunFams [37], or functionally important residues, such as EFICAz [38] and ConFunc [7]. In theory, the functional domains/residues-based approaches should be more accurate than sequence similarity based methods, because the function of a protein is often determined by a functional domain or even a small number of functionally important residues rather than by the entire sequence [33]. However, identifying the functional domain/residues associated with a given function is not a trivial task, and often requires the construction of high quality multiple sequence alignments (MSAs). For example, EFICAz [38], a method developed by Tian and Skolnick that infers enzyme functions by detecting the functionally discriminating residues (FDRs) specific to a given enzyme function, requires the construction of high-quality seed MSAs for sequences with the same enzymatic function. However, for protein function annotations defined by Gene Ontology (GO) consortium [39], it is not practical to prepare a high quality MSA for each GO term. This makes these functional domains/residues-based methods not as convenient as the PSI-BLAST *E*-value based methods that use only the PSI-BLAST output for inferring functions.

In this study, we aimed to develop an algorithm that is as approximately accurate as the functional residue-based methods, while in the meantime can be also as convenient as the *E*-value-based methods. Here, we present GoFDR, a sequence alignment-based algorithm that adopts the FDR approach used by EFICAz for predicting protein function, while avoiding the complicated MSA construction work required by EFICAz by using the query sequence-based MSA directly from PSI-BLAST output. Using the PSI-BLAST-based MSA, GoFDR identifies all GO terms associated with the sequences in the MSA, and determines the FDRs for each GO term, from which a position specific scoring matrix (PSSM) is constructed. It then uses the PSSM to score the query sequence for its association with the GO term, followed by a raw score adjustment step to convert the raw score into a probability. The score conversion is an essential step in GoFDR. Using a carefully designed benchmark, we showed that GoFDR outperformed three sequence-based methods: GOtcha, PFP, the two *E*-value-based methods, and ConFunc, a functional-residue based algorithm. GoFDR's superior performance was further validated by the 2nd Critical Assessment of Functional Annotation (CAFA2): it was ranked one of the top methods among 56 participating teams according to the preliminary evaluation report released by CAFA2 organizers. Finally, we applied GoFDR to the complete human proteome, and showed that the predictions made by GoFDR with high confidence significantly expanded current annotations.

2. Material and methods

2.1. Data collection

GO annotations [39] was downloaded from UniProt-GOA website (http://www.ebi.ac.uk/GOA) on 2013-7-1. Gene Ontology file was obtained from Gene Ontology Consortium website (http://geneontology.org/page/download-ontology) on 2013-06-25, and was used to expand GO annotations by tracing back to parent GO terms. The total number of GO annotations in GOA is 7.87×10^8. After filtering GO annotations with evidence code of "IEA" or "RCA", we obtained a total number of 5.56×10^6 high-quality GO annotations.

2.2. Algorithm design of GO-FDR

There are four steps in GO-FDR: preparation of a query sequence-based multiple sequence alignment (MSA), determination of functionally discriminating residues (FDRs) for a target GO term and construction of a position specific scoring matrix (PSSM) for the FDRs, scoring the query protein using the PSSM, and adjusting the raw score into probability. Below we describe these four steps in details.

(1) Preparation of a query sequence-based MSA by database search

BLAST (blastp version 2.2.26+) or PSI-BLAST (psiblast 2.2.27+) (three iterations) search with default parameters was run with a query sequence to produce the query sequence-based MSA. For both BLAST and PSI-BLAST searches, the *E*-value threshold for selecting hit sequences was 0.01 and the maximum number of hit sequences was 20,000. The sequence database was UniRef90 [40] (released in Jul 2013) that consists of approximately 1.50×10^7 proteins.

(2) Identification of FDRs for a given GO term, and construction of the FDR-PSSM

After mapping GO annotations to the sequences in the MSA, we identified all relevant GO terms to be predicted. Here, only high-quality GO annotations were used. Then, for each GO term we adopted EFICAz's approach with modifications to identify the corresponding FDRs. For details about EFICAz, refer to [38]. Among all sequences with at least one GO term in the MSA, we identified those with the target GO term and those without, and used their aligned sequences in the MSA to prepare the homo-functional ad the hetero-functional MSA, respectively. Then, for each position i in the MSA we calculated a relative entropy (RE):

$$\mathrm{RE}(i) = \sum_{\mathrm{AA}=\{\mathrm{A,C,D...,Y}\}}^{\mathrm{n}=20} p(i,\mathrm{AA})\ \log\frac{p(i,\mathrm{AA})}{q(i,\mathrm{AA})}$$

where $p(i,\mathrm{AA})$ and $q(i,\mathrm{AA})$ are the frequency of amino acid AA at position i in the homo-functional and the hetero-functional MSA, respectively, and were computed as

$$\mathrm{Freq}_{\mathrm{AA,homo}}(i) = \frac{n_{\mathrm{AA,homo}}(i)}{N_{\mathrm{homo}}+1} + \mathrm{Freq}_{\mathrm{AA,bg}}$$

$$\mathrm{Freq}_{\mathrm{AA,hetero}}(i) = \frac{n_{\mathrm{AA,hetero}}(i)}{N_{\mathrm{hetero}}+1} + Freq_{\mathrm{AA,bg}}$$

$n_{\mathrm{AA,homo}}(i)$ and $n_{\mathrm{AA,hetero}}(i)$ are the number of sequences in the homo-functional and the hetero-functional MSA with amino acid AA at position i, respectively, and N_{homo} and N_{hetero} are the total number of homo-functional and hetero-functional sequences, respectively. $\mathrm{Freq}_{\mathrm{AA,bg}}$ refers to the background frequency of amino acid AA at position i in the whole MSA. When a gap was

encountered, each of 20 AAs would be added 1/20. After the calculation of RE at all positions, we computed the average and standard deviation, and then the *Z*-scores of RE for each position. The higher the *z*-score threshold, the more powerful the selected residues in distinguishing homo-functional sequences from hetero-functional sequences. However, higher *z*-score threshold would also result in smaller number of selected FDRs. To balance the functional discriminating power of the selected FDRs and the number of selected FDRs, we chose a *z*-score threshold of 1.0. Those positions with a *Z*-score greater than 1.0 were considered FDRs. The PSSM for the FDRs were constructed by computing the log-odds of the frequency of each of 20 AAs in the homo-functional vs. in the hetero-functional MSA as the followings:

$$\log\_\text{odds}(i, \text{AA}) = log\frac{p(i, \text{AA})}{q(i, \text{AA})}$$

where $p(i,\text{AA})$ and $q(i,\text{AA})$ are the frequency of amino acid AA at position i in the homo-functional and the hetero-functional MSA, respectively.

(3) Scoring the query protein using the FDR-PSSM

We identified the amino acid AA of the query sequence at each of the corresponding positions of the FDRs, and then applied the PSSM to score the query sequence by averaging the log-odds according to the PSSM. A higher score indicates that the query sequence is more similar to the sequences in the homo-functional MSA that to those in the hetero-functional MSA, and is therefore more likely to be annotated with the target GO term.

(4) Raw score adjustment

In general, predictions with higher raw scores are more likely to be true. However, the raw scores are not probabilities, and cannot tell us how likely the corresponding predictions are true. On the other hand, the raw scores may also be biased by the corresponding homo-functional MSA from which the FDR-PSSM is constructed. In order to compare different predictions, we need to convert raw scores into probabilities. Since each prediction is associated with a homo-functional MSA, by grouping predictions based on the properties of their corresponding homo-functional MSA, we can then prepare a score-to-probability table for each group of predictions. We have experimented different properties of the homo-functional MSAs, and found three that are correlated to prediction accuracy: the category of the target GO term associated with the homo-functional MSA (e.g., BP or MF), the frequency of the sequences in the homo-functional MSA among all functionally known sequences, and the maximum sequence identity of the sequences in the homo-functional MSA to the query sequence. Thus, for each combination of the above three properties, we prepared a score-to-probability table.

Specifically, we first applied GoFDR to a large number of training sequences, and computed raw scores for each of the predictions. Then, we divided all predictions into groups according to the above-mentioned three properties. For each group, we sorted the predictions according to their raw scores, and divided them into subgroups with pre-defined raw score ranges. For each raw score range, we then computed the percentage of true positive predictions, and considered it as the probability corresponding to the raw scores falling into the range. Thus, a score-to-probability table was prepared. In real practice, given a prediction, we first located the score-to-probability table based on the homo-functional MSA associated with the prediction; then, we identified the score range for the raw score, and obtained the corresponding probability. Finally, after converting all prediction raw scores of a query sequence into probabilities, we further adjusted the probability for each GO term by considering the parent–child GO term relationships, i.e., the probability for a given GO term should not be less than that for any of its child GO term, and if such case was found, then the probability of that GO term would be replaced by the probability of its child GO term.

2.3. Benchmark GoFDR

2.3.1. Benchmark dataset and GoFDR application

We selected 18,520 sequences from UniRef50 [40] (released in Jul 2013) that have been annotated with more than 3 none-IEA GO terms in each of the three GO categories as the benchmark dataset. For each sequence, we run a BLAST or PSI-BLAST search against UniRef90 to prepare a query sequence-based MSA. For most of the query sequences, database search would result in the identification of one or more functionally known sequences that have high sequence identity to the query sequence, making it too easy for inferring protein functions. Thus, when benchmarking GoFDR, all functionally known sequences with above 60% sequence identity to query sequences were removed from the query sequence-based MSAs. We randomly divided all sequences into 10 groups. Then, each time we prepared score-to-probability tables using 9 groups of sequences as training sequences, and applied the tables to convert the raw scores of the prediction made for the remaining one group of sequences into probabilities. This process was repeated 10 times, such that we obtained the probabilities for the predictions made for all 18,520 sequences.

2.3.2. Algorithms to be compared with GoFDR

We compared GO-FDR with three baseline methods and three sequence-based methods. The three baseline methods were the simple use of maximum sequence identity (max-ID), minimum *E*-value (min-*E*) and GO term frequency (GO-freq), respectively. The max-ID methods uses the maximum sequence identity of the sequences with the target GO term to the query sequence as the prediction score for the target GO term, while the min-*E* method use the maximum of $-\log(E\text{-value})$ as the prediction score. The GO-freq method uses the frequency of the sequences with the target GO term among all functionally known sequences in the MSA as the prediction score.

The three sequence-based methods were PFP [6], GOtcha [5], and ConFunc [7]. Because the source codes of these three methods were not available, we wrote the codes for each of these methods following the exact descriptions provided in the respective papers. Here, we briefly describe their algorithms. For details, refer to the original paper. PFP (2009 version [6]) records the $-\log(E\text{-values})$ of all annotated homologous sequences from a BLAST or PSI-BLAST search, and uses a Function Association Matrix to score each GO term and propagate the scores to parent GO terms. The FAM computes co-occur frequency of two GO terms and give a probability whether they will be associated to the same sequence in UniProt database. It is a typical sequence based prediction method that integrates both alignment information and Naïve Bayesian probability. GOtcha [5] also infers protein functions using the *E*-values produced by database search. It combines both the *E*-values of the GO terms and the GO hierarchical structure to score a target GO term, and further uses a "*training–testing*" strategy to convert the raw scores to *P*-scores. ConFunc [7] identifies the conserved residues among the sequences with a given GO term, and generates a GO specific PSSM. When scoring a query protein, it applies the PSSM to both the query sequence and re-shuffled query sequence to produce a p-value to indicate the association of the query sequence with the target GO term. In the original paper, ConFunc uses MUSCLE [41] to construct a high quality MSA in order to identify conserved residues. Here, the input MSA for ConFunc was simply the BLAST or PSI-BLAST-based alignment. Accordingly, the performance of ConFunc tested here does not reflect its real

performance. Because both PFP and Gotcha require the training sequences to convert the final score, we followed the 10-fold cross validation procedure described in the benchmark of GoFDR for these two methods. For all the six methods compared here, similar to the treatment in benchmarking GoFDR, we removed functionally known sequences that are above 60% sequence identity to the query sequence in the BLAST or PSI-BLAST search output.

2.3.3. Performance evaluation measures

We used the precision–recall curve to evaluate the performance of a method in predicting protein functions. Specifically, we sorted the prediction scores from high to low, and then computed the precision and recall at each score threshold and plotted the precision–recall curve. The overall performance of the precision–recall curve was evaluated by the F_{max} measure which is the maximum of the F measure along the precision–recall curve. At each threshold, the F-measure is computed using the corresponding precision and recall as $F_{measure} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$. The higher the F_{max}, the better performance a method has.

2.4. CAFA2 experiment

The Critical Assessment of Function Annotation (CAFA) project was a community effort to build a common standard to evaluate the prediction performance of different function prediction methods. The 1st CAFA was launched in 2011, and the evaluation report was published in 2013 [42]. In Aug 2013, CAFA2 was launched and 100,816 functional unknown or partially unknown protein sequences from 27 species were released to the community. Each participating team was asked to submit predicted functions with a confidence score ranging from 0 to 1 before Jan 20, 2014. The organizing committee collected the added annotations of these sequences within a five-month period after the submission deadline, and used these annotations to evaluate the performance of each participating method. We participated CAFA2 by using the GoFDR algorithm to predict functions for the released sequences, and our submitted predictions were under the model name of "Tian-Lab model 1".

Our prediction procedures were as the followings. First, we expanded high-quality GO annotations in the database by including the "confirmed" "IEA" annotations. A "confirmed" IEA annotation means that the sequence with the "IEA" annotation must have a homologous sequence that has the same annotation of non-IEA evidence code and is above 60% sequence identity to the "IEA" annotated sequence. This results in the expansion of GO annotation from 5.56×10^6 to 1.67×10^7. Next, a PSI-BLAST search with each of the target sequences was conducted against UniRef90 database to prepare query sequence-based MSAs. Then, GoFDR was applied to compute a probability score for all GO terms relevant to a target sequence. Here, the score-to-probability tables were prepared by using all 18,520 sequences in the previously mentioned benchmark dataset as the training sequences.

2.5. Application of GoFDR to human proteome sequence

We download the human proteome sequence database from UniProt (Apr-2015 release). It consists of a total number of 20,882 protein sequences that are the representative products in the human genome [40]. GoFDR was applied to predict the functions of these sequences in the same way as we described in CAFA2 experiment.

2.6. GoFDR server & source code

We have developed a webserver for GoFDR at http://gofdr.tianlab.cn, where genome-wide function predictions for the complete proteome of several model organisms can be retrieved and downloaded. In addition, pre-predicted functions of proteins with UniProt ID are provided in the webserver. Packaged GoFDR source code are also available for downloading on this website.

3. Results

3.1. A brief description of GoFDR

GoFDR is a sequence-based method for predicting Gene Ontology (GO)-based functions. The prediction pipeline by GoFDR is shown in Fig. 1. For details about its algorithm design, refer to the Method section. Here, we briefly describe the workflow of GoFDR. GoFDR takes the input of a query sequence-based MSA produced directly from BLAST or PSI-BLAST search. After mapping GO annotations to all homologous sequences in the MSA, GoFDR identifies all relevant GO terms to be predicted. For each GO term, GoFDR basically compares the sequence conservation in the homo-functional MSA (aligned sequences with the GO term) and the hetero-functional MSA (aligned known sequences without the GO term) to identify a number of functionally discriminating residues (FDRs) for distinguishing the homo-functional sequences from the hetero-functional sequences, similar to the approach introduced by EFICAz [38]. Different from EFICAz that requires exact match of the query sequence to the FDRs in order to infer functions, GoFDR builds a PSSM for the FDRs, and then applies the PSSM to score the query sequence for its association with the target GO term. Finally, GoFDR converts the raw score of a prediction into a probability according to a pre-constructed score-to-probability table from training sequences.

3.2. The use of raw score adjustment is an essential step in GoFDR

To evaluate the performance of GoFDR in predicting protein functions, we prepared a large benchmark dataset consisting of 18,520 sequences. These sequences were selected from UniRef50 in which sequences have less than 50% sequence identity to each other. We required these sequences all be annotated with at least three GO terms with non "IEA" evidence code in each of the three GO categories (MF, BP and CC). These sequences therefore comprise of a representative set of functionally known sequences in the database.

We first compared the raw scores produced by GoFDR with three baseline methods—max-ID, min-*E*, and GO-freq, referring to the use of the maximum sequence identity, the minimum *E*-value of the homologous sequences with the target GO term to the query sequence, and the frequency of the target GO term among all functionally known sequences in the query sequence-based MSA, respectively. Note here homologous sequences with above 60% sequence identity to the query sequence were removed from the query sequence-based MSA, because all methods would produce accurate predictions with the inclusion of those sequences. The min-*E* method was significantly worse than the other two baseline methods for predicting protein functions, and its performance was further deteriorated for PSI-BLAST-based *E*-values (Fig. 2). This is expected, as PSI-BLAST *E*-values actually reflect the similarity of the hit sequences to the sequence profile generated from the hit sequences produced in the previous iteration rather than to the query sequence, and are less correlated to function similarity as a consequence. However, what's interesting here was that the other two baseline methods were actually not

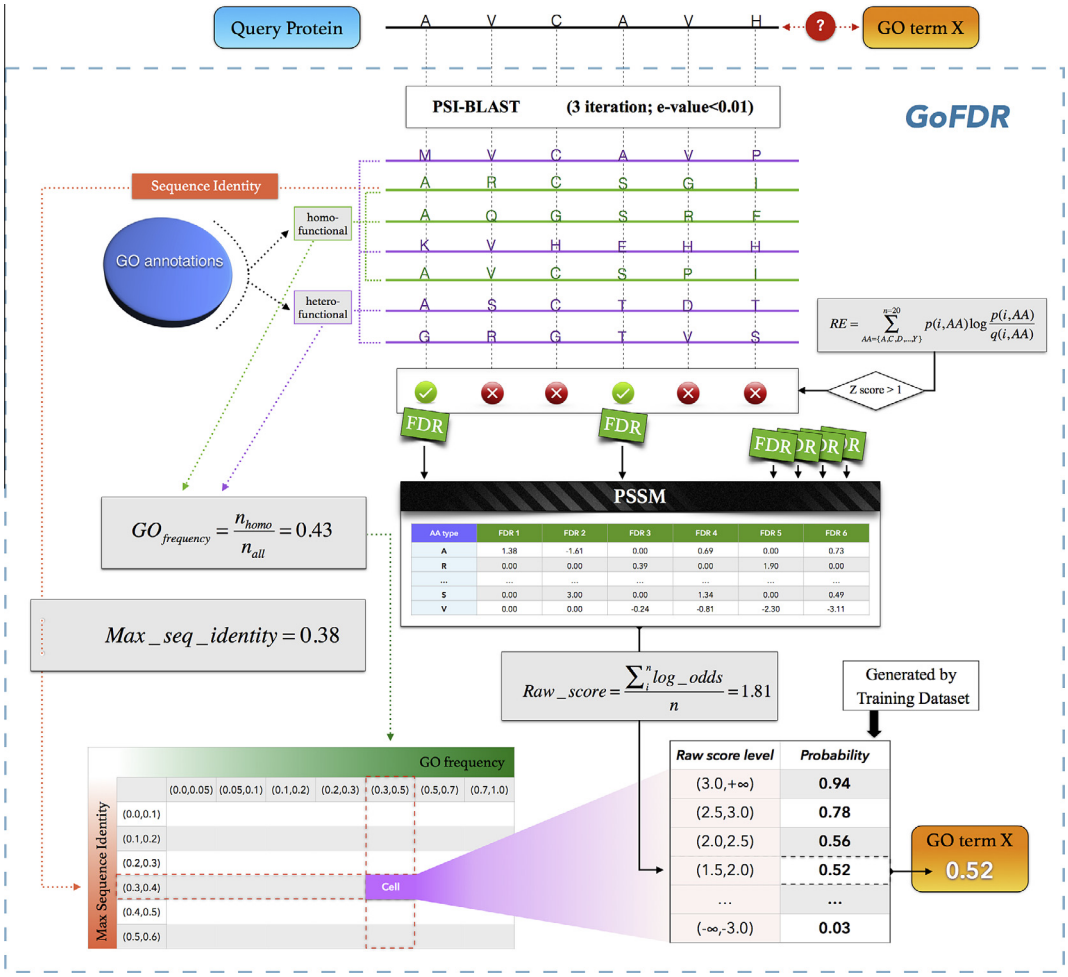

Fig. 1. The workflow of GoFDR.

bad predictors, and the simple use of GO term frequency even outperformed the raw score produced by GoFDR (Fig. 2). For example, the F_{max} for BP and CC GO terms predicted by GO-freq using BLAST-based MSA was 0.486 and 0.609, respectively, in contrast to 0.463 and 0.573 by the raw score of GoFDR, respectively. For PSI-BLAST-based MSA, GO-freq also outperformed the raw score of GoFDR for predicting BP and CC GO terms. Another finding was that except for max-ID, all methods performed better with BLAST-based MSA than with PSI-BLAST-based MSA (Fig. 2). This can be interpreted as that the inclusion of more homologous sequences by PSI-BLAST complicates the task of identifying functional relationships using these simple methods.

The above findings confirmed the raw score of GoFDR should be adjusted before different predictions could be compared with each other. Given that both the GO term frequency and the maximum sequence identity of the sequences with the GO term to the query sequence are strongly correlated with function similarity, we corrected the raw scores by considering these two types of information. We divided the 18,520 sequences in the benchmark dataset randomly into 10 groups, and selected 9 groups of sequences as training sequences to prepare a score-to-probabilities table for each combination of GO term frequency and the maximum sequence identity of the GO term (for details, refer to the Method section). Then we applied the tables to convert the raw scores of the predictions made for the remaining group of sequences into probabilities. This process was repeated 10 times to convert the raw scores of all predictions into probabilities. The conversion of raw scores into probabilities significantly improves the prediction performance of GoFDR (Fig. 3). Before the adjustment, the F_{max} for the predictions of GO terms in MF, BP, and CC categories by GoFDR using BLAST-based MSAs was 0.616, 0.463, and 0.573, respectively. After the adjustment, it was improved to 0.679, 0.540, and 0.650, respectively. The improvement was more significant for GoFDR using PSI-BLAST-based MSAs: before the adjustment, the performance of GoFDR using PSI-BLAST-based MSAs was clearly worse than using BLAST-based MSAs; after the adjustment, it was even slightly better than that using BLAST-based MSAs (Fig. 3). The performance of GoFDR using adjusted scores also made it significantly superior to the two baseline methods—GO-freq and max-ID (Fig. 3).

3.3. The comparison of GoFDR with three sequence based methods

In addition to the three baseline methods, GoFDR was also compared with three published sequence-based methods—PFP [6], GOtcha [5], and ConFunc [7]. The former two are based on the integration of *E*-values produced by database search, while the latter was based on the identification of conserved residues specific to a given function in the MSA. Note that high-quality MSA construction was a necessary step in the original ConFunc method. Here, for convenience the BLAST or PSI-BLAST-based MSA was inputted to the ConFunc method. Thus, the performance of ConFunc here does not reflect the performance of the original method. Using the same benchmark, we found that GoFDR outperformed these three meth-

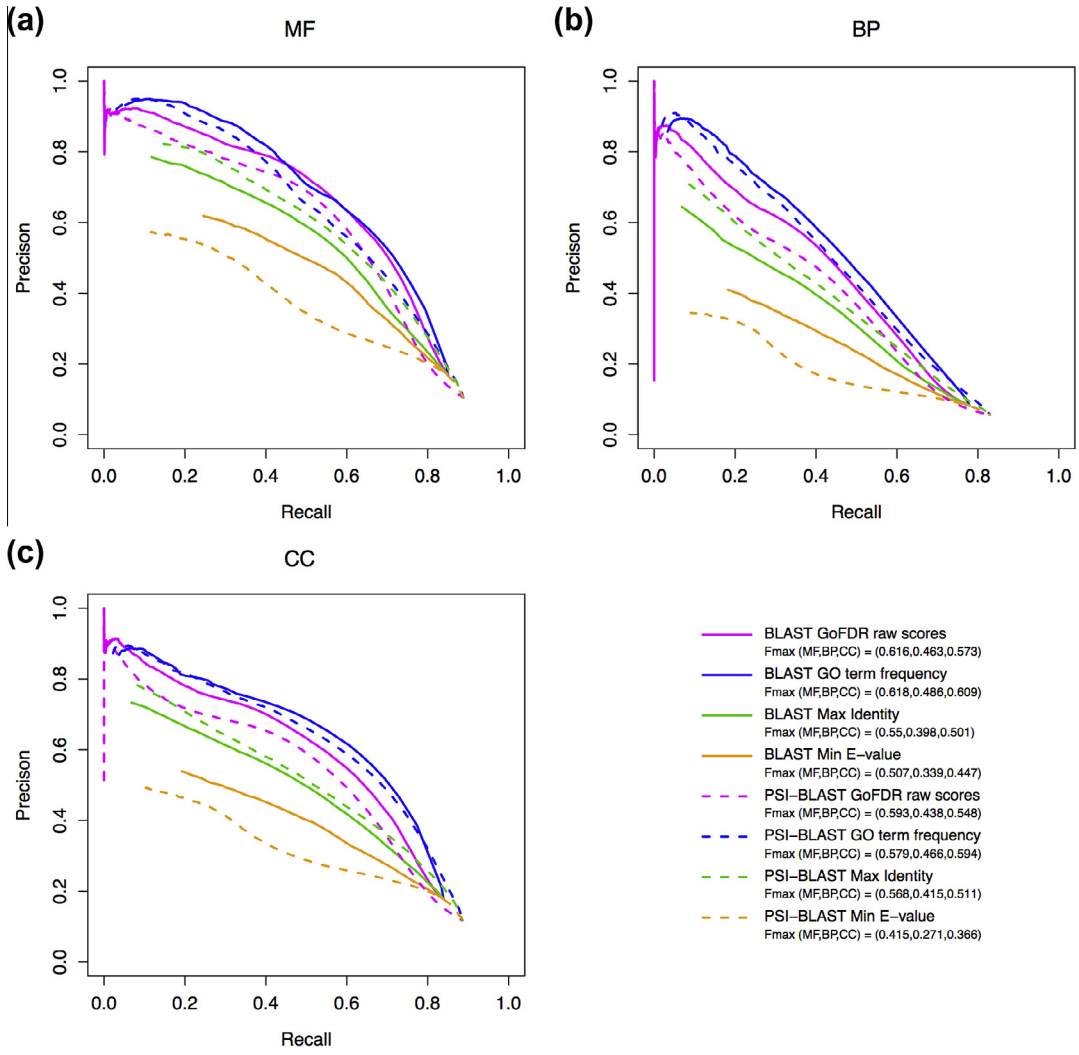

Fig. 2. The precision–recall curves for GoFDR's raw score and three baseline methods in each of MF (a), BP (b) and CC (c) categories using either BLAST (solid lines) output or PSI-BLAST output (dashed lines). Max ID: the maximum sequence identity of the sequences with the target GO term to the query sequence. Min *E*, the maximum −log(*E*-value) of the sequences with the target GO term. GO-freq: the frequency of the target GO term in the query sequence-based MSA.

ods with a significant margin, especially for PFP and ConFunc (Fig. 4). For example, the F_{max} score of GoFDR for GO terms in the MF category using PSI-BLAST-based MSAs was 0.685, in contrast to 0.567 and 0.497 for PFP and ConFunc, respectively. GOtcha's performance was closer to GoFDR's, but was evidently worse (Fig. 4). It is worth noting that the performance of all the three sequence-based methods became worsened when switch the use of BLAST output to the use of PSI-BLAST output; in contrast, GoFDR was able to achieve a better performance using PSI-BLAST based MSAs.

We further investigated the performance of different methods for GO terms with different sizes and different frequency. By dividing all predictions into groups according to the size of the target GO terms or the frequency of target GO terms in the query sequence-based MSAs produced by PSI-BLAST, we computed the F_{max} scores for all methods in each group. GoFDR was the best method at all GO term sizes, and at all ranges of GO term frequency (Fig. 5). The advantage of GoFDR over the other methods was most significant for GO terms with very small sizes (1–5 annotated genes) or with low frequencies (0–0.05), which represent the most challenging category of function prediction. Interestingly, when we divided the predictions into separate groups according to the frequency of the target GO terms, the performance of GOtcha and PFP was even worse than the two baseline methods (max-ID and min-*E*) (Fig. 5). This is likely because both PFP and GOtcha combine the *E*-values of all sequences annotated to a GO term for inferring function, and are therefore affected by the frequency of the target GO terms; in contrast, the maximum sequence identity or the minimum *E*-value approach is independent of the frequency of the target GO terms. The significantly improved performance of the min-*E* method when its performance based on individual groups of predictions as compared with all predictions also highlighted the importance of raw score conversion: also implies that by controlling certain features, even very simple methods may achieve significant performance improvement. In conclusion, the above results demonstrated the advantage of GoFDR over existing methods.

3.4. The performance of GoFDR in CAFA2

Although we have demonstrated the usefulness of GoFDR using the above-described benchmarks, it may still be argued that there may exist biases in the selection of the benchmark dataset, the selection of evaluation measures, or the way to run the other methods by us. A common benchmark dataset evaluated by independent researchers would provide unbiased estimation of different methods. The Critical Assessment of Function Annotation (CAFA) project represents a community effort to develop a com-

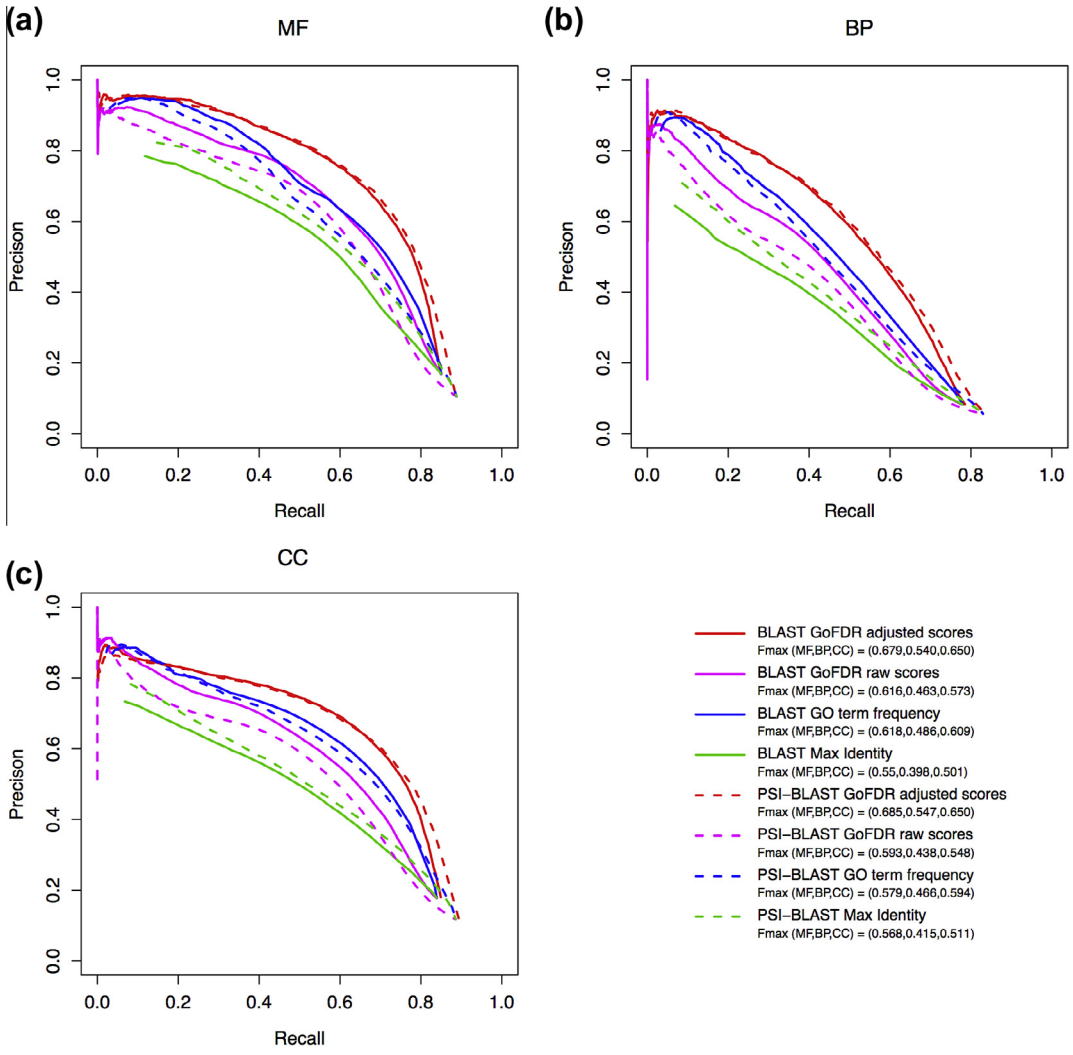

Fig. 3. The comparison of GoFDR's adjusted score with GoFDR's raw score and two baseline methods (GO-freq and max-ID). The plots are similar to those in Fig. 2. See Fig. 2 legend for details.

mon standard for evaluating the performance of function prediction methods. In Aug 2013, the CAFA2 project released over 100,000 sequences to the protein function prediction community. These sequences are either functionally unknown or only partially known, and are from 27 different species. According to the organizers of CAFA2, 54 teams submitted the predicted functions for these sequences using 126 prediction models. After the submission, the organizers collected all new experimentally determined function annotations added to these sequences over a five-month period of time, and then evaluated the performance of each submitted model using the new annotations from both EBI and all benchmark dataset. The classification of two validation datasets was predefined by CAFA2 organizers. We also participated CAFA2, and applied GoFDR to make predictions for all released target sequences. Our predictions were named under "Tian-Lab Model 1" in CAFA2.

According to the preliminary evaluation report shared by the CAFA organizers with all participants, "Tian-Lab Model 1" was one of the top methods in CAFA2. In this report, "Tian-Lab Model 1" was the only model that ranked top 10 according to the F_{max} scores in each of the three categories of GO term predictions for both the EBI benchmark and the all benchmark (Fig. 6). In the final evaluation report shared by CAFA2 organizers, "Tian-Lab Model 1" was also ranked one of the top methods, and was ranked the top method in the molecular function category (personal communications). Thus, CAFA2 provided an unbiased benchmark to evaluate and compare the performance of different methods, and the excellent performance of "Tian-Lab Model 1" clearly demonstrated the usefulness of GoFDR in predicting protein functions.

3.5. The application of GoFDR to human proteome

As an application of GoFDR, we applied it to human proteome sequences. The human proteome sequences downloaded from UniProt include a total number of 20,882 human protein sequences. After the propagation of GO annotations using the parent–child relationships in GO graphs, there are 11,610, 10,792 and 12,096 human proteins annotated with at least one non-IEA GO term in BP, MF and CC categories, respectively. As another validation of GoFDR, we treated those sequences as testing sequences, and investigated the performance of GoFDR. Here, PSIB-BLAST-based MSAs were used, and the homologous sequences with above 60% sequence identity to the query sequence in each MSA were removed. The F_{max} score of GoFDR for predictions of GO terms in the MF, BP, and CC category was 0.621, 0.476, and 0.593, respectively (Fig. 7a), further confirming that GoFDR was able to predict protein functions with reasonably good accuracy. The F_{max} scores of GoFDR obtained using human proteome sequences were lower

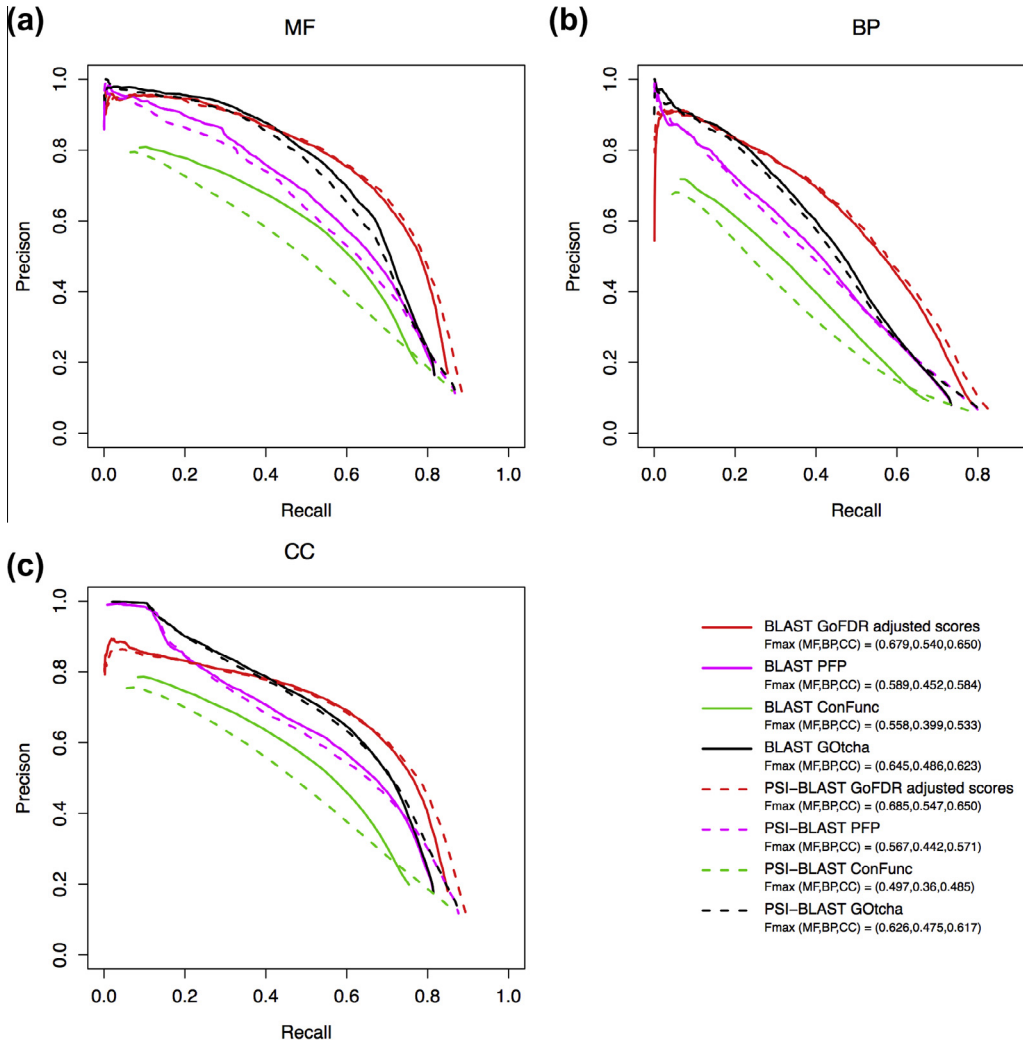

Fig. 4. The precision–recall curves for GoFDR, PFP, GOtcha, and ConFunc. The plots are similar to those in Fig. 2. See Fig. 2 legend for details.

than those using 18,520 sequences in previously describe benchmark, which were 0.685, 0.547, and 0.650 for MF, BP, and CC GO terms, respectively. This was attributed to the difference in the distribution of GO terms size in the two datasets: in the original benchmark dataset, there are a small number of GO terms that are annotated to large number of sequences; in contrast, the distribution of GO term size in the human proteome sequences is more evenly distributed. Thus, the validation of GoFDR using human proteome annotations should reflect the practical performance of GoFDR for genome sequence annotations. In addition, we also checked the precision score at different threshold of the adjust probability score in the human proteome validation dataset. When the adjusted probability produced by GoFDR was above 0.5, the corresponding precision for MF, BP and CC GO term predictions was 0.685, 0.559, and 0.512, respectively (Fig. 7b). Thus, in the following prediction of human proteome sequence, we used the adjusted probability score of 0.5 as a cutoff to select high confidence predictions.

When applying GoFDR to predict the function for human proteome sequences, we followed the same procedure we did in CAFA2. However, if a GO term was found to include a sequence with above 60% sequence identity to the query sequence, then we simply assigned a probability score of 0.9 to the GO term, and would not use GoFDR to predict that GO term. Here, we selected all predictions with a probability score of above 0.5 as high-confidence predictions as high confidence predictions. Before the prediction, there are 14,841 proteins with 9.28×10^5 annotations (non-IEA annotations + "confirmed" IEA annotations); after the prediction, there are now 19,730 proteins with 1.58×10^6 annotations plus high confidence predicted annotations, i.e., GoFDR added about 660,000 predictions (the complete list of all predictions with adjusted probability scores can be downloaded from http://gofdr.tianlab.cn). The average number of annotations for each protein sequence increased from 63 to 80.

The GoFDR predictions were made based on the annotations released in 2013 by UniProt-GOA. Here, we downloaded the newest annotations of human proteome sequences from UniPrt-GOA (2014-12-25 release), and investigated how many of the newly added non-IEA annotations were predicted with high confidence by GoFDR. After filtering out the annotations made for newly created GO terms and all non IEA plus "confirmed" IEA annotations in the 2013 release, we obtained a total number of 22,398, 143,302, and 44,777 new annotations in MF, BP, and CC categories, respectively. Among these new annotations, about 33%, 19%, and 25% were predicted by GoFDR with high confidence, respectively; in comparison, about 28%, 24% and 20% were annotated with IEA evidence code in the 2013 release of GOA, respectively (Fig. 7c). The annotations that were predicted by both GoFDR and IEA account for 20%, 10%, and 12% of the newly added annotations in MF, BP, and CC categories, respectively (data not shown). Since IEA

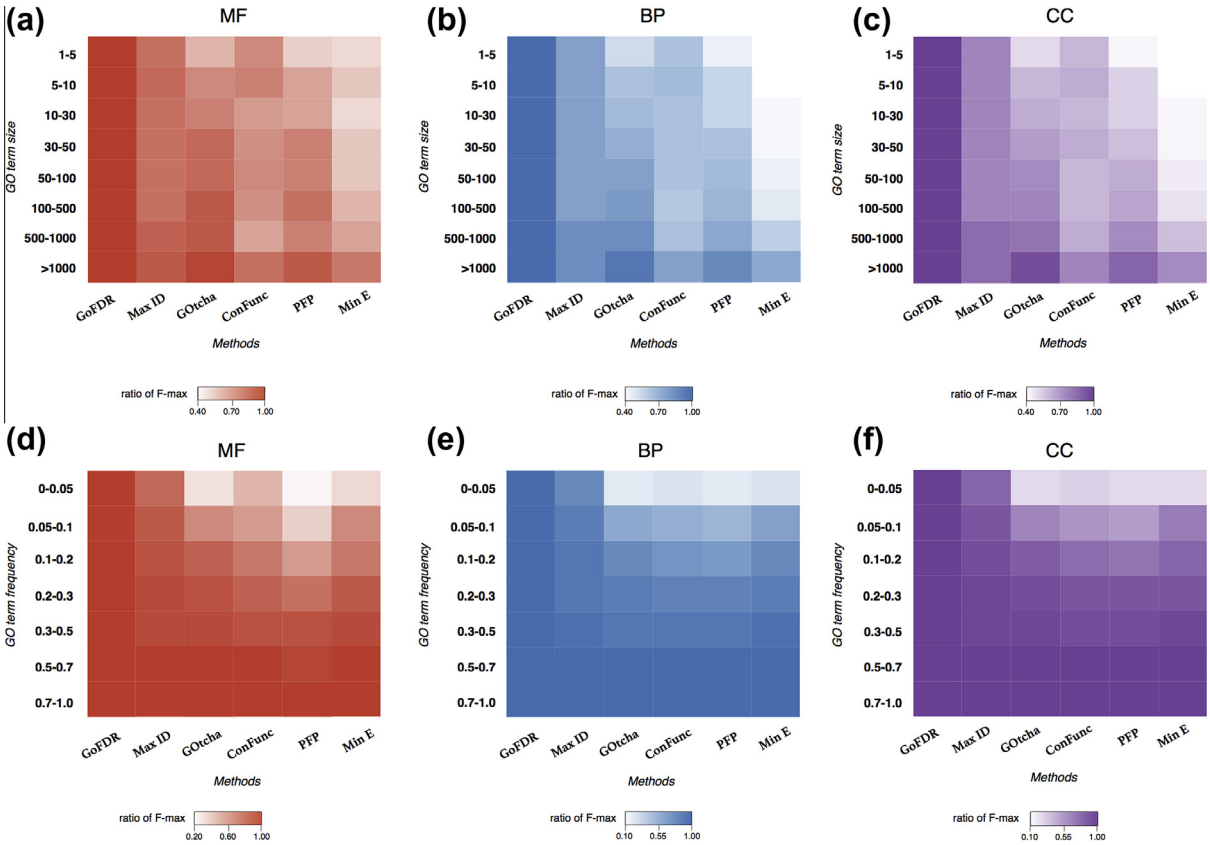

Fig. 5. The comparison of the F_{max} scores of six methods using predictions based on PSI-BLAST output at different GO term size(a, b, c) and different GO term frequency(d, e, f) level. The F_{max} score of each method was compared to that of GoFDR's at the same level, and the ratios of F_{max} scores at different levels is shown in heatmaps. For GO term size level, 10–30 means the target GO term has 10–30 annotated genes in the benchmark dataset. For Go term frequency level, 0.1–0.2 means the sequences with the target GO term account for 10–20% of functionally known sequences in a query sequence-based MSA produced by PSI-BLAST.

annotations were made based on keywords matching that tends to generate extremely large number of annotations, the fact that more newly added annotations in the MF and CC categories were predicted by GoFDR than by IEA suggested well illustrated the sensitivity of GoFDR. We further checked the statistics of the high confidence GoFDR predictions validated by newly added annotations, and found the target GO term frequency for most predictions is less than 0.1, demonstrating GoFDR's ability to make accurate predictions for the most challenging category of cases. The target GO term frequency for most newly added annotations that were not predicted by GoFDR was also less than 0.1 with the median at 0.033, indicating that they were difficult to be predicted. Here, we give two examples for the validated predictions. P59540 (taste receptor type 2 member 46 encoded by TAS2R46 gene) was predicted by GoFDR to have GO:0033038 "bitter taste receptor activity" in MF category, which was added to the current release as a non-IEA annotation. Another example is P16050, arachidonate 15-lipoxygenase, that was predicted by GoFDR with GO:0004052 "arachidonate 12-lipoxygenase activity"; this annotation is now with a non-IEA evidence code in the current release.

We also investigated how many of our high confidence predictions have not been included in the current release of non-IEA annotations, not predicted with IEA evidence code, and considered them as novel predictions. We found more than half of our predictions are novel predictions (Fig. 7d). A quick literature search found support for some of the novel predictions. For example, Q12955 (gene name Ankyrin-3 (ANK3), also named Ankyrin-G) was predicted to have GO:0045760 "positive regulation of action potential". Ankyrin-G has been found to be required for the normal clustering of voltage-gated sodium channels at the axon hillock and for action potential firing [43], supporting our prediction. Therefore, the novel predictions from GoFDR can provide biologists with new hypotheses about the function of proteins they are interested in.

4. Discussion

In this study, we have developed an alignment-based method named GoFDR for protein function prediction from the query sequence-based MSA produced by BLAST or PSI-BLAST search. We have rigorously tested GoFDR's performance, and have shown using a large benchmark dataset that it outperformed three existing sequence-based methods. In addition, we have also shown that GoFDR was ranked one of the top methods in CAFA2, a function annotation assessment experiment participated by 54 teams. Given its excellent performance and the convenient use of PSI-BLAST or BLAST output, GoFDR is of great value not only for large-scale genome sequence annotation, but also for discovering novel functions for genes/proteins of interest.

There are two key steps in GoFDR. One is the identification of GO term-specific FDRs from the query sequence-based MSA. Another is the raw score adjustment. Both steps are essential for GoFDR. Unlike *E*-values or sequence identities that measures the similarity between two sequences regardless of whether the two sequences have the same function or not, the FDRs defined in GoFDR are determined through comparing sequence conservation within sequences with the GO term to that within sequences without the GO term, and are therefore specific to the target GO term. However, there are a number of factors that may affect the identified FDRs, such as the number of sequences with the GO term, the

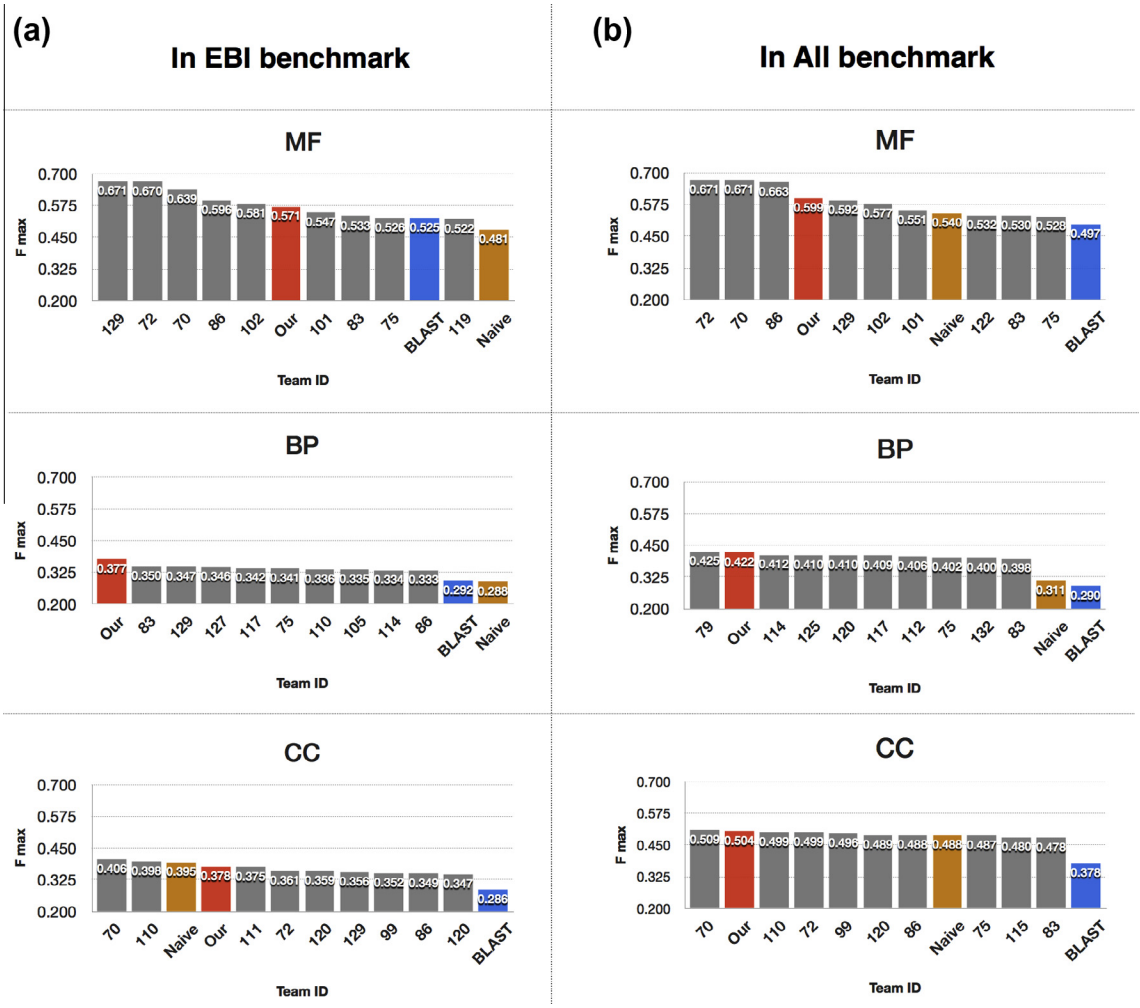

Fig. 6. The top 10 prediction models in each of MF, BP, and CC categories in CAFA2 according to the preliminary evaluation report shared by CAFA2 organizers. The bootstrapped F_{max} scores of each method in each category ear shown in barplots for both EBI benchmark dataset (a) and all benchmark dataset (b).

relative sequence dissimilarity between sequences with the GO term and those without, and the cutoff for selecting FDRs, etc., which in turn would have an impact on the raw score of the query sequence produced by GoFDR. In addition, the similarity between the query sequence and the sequences with the target GO term may also affect the raw scores. While further work could be done to develop more sophisticated methods that take into consideration of these factors when identifying FDRs, an alternative approach would be to find a way to adjust the raw scores by controlling some of the above-mentioned factors, which is a implementation of the raw score adjustment in GoFDR.

In the raw score adjustment step of GoFDR, a list of score-to-probability tables were generated for different combinations of GO category, GO term frequency and the maximum sequence identity between the sequences with the GO term to the query sequence using a large number of training sequences. With the score-to-probabilities table, the raw score of a query sequence could then be easily converted into a probability by locating the corresponding table based on the information of the target GO term. This step has proven to be critical for the performance of GoFDR. In fact, the adjustment step was not only useful for GoFDR, but also for other methods. For example, we observed that the performance of the two baseline methods—the minimum *E*-value and the maximum sequence identity method were significantly improved if the GO term frequency was controlled. This highlighted the importance of the adjustment step, and also implied that there is room for making further improvement by using more refined adjustment procedures. In addition to optimizing the FDR identification step and the score adjustment steps to further improve GoFDR, we can also integrate the components of the other successful methods in the future development of GoFDR. For example, GOtcha [5] integrates GO hierarchal structure when inferring protein functions, while PFP [6] considers the association between different GO terms. These lines of information could also be integrated in GoFDR to enhance its performance.

GoFDR is a sequence-based method, and requires only the input of query sequence-based MSAs. However, there are currently enormous amount of functional genomics data available for model organisms. Although the performance of omics data-based methods is generally worse than sequence-based methods, using only sequence information while ignoring the genomics data should not be the solution for predicting protein functions. As shown in the CAFA2 preliminary evaluation report and also shown in our benchmark, predicting BP GO terms is a more difficult task than predicting MF GO terms. BP GO terms describe the relationships between genes, while MF GO terms describe the properties of a gene. It can be easily imagined that the property of a gene is determined by itself, while the relationship of a gene with other genes is

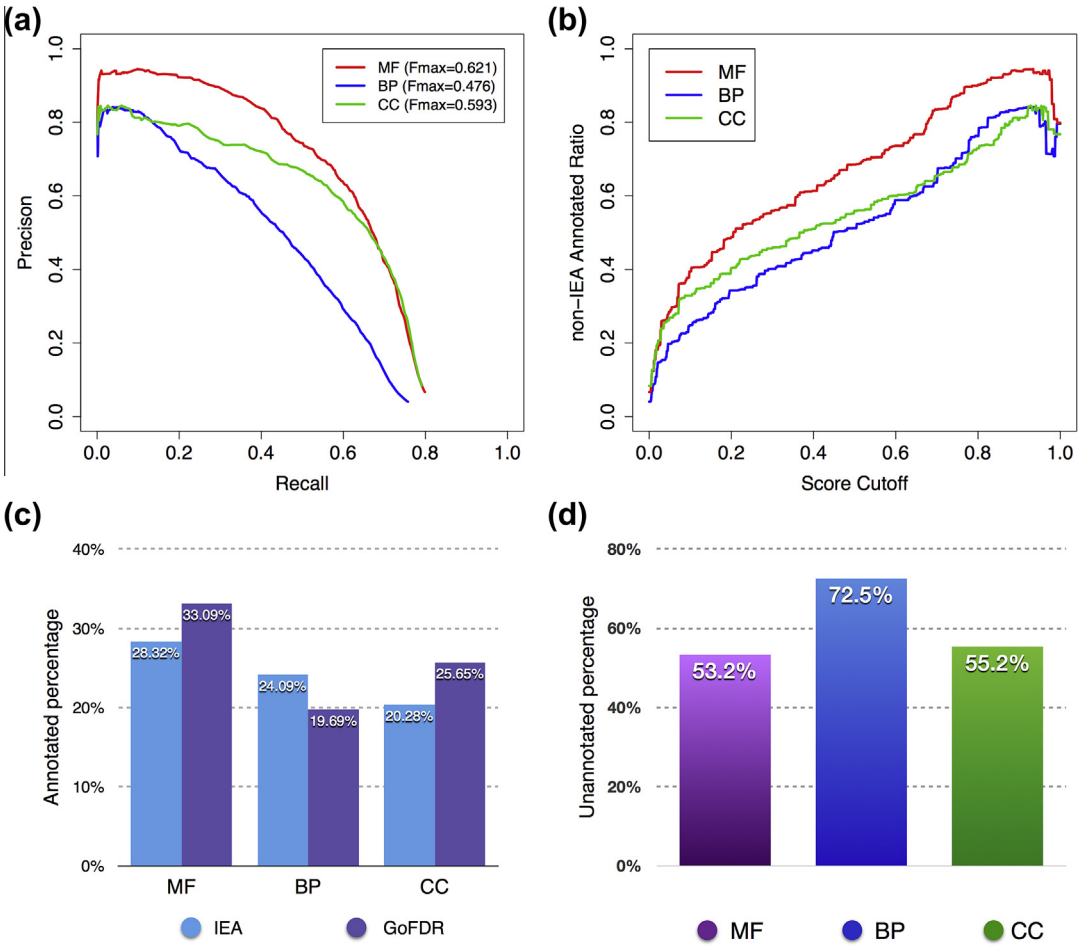

Fig. 7. (a) The precision–recall curves of GoFDR in predicting human proteome sequences. (b) The precision score of GoFDR at different cutoff of adjusted probabilities in predicting human proteome sequences. (c) The percentage of newly added non-IEA annotations that were annotated by old release IEA annotations or predicted by GoFDR with high confidence. (d) The percentage of GoFDR's high confidence predictions in MF, BP, CC category that are novel predictions. A high confidence prediction is considered novel if it has not been validated by current release of GOA annotation, nor were annotated by GOA with IEA evidence code in the 2013 release.

not only determined by itself, but also by other genes, and are consequently not necessarily inferred from only the sequence of a gene. The reason why sequence-based methods can still make reasonably good predictions for some BP GO terms is because there exists strong correlation between these BP terms and some MF GO terms. Thus, besides the continuing development of sequence-based methods such as GoFDR, new tools are in dire need to take advantage of both sequence data and omics data for making function predictions, especially for making BP GO terms.

GoFDR is based on PSI-BLAST search output. Once PSI-BLAST is done, applying GoFDR only takes a couple of seconds. However, given the huge number of sequences in UniRef90 (about 1.50×10^7 proteins), running PSI-BLAST with three iterations is computationally expensive, and may take from several minutes to 10 or 20 more minutes for different query sequences. This will be a significant factor to limit the online application of GoFDR. Recently, another version of BLAST, RPS-BLAST, has been included in the BLAST software release packages. RPS-BLAST searches against a collection of protein domain databases with the query sequences, which takes less than a second to complete. We are currently working to extend GoFDR's application to the search output of RPS-BLAST, which is expected to significantly reduce the running time for making predictions using GoFDR.

Authors' contributions

WT conceived and supervised the study. QG designed the algorithm, carried out benchmark analysis, and submitted predictions to CAFA2. WN conducted the prediction for human proteome sequences, analyzed the evaluation report in CAFA2 and developed the web server. WN, QG and WT drafted the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China [31471245, 91231116, 31071113, 30971643] ; the Specialized Research Fund for the Doctoral Program of Higher Education of China [20120071110018] ; the Innovation Program of Shanghai Municipal Education Commission [13ZZ006]; the Shuguang Program of Shanghai Municipal Education Commission [13SG05].

Competing interests

We declare that we have no competing interests.

Acknowledgements

We thank the organizers of the CAFA2 experiment for sharing the preliminary results with us.

References

[1] M. Falda et al., Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms, BMC Bioinformatics 13 (Suppl 4) (2012) S14.
[2] S.M. Sahraeian, K.R. Luo, S.E. Brenner, SIFTER search: a web server for accurate phylogeny-based protein function prediction, Nucl. Acids Res. (2015) gkv461.
[3] A. Conesa et al., Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (18) (2005) 3674–3676.
[4] M. Punta, Y. Ofran, The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function, PLoS Comput. Biol. 4 (10) (2008) e1000160.
[5] D.M. Martin, M. Berriman, G.J. Barton, GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes, BMC Bioinformatics 5 (1) (2004) 178.
[6] T. Hawkins et al., PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data, Proteins: Struct., Funct., Bioinf. 74 (3) (2009) 566–582.
[7] M.N. Wass, M.J.E. Sternberg, ConFunc—functional annotation in the twilight zone, Bioinformatics 24 (6) (2008) 798–806.
[8] D. Piovesan et al., BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences, Nucleic Acids Res. 39 (suppl 2) (2011) W197–W202.
[9] W.T. Clark, P. Radivojac, Analysis of protein function and its prediction from amino acid sequence, Proteins: Struct., Funct., Bioinf. 79 (7) (2011) 2086–2096.
[10] B.E. Engelhardt et al., Protein molecular function prediction by Bayesian phylogenomics, PLoS Comput. Biol. 1 (5) (2005) e45.
[11] B.E. Engelhardt et al., Genome-scale phylogenetic function annotation of large and diverse protein families, Genome Res. 21 (11) (2011) 1969–1980.
[12] P. Gaudet et al., Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium, Brief. Bioinformatics 12 (5) (2011) 449–462.
[13] D. Cozzetto et al., Protein function prediction by massive integration of evolutionary analyses and multiple data sources, BMC Bioinformatics 14 (Suppl 3) (2013) S1.
[14] C.S. Funk et al., Evaluating a variety of text-mined features for automatic protein function prediction with GOstruct, J. Biomed. Sem. 6 (1) (2015) 9.
[15] A. Vazquez et al., Global protein function prediction from protein-protein interaction networks, Nat. Biotechnol. 21 (6) (2003) 697–700.
[16] K.M. Borgwardt et al., Protein function prediction via graph kernels, Bioinformatics 21 (suppl 1) (2005) i47–i56.
[17] R. Sharan, I. Ulitsky, R. Shamir, Network-based prediction of protein function, Mol. Sys. Biol. 3 (1) (2007).
[18] L.J. Jensen et al., Prediction of human protein function from post-translational modifications and localization features, J. Mol. Biol. 319 (5) (2002) 1257–1265.
[19] E. Nabieva et al., Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps, Bioinformatics 21 (suppl 1) (2005) i302–i310.
[20] H. Lee et al., Diffusion kernel-based logistic regression models for protein function prediction, OMICS 10 (1) (2006) 40–55.
[21] H.N. Chua, W.-K. Sung, L. Wong, Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions, Bioinformatics 22 (13) (2006) 1623–1630.
[22] S. Mostafavi, Q. Morris, Combining many interaction networks to predict gene function and analyze gene lists, Proteomics 12 (10) (2012) 1687–1696.
[23] M. Hulsman, C. Dimitrakopoulos, J. de Ridder, Scale-space measures for graph topology link protein network architecture to function, Bioinformatics 30 (12) (2014) i237–i245.
[24] O.G. Troyanskaya et al., A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae), Proc. Natl. Acad. Sci. 100 (14) (2003) 8348.
[25] R.L. Tatusov, M.Y. Galperin, D.A. Natale, E.V. Koonin, The COG database: a tool for genome-scale analysis of protein functions and evolution, Nucleic Acids Res. 28 (4) (2000) 33–36.
[26] W. Tian et al., Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function, Genome Biol. 9 (Suppl. 1) (2008) S7.
[27] M. Tasan et al., An en masse phenotype and function prediction system for Mus musculus, Genome Biol. 9 (Suppl 1) (2008) S8.
[28] O.G. Troyanskaya et al., A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae), Proc. Natl. Acad. Sci. 100 (14) (2003) 8348–8353.
[29] Z. Barutcuoglu, R.E. Schapire, O.G. Troyanskaya, Hierarchical multi-label prediction of gene function, Bioinformatics 22 (7) (2006) 830–836.
[30] A. Vinayagam et al., Applying Support Vector Machines for Gene Ontology based gene function prediction, BMC Bioinformatics 5 (12) (2004) 1057–1065.
[31] S.F. Altschul et al., Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.
[32] S.F. Altschul et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.
[33] W. Tian, J. Skolnick, How well is enzyme function conserved as a function of pairwise sequence identity?, J Mol. Biol. 333 (4) (2003) 863–882.
[34] L.B. Koski, G.B. Golding, The closest BLAST hit is often not the nearest neighbor, J. Mol. Evol. 52 (6) (2001) 540–542.
[35] Z. Wang, R. Cao, J. Cheng, Three-level prediction of protein function by combining profile-sequence search, profile-profile search, and domain co-occurrence networks, BMC Bioinformatics 14 (Suppl 3) (2013) S3.
[36] D.A. de Lima Morais et al., SUPERFAMILY 1.75 including a domain-centric gene ontology method, Nucleic Acids Res. (2011).
[37] R. Rentzsch, C.A. Orengo, Protein function prediction using domain families, BMC Bioinformatics 14 (Suppl 3) (2013) S5.
[38] W. Tian, A.K. Arakaki, J. Skolnick, EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference, Nucleic Acids Res. 32 (21) (2004) 6226–6239.
[39] Gene. Ontology, C., *Gene Ontology annotations and resources*, Nucleic Acids Res. 41 (D1) (2013) D530–D535.
[40] B.E. Suzek et al., UniRef: comprehensive and non-redundant UniProt reference clusters, Bioinformatics 23 (10) (2007) 1282–1288.
[41] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics 5 (4) (2004) 1–19.
[42] P. Radivojac et al., A large-scale evaluation of computational protein function prediction, Nat. Methods 10 (3) (2013) 221–227.
[43] K.L. Hedstrom et al., Neurofascin assembles a specialized extracellular matrix at the axon initial segment, J. Cell Biol. 178 (5) (2007) 875–886.