Bioinformatics, 35(5), 2019, 753–759
doi: 10.1093/bioinformatics/bty704
Advance Access Publication Date: 25 August 2018
Original Paper

Sequence analysis

Phylo-PFP: improved automated protein function prediction using phylogenetic distance of distantly related sequences

Aashish Jain[1] and Daisuke Kihara [1,2,*]

[1]Department of Computer Science and [2]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on April 5, 2018; revised on July 30, 2018; editorial decision on August 10, 2018; accepted on August 23, 2018

Abstract

Motivation: Function annotation of proteins is fundamental in contemporary biology across fields including genomics, molecular biology, biochemistry, systems biology and bioinformatics. Function prediction is indispensable in providing clues for interpreting omics-scale data as well as in assisting biologists to build hypotheses for designing experiments. As sequencing genomes is now routine due to the rapid advancement of sequencing technologies, computational protein function prediction methods have become increasingly important. A conventional method of annotating a protein sequence is to transfer functions from top hits of a homology search; however, this approach has substantial short comings including a low coverage in genome annotation.

Results: Here we have developed Phylo-PFP, a new sequence-based protein function prediction method, which mines functional information from a broad range of similar sequences, including those with a low sequence similarity identified by a PSI-BLAST search. To evaluate functional similarity between identified sequences and the query protein more accurately, Phylo-PFP reranks retrieved sequences by considering their phylogenetic distance. Compared to the Phylo-PFP's predecessor, PFP, which was among the top ranked methods in the second round of the Critical Assessment of Functional Annotation (CAFA2), Phylo-PFP demonstrated substantial improvement in prediction accuracy. Phylo-PFP was further shown to outperform prediction programs to date that were ranked top in CAFA2.

Availability and implementation: Phylo-PFP web server is available for at http://kiharalab.org/phylo_pfp.php.

Contact: dkihara@purdue.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Proteins are drivers of almost all biological processes in the cell. Therefore, elucidating function of an individual protein is key to understanding how a biological system operates through functional interactions of component proteins. Ultimately, the biological function of a protein needs to be determined experimentally; however, a hypothesis is needed to design an assay that determines whether a target protein has a particular function. Computational function prediction can provide valuable information when biologists build such hypotheses. As genome sequencing has become routine due to the rapid advancement of sequencing technologies (Mardis, 2013), function prediction has become increasingly important. Computational function prediction methods are also useful for analyzing omics data including gene expression and protein–protein interaction data.

In addition to function prediction methods that use protein sequence information, there are other types of methods that consider gene co-expression patterns, phylogenetic profiles, three dimensional (3D) structures of proteins, as well as protein–protein interaction networks (Hawkins and Kihara, 2007). These non-sequence-based methods can often identify functional relationships of proteins that are not obvious from sequence similarity. However, non-sequence information is not always available and thus has limited applicability.

Recently there is an increasing momentum for developing function prediction methods driven by successful organization of a community-wide objective assessment of protein function prediction, the Critical Assessment of Function Annotation (CAFA) (Jiang *et al.*, 2016; Radivojac *et al.*, 2013). In CAFA, participants predict function (GO terms or other ontology terms specified by the organizers) of many target proteins (48298 and 100816 proteins in CAFA1 and CAFA2, respectively). Then, predictions are evaluated only for newly annotated GO terms to the target proteins after a waiting period of over six months from the prediction submission. This process is designed for assessing methods' capability of predicting new functions rather than retrieving known functions from existing data sources. Three rounds of CAFA have been held so far, CAFA1 in 2010–2011, CAFA2 in 2013–2014 and CAFA3 in 2016–2017, for which the official evaluations were reported for the first two.

PFP (Hawkins *et al.*, 2006, 2009) is one of the pioneer methods, which makes use of sequences with a wide range of similarity to a query ranging from significant hits to very weakly similar ones up to an E-value of 125, far larger than conventionally used thresholds, e.g. 0.001. GO terms are extracted from all the retrieved sequences; however, to reduce the risk of predicting unrelated GO terms taken from weakly similar sequences, sequences are weighted by their E-values. PFP also considers the co-occurrence of GO terms, which is statistics of GO term pairs that frequently co-occur in annotation of the same sequence. PFP was one of the top ranked function prediction methods in CAFA and the top in the Critical Assessment of Protein Structure Prediction (CASP) function prediction category in 2007 (Lopez *et al.*, 2007).

Here, we present a new method, Phylo-PFP, which significantly improves prediction performance over PFP by incorporating phylogenetic information in defining sequence similarity. We first show that the E-values of the sequences do not largely agree with the distances defined by phylogenetic trees to a surprising extent. Then, we show that weighting sequence by considering the phylogenetic distance can substantially improve GO term prediction accuracy. Predictions by Phylo-PFP were evaluated on a dataset of 1702 non-redundant protein sequences and showed better performance than the original PFP as well as several other existing methods. To compare its performance among the best programs available to date, Phylo-PFP was used to predict functions of target sequences in CAFA2. We show that Phylo-PFP outperforms all the top methods used in CAFA2, having the highest score in all three GO categories, Molecular Function (MF), Biological Process (BP) and Cellular Component (CC).

2 Materials and methods

2.1 Overview of the phylo-PFP method

Figure 1 illustrates the workflow of Phylo-PFP. For a query protein sequence, Phylo-PFP searches similar sequences from a reference sequence database with PSI-BLAST (maximum iteration set to 3). In this retrieval, top 500 sequences are retrieved or until an E-value of up to 125 is reached. Collecting diverse sequences with a large E-value has two advantages: First, as demonstrated in the original version of PFP, the E-value cutoff will capture a larger breadth of sequences, which is particularly effective when closely annotated homologs to the query do not exist in the database. Also, for Phylo-PFP, having many sequences help in constructing meaningful phylogenetic trees, which is a key new feature of the Phylo-PFP algorithm.

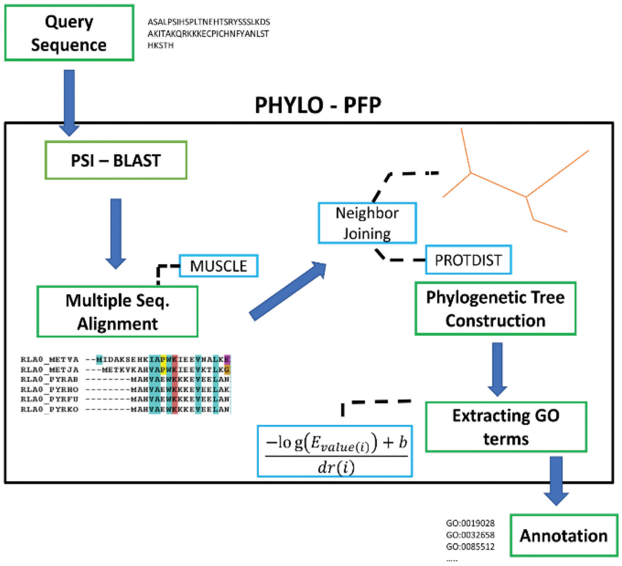
Query
Sequence
ASALPSIHSPLTNEHTSRYSSSLKDS
AKITAKQRKKKECPICHNFYANLST
HKSTH
PHYLO - PFP
PSI – BLAST
MUSCLE
Multiple Seq.
Alignment
RLA0_METVA --MIDAKSEHKIAPWKIEEVNALKE
RLA0_METJA ---METKVKAHVAPWKIEEVKTLKG
RLA0_PYRAB --------MAHVAEWKKKEVEELAN
RLA0_PYRHO --------MAHVAEWKKKEVEELAK
RLA0_PYRFU --------MAHVAEWKKKEVEELAN
RLA0_PYRKO --------MAHVAEWKKKEVEELAN
Neighbor
Joining
PROTDIST
Phylogenetic Tree
Construction
$\frac{-\log g(E_{value(i)})+b}{dr(i)}$
Extracting GO
terms
GO:0019028
GO:0032658
GO:0085512
.....
Annotation

Fig. 1. Overview of Phylo-PFP algorithm

The next step of Phylo-PFP is to rank retrieved sequences using a weighting factor that considers the phylogenetic distance among them. This step is the key difference from the original PFP, which simply uses the raw E-value to rank sequences. There are three steps in constructing a phylogenetic tree: (i) A multiple sequence alignment (MSA) is computed for the retrieved sequences using MUSCLE (Edgar, 2004). (ii) From the MSA, a pairwise sequence alignment for each sequence pair is extracted, from which a distance matrix is computed using PROTDIST (Felsenstein, 1981) in the PHYLIP package with the Jones-Taylor-Thornton model. (iii) With the set of computed distances, a phylogenetic tree is constructed using the neighbor joining (NJ) method implemented in PHYLIP. Following the tree construction, a distance *dr* is defined between the query protein and each protein *k* as the sum of the branch lengths between them on the tree, which is scaled to a value between 0 and 100 as

$$dr = \frac{distance(k) - \min_i distance(i)}{\max_i distance(i) - \min_i distance(i)} {*}100 \qquad (1)$$

Using the phylogenetic distance *dr* and the E-value, a retrieved sequence *i* is ranked with a weight named the Evolutionary distance-normalized Log E-value (ELE) in the descending order:

$$ELE(i) = \frac{-\log_{10}(E - value(i)) + b}{dr(i)} \qquad (2)$$

where *E-value(i)* is the E-value of the sequence *i*, *b* is the constant, $\log_{10}(125)$, which is an offset added to make the numerator of the equation a non-negative value up to an E-value of 125, and *dr*(*i*) is the phylogenetic distance of the sequence *i*. The numerator is the weight used in the original PFP. In Phylo-PFP, the numerator is normalized by *dr*(*i*), i.e. a sequence that has a large distance on the phylogenetic tree receives a discounted weight, which brings the contribution of the sequence lower when the score for predicted GO terms are computed. Using ELE, a function (GO term) f_a is scored for a query sequence as

$$s(f_a) = \sum_{i=1}^{N} \sum_{j=1}^{Nfunc(i)} \left(ELE(i)P(f_a|f_j)\right) \tag{3}$$

where N is the number of sequences retrieved from the sequence database within an E-value of 125, *Nfunc(i)* is he number of GO terms annotating the sequence *i*, *ELE(i)* is the weight defined in Eq. 2, and $P(f_a|f_j)$ is the functional association (Hawkins *et al.*, 2009), a conditional probability that GO term f_a is in annotation of a sequence that is also annotated with GO term f_j. The function association allows predicting GO terms that do not appear in annotations of retrieved sequences. Associations are also computed between terms across different categories, e.g. terms in MF and BP. Associations with a probability of 0.9 or higher were considered. Each GO term in the final prediction is also given a confidence score, which is computed by normalizing ELE for all GO terms belonging to the same category. The Eq. 2 is an update from the original PFP score (Hawkins *et al.*, 2009). In PFP, the score is

$$s(f_a) = \sum_{i=1}^{N} \sum_{j=1}^{Nfunc(i)} (-\log(E - value(i)) + \ b)P(f_a|f_j) \tag{4}$$

where *E-value(i)* is the E-value of sequence *i*. In Supplementary Table S1, the computational time of Phylo-PFP in comparison with PFP and PSI-BLAST is provided.

2.2 Constructing the annotation database

For any function prediction method, it is crucial to have a comprehensive annotation database that keeps known GO terms for sequences, as the method depends on it in extracting GO terms from PSI-BLAST hits. We integrated several data sources to form our annotation database for Phylo-PFP. The primary database used was the UniProtKB/Swiss-Prot including Non-IEA (Inferred from Electronic Annotation) annotations (Boutet *et al.*, 2016). In addition we integrated annotations from UniPathway (Morgat *et al.*, 2012), TIGRFAMs (Haft *et al.*, 2013), SMART (Letunic and Bork, 2018), Reactome (Fabregat *et al.*, 2018), PROSITE (Sigrist *et al.*, 2013), ProDom (Bru *et al.*, 2005), PRINTS (Attwood, 2012), PIRSF (Nikolskaya *et al.*, 2006), Pfam (Finn *et al.*, 2016), InterPro (Finn *et al.*, 2017) and HAMAP (Pedruzzi *et al.*, 2015).

2.3 Non-redundant benchmark dataset

Target sequences for the benchmark dataset were selected from UniProt Reference Clusters (UniRef), which provides a clustered set of sequences from UniProt Knowledgebase (Suzek *et al.*, 2015). We used the UniRef50 clusters of 8/25/2016, in which sequences with more than 50% identity to each other are clustered. We selected a representative sequence from each cluster which fulfills two conditions: a cluster must include more than 1500 sequences, and the representative protein is annotated in UniProt. Representative sequences were removed if it had more than 500 hits with an E-value 0.0 in the third round of PSI-BLAST as these sequences have many highly similar sequences which makes their function prediction easy. This procedure yielded 1702 sequences for the benchmark dataset. We also constructed another benchmark dataset by clustering these 1702 sequences with 30% sequence identity cutoff.

2.4 CAFA2 dataset

We also tested Phylo-PFP on the dataset from CAFA2 (Jiang *et al.*, 2016). CAFA2 released 100 816 target protein sequences but predictions were evaluated only for 1776 sequences which newly accumulated GO terms during the waiting period. Among the 1776 sequences, 419 sequences had MF GO terms, 860 sequences had BP GO terms and 1259 sequences had CC GO terms. For replicating participation in CAFA2 with Phylo-PFP, the benchmark sequence dataset as well as the ground truth of the annotation were obtained from the Supplementary data at https://figshare.com/articles/Supplementary_Data_for_CAFA2/2059944/1. When we ran Phylo-PFP, we used the UniProt database of August 2013 (a version released before the CAFA2 target sequences were released to participants), so that annotations newly added after the release were not included.

3 Results

We first discuss that relationship between the E-value and phylogenetic distance of sequences. Then, we present prediction results of Phylo-PFP on the two datasets.

3.1 New sequence weight and sequence similarity

First, we examined to what extent the new sequence weight ELE (Eq. 2) correlates sequence similarity score computed with E-value used in the original PFP. In phylogenetic studies, difference between sequence similarity scores and the phylogenetic distance has been a focus of interest (Cantarel *et al.*, 2006; Smith and Pease, 2017). Smith and Pease discussed cases when a sequence similarity score, *−log(E-value)*, which is also used in the original PFP, does not capture evolutionary related sequences (Smith and Pease, 2017). Eisen showed examples when the phylogenetic distance is expected to perform better than a sequence similarity-based score in predicting gene function (Eisen, 1998).

Figure 2a shows the distribution of the Pearson's correlation coefficients between ELE and −log*(E-value)* for PSI-BLAST hits for 1702 sequences in the benchmark dataset. For each query sequence in the benchmark dataset, similar sequences to the query were retrieved from the database with PSI-BLAST up to an E-value of 125, and correlation between the sequence similarity score and ELE was computed and summarized in a histogram. For cases of very similar sequences to the query with an E-value of 0, a very small number (1e−1000) was assigned.

Although there is a small peak at the highest correlation bin of 1.0, the highest peak in the plot was observed at a very weak correlation of around 0.1. 53.58% were less than 0.2. Due to these very weak correlation, the mean correlations values were modest, 0.234.

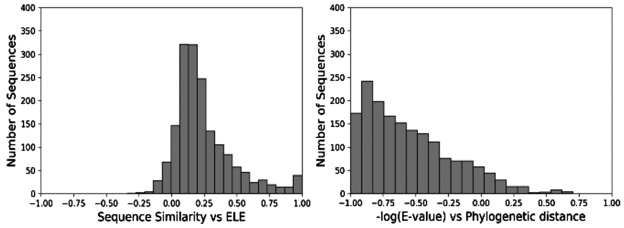

Fig. 2. (**a**) Histogram of Pearson's correlation coefficients computed between −log*(E-value)* and ELE of PSI-BLAST hits for the dataset of 1702 sequences. (**b**) Histogram of Pearson's correlation coefficients between −log*(E-value)* and the phylogenetic distance

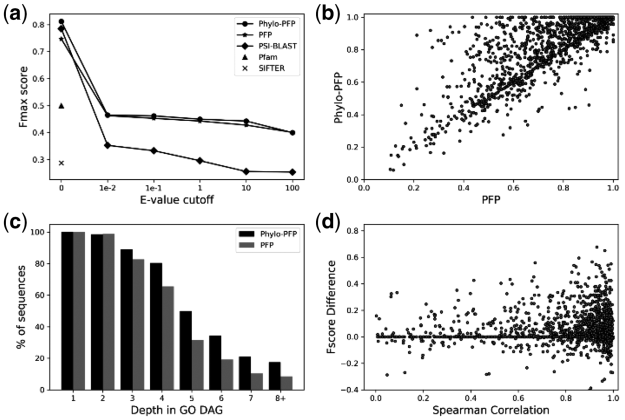

Fig. 3. Prediction performance of Phylo-PFP on the benchmark dataset of 1702 non-redundant proteins. (**a**) Performance comparison with different E-value cutoffs applied to PSI-BLAST hits in terms of the Fmax score. Phylo-PFP (circles) was compared with PFP (stars), PSI-BLAST (diamonds), Pfam (triangle) and SIFTER (cross). Sequence hits that have an E-value smaller (i.e. more significant) than the E-value cutoff are removed and not used for extracting GO terms. (**b**) Comparison of predictions by Phylo-PFP and PFP for individual proteins. Fmax scores were compared. (**c**) The depth of correctly predicted GO terms with an E-value cutoff of 1e−2 by Phylo-PFP and PFP were compared. The x-axis represents the depth of the correctly predicted GO terms in the GO graph. If a GO term has multiple parental terms with different depths, the smallest depth for the term was considered. Predictions with a confidence score of 0.9 or higher were considered. If a sequence had multiple correctly predicted GO terms of different depths, the sequence was counted for all the depths. The right most bars, 8+, are for depths of 8 or larger. (**d**) Difference of Fmax scores of Phylo-PFP and PFP against the Spearman's correlation between the PSI-BLAST hits ranks of the two methods. Each data point corresponds to a protein sequence in the benchmark dataset

The same trend was observed in Supplementary Figure S1, which is a histogram of correlation between the BLAST Bit score and ELE. Smith and Pease showed similar plots of correlation between the evolutionary distance and –log*(E-value)* for simulated protein sequences (Smith and Pease, 2017), which corresponds to Figure 2b in our analysis. Compared with their results (Fig. 3a and b in their paper), which showed high correlations between the two values, our results on real sequences show a diverse distribution of correlations including many cases that had almost no correlation. The average correlation in Figure 2b was −0.546. The different result between the plots by Smith and Pease and Figure 2b in this work is probably due to the different ways that the sequence datasets were constructed. The sequence dataset in the former work was simulated based on a molecular model (Smith and Pease, 2017) while in the current work sequences were collected from real database searches up to a very weak similarity of an E-value of 125. Another difference is that while the plots by Smith and Pease are computed only for two query proteins, the current work summaries 1702 proteins showing that there are sequences of a high correlation but with a larger number of sequences with weak correlations. In Supplementary Figure 2, three examples of proteins with a high and low correlations are shown.

3.2 Performance of phylo-PFP on the non-redundant benchmark dataset

Next, we evaluated prediction performance of Phylo-PFP on the benchmark dataset. The prediction performance was compared with the original PFP and three other existing methods as reference, PSI-BLAST, Pfam and SIFTER (Sahraeian *et al.*, 2015). Details of these methods are provided in Supplementary Note S1. SIFTER was chosen because it considers a phylogenetic tree to transfer function from similar proteins to the query. When we ran Phylo-PFP and PFP, we removed the query sequence itself and sequence hits with an E-value of 0 from the PSI-BLAST run. The performance of the methods were evaluated with the Fmax score, which is the average F1-score at a method's score cutoff that gives the maximum F1-score to the entire set of target proteins (i.e. the method's score cutoff was not optimized differently for each target. See Supplementary Note S2). Fmax score was used because it is a main evaluation metric used in CAFA.

Table 1. The Fmax score of the five methods on the benchmark dataset

Method	Fmax (no cutoff)	Fmax 1e−2 cutoff
Phylo-PFP	0.812	0.465
PFP	0.747	0.463
PSI-BLAST	0.785	0.353
PFam	0.500	–
SIFTER	0.288	–

Note: Fmax scores were computed at two E-value cutoffs of PSI-BLAST search, with no cutoff and 1e−2. Only one score was provided for Pfam and SIFTER since they do not use a database search results from PSI-BLAST.

Table 1 summarizes the Fmax score of the five methods. Phylo-PFP showed the highest Fmax score, 0.812 followed by PSI-BLAST with a score of 0.785. The rest of the methods were ranked in the order of PFP, Pfam and SIFTER. The performance difference between Phylo-PFP and the other methods was statistically significant (Supplementary Table S2). For further comparison, in Figure 3a we removed sequence hits up to a certain E-value, 1e−2, 1e−1, 1, 10 and 100 from the PSI-BLAST search for the three methods, Phylo-PFP, PFP and PSI-BLAST, and predicted GO terms from remaining sequence hits. This is to simulate situations when a query protein does not find any significant hits. Pfam and SIFTER results do not change by E-value cutoffs, because PSI-BLAST is not used in their algorithms. It is apparent that Phylo-PFP and PFP performed substantially better than PSI-BLAST. At an E-value cutoff of 1e−2, the Fmax scores of the three methods were 0.465, 0.463 and 0.353, respectively (Table 1). It caught our attention that the Fmax score of PFP was worse than PSI-BLAST with no cutoff, which is probably due to the nature of this particular benchmark dataset, where query sequences have a sufficient number of highly similar sequences because they are collected from clusters of Uniref50. However, when sequence hits were limited to an E-value of 1e−2 or lower, PFP showed its superior ability to PSI-BLAST as consistent with the earlier benchmark studies of PFP (Chitale *et al.*, 2013; Hawkins *et al.*, 2006, 2009). Comparing Phylo-PFP and PFP, Phylo-PFP performed better with no cutoff (0 in the plot) and cutoffs of 1e−2, 1e−1, 1 and 10. The margin between the two methods was largest when no E-value cutoff was applied. This implies that the sequence hits reranking with the ELE weight was more effective when closely similar sequences were more correctly ranked. At the cutoff of 100, Phylo-PFP and PFP showed almost identical Fmax score, 0.400, and 0.401, respectively.

As described in Section 2, Phylo-PFP uses an E-value cutoff of 125 for retrieving sequences. Supplementary Table S3 provides results using two other cutoff values, 100, and 150, which gave similar results but 125 had the highest Fmax score among them.

Additionally, we also used HHblits (Remmert *et al.*, 2011) and MMseqs2 (Steinegger and Söding, 2017) instead of PSI-BLAST in Phylo-PFP with the same parameters as we used PSI-BLAST, i.e. up to

Table 2. Confidence scores of correct GO terms for P03423 by Phylo-PFP and PFP

Correct GO terms	Phylo-PFP Confidence Score	PFP Confidence Score
GO: 0044228 (Host cell surface)	0.99	0.06
GO: 0055036 (virion membrane)	0.99	0.06
GO: 0046718 (viral entry into host)	1.00	0.11
GO: 0005576 (extracellular region)	0.99	0.13
GO: 0016021 (integral component of membrane)	1.00	0.09
GO: 0030683 (evasion or tolerance by virus of host immune response)	1.00	0.11
GO: 0019062 (virion attachment to host cell)	1.00	0.11
GO: 0019012 (virion)	1.00	0.12
GO: 0016032 (viral process)	1.00	0.11
GO: 0016020 (membrane)	0.694	0.17
GO: 0046462 (diaminopimerate metabolic process)	0.55	1.00
GO: 0009089 (lysine biosynthetic process via diaminopimerate)	0.55	1.00

Note: The last two GO terms are incorrect terms, predicated with the maximum confidence by PFP.

three iterations and an E-value cutoff of 125 with 500 maximum sequence hits. Interestingly, Phylo-PFP with MMseqs2 exceeded the Phlyo-PFP's performance with a Fmax of 0.842. Comparison of the two methods for each benchmark sequence are shown in Supplementary Figure S3. Phylo-PFP-HHblits had an Fmax score of 0.633.

We further tested the methods on a dataset with 30% identity cutoff, which included 1234 sequences. The results were consistent with Table 1 (Supplementary Table S4).

In the subsequent panels in Figure 3, we analyzed the difference between Phylo-PFP and PFP from several different angles. Figure 3b shows a direct comparison of Fmax score of individual proteins in the benchmark dataset. Phylo-PFP showed larger or the same Fmax score than PFP for 83.72% of the sequences. Often the gain by Phylo-PFP over PFP was large; for 89 (5.23%) sequences the improvement of the score was more than 0.3 and the maximum Fmax score increase observed was 0.677 (from0.212 to 0.889). Phylo-PFP achieved the perfect score of 1.0 for 529 proteins while it was 338 for PFP. On the other hand, the deterioration of the score by Phylo-PFP was relatively small. For only 5 (0.29%) sequences the decrease in the score was more than 0.3.

In Figure 3c, we examined the information content of predicted functions by Phylo-PFP and PFP quantified as the depth of correctly predicted GO terms. GO terms are organized in a directed acyclic graph ordered from general functional terms to more specific functions (Consortium, 2015). Thus, correct predictions of GO terms at larger depth (closer to leaves) are more valuable than prediction of shallower GO terms. In the plot, the results from the E-value cutoff of 1e−2 (Fig. 3a) was used and only high confidence predictions with a confidence level over 0.9 were considered. It is shown in Figure 3c that Phylo-PFP predicted more terms at larger depths than PFP. Phylo-PFP predicted correct GO terms at a depth of five or deeper for 1.58 times more sequences than PFP. When only depths of eight or deeper were considered, Phylo-PFP predicted GO terms in 2.03 times as many cases as PFP.

In the last panel, Figure 3d, we examined the difference of the prediction performance (Fmax score) for each target protein between Phylo-PFP and PFP relative to the amount of the difference in the ranks of PSI-BLAST-retrieved sequences. Since Phylo-PFP and PFP use the same set of retrieved sequences from a PSI-BLAST search with only difference being ranking of the sequences due to the different scoring schemes used by the two methods, the performance difference may be correlated to the difference of the sequence ranks. The difference of the sequence rankings was evaluated by the Spearman's correlation (x-axis). We expected that a large improvement of prediction accuracy occurs when a large sequence ranking difference is observed, which should result in a small correlation coefficient. However, the trend seems to be rather opposite. Large Fmax score improvements were observed more frequently when the correlation values are close to 1.0, which indicates a small difference in the sequence rankings of the two methods. This may be implying that an improvement occurs when a small number of key sequences are adjusted in their ranks. In Supplementary Note S5, we further tested statistical significance of rank change by Phylo-PFP, and found a significant change for only 13.6% of the cases.

3.3 Case studies

In this section we discuss an illustrative case of Phylo-PFP's prediction. The focus is to examine how Phylo-PFP improved prediction over PFP by reranking PSI-BLAST sequence hits by the ELE weight.

The query protein used is human respiratory virus surface glycoprotein G (UniProt ID: P03423). This protein is present on the virus surface, for which GO terms such as virion membrane (GO: 0055036), virion (GO: 0019012), host cell surface (GO: 0044228), extracellular region (GO: 0005576), integral component of membrane (GO: 0016021) and membrane (GO: 0016020) are annotated in the CC category. The protein helps in attachment of the virus to the host cell membrane by interacting with heparan sulfate, initiating viral infection. This corresponds to GO annotations of virion attachment to host cell (GO: 0019062), viral process (GO: 0016032), evasion/tolerance by virus of host immune response (GO: 0030683) and viral entry into host cell (GO: 0046718) in the BP category. Phylo-PFP showed a high prediction accuracy, an Fmax score of 0.803 while it was 0.042 by PFP. If we calculate Fmax using the optimal score cutoff for this particular protein, then Phylo-PFP score was increased to 0.958, while PFP score increased to 0.741, still lower Phylo-PFP.

As shown in Table 2, Phylo-PFP predicted most of the correct GO terms with a high confidence score of 0.99–1.00, while PFP predicted them with a low score of 0.06–0.11. PFP instead predicted the incorrect terms, diaminopimelate metabolic (GO: 0046451) and lysine biosynthetic process via diaminopimelate (GO: 0009089) with the highest score of 1.0. These two incorrect GO terms came from sequence hits of diaminopimelate epimerase, which had a significant E-value (e.g. 1e−26 for bacterial diaminopimelate epimerase, UniProt ID: A6VQR8). In contrast to PFP, Phylo-PFP moved the ranks of Epstein-Barr virus envelope glycoproteins (e.g. Q3KST4, P03200 and P68344) higher, which are virus envelope proteins similar to the query protein. Figure 4 depicts how the

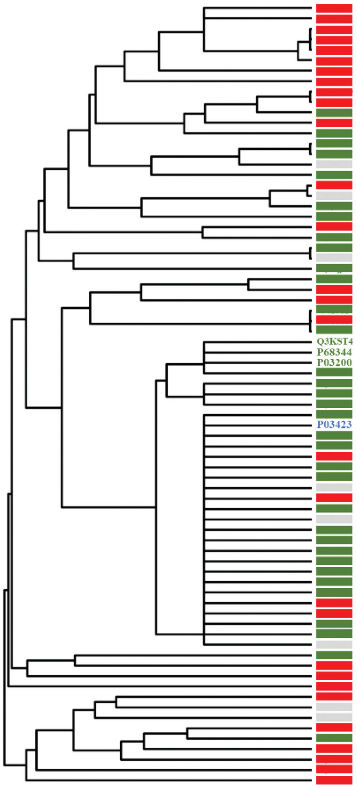

Fig. 4. Visualization of sequence rank changes by ELE used in Phylo-PFP relative to the functional similarity to the query protein. The dendrogram shows functional similarity of each sequence to the query protein (shown in blue), which was quantified with the funSim score (see Supplementary Note S4) of GO terms annotations of two proteins. Top 75 sequences of highest functional similarity to a query protein are shown. In comparison with sequence hit ranks in PFP, sequences that went up or down in their ranking in Phylo-PFP are shown in green and red, respectively. UniProt IDs are shown for proteins that are mentioned in the text. The query is human major surface glycoprotein G (UniProt ID: P03423)

Table 3. Comparison of E-value, phylogenetic distance and ELE of a few key PSI-BLAST hits for a query protein, P03423

Prot. ID	E-val.	Rel(PFP)[a]	Phyl[b]	Rel(ELE)[c]	Func.
A6VQR8	1E−26	1.0	31.6	1.0	D.e[d]
Q5N013	2E−22	3	35.7	0.748	D.e
Q3KST4	2.2	0.027	13.9	0.142	E.-b.[e]
P03200	0.23	0.097	15.8	0.198	E.-b.
P68344	0.55	0.084	14.1	0.188	E.-b.

[a] The weight of the sequence used in PFP, i.e. –log(E-value) + b, relative to A6VQR8.

[b] The phylogenetic distance.

[c] ELE relative to A6VQR8. Supplementary Table S5 provides more data for these five sequences.

[d] Diaminopimelate epimerase.

[e] Epstein-Barr virus envelope glycoprotein.

Table 4. The Fmax score of predictions for the CAFA2 dataset by Phylo-PFP in comparison with top performing methods in CAFA2

Method	MF	BP	CC
MS-KNN	0.595	0.363	0.455
EVEX	0.593	–	0.468
Paccanaro Lab	–	0.372	–
Tian Lab	0.591	0.367	0.462
Orengo-FunFams	0.569	0.352	0.438
Go2Proto	0.563	–	–
SIFTER	0.561	–	–
INGA-Tosatto	0.555	0.347	–
Jones-UCL	0.554	0.352	0.450
Argot2	0.544	0.351	–
Gough Lab	–	0.352	0.458
PULP	–	0.350	0.441
Rost Lab	–	–	0.442
IASL	–	–	0.439
PFP	0.574	0.348	–
CONS	–	–	0.446
BLAST	0.473	0.251	0.347
Phylo-PFP	**0.606**	**0.380**	**0.506**

Note: Results of the three GO categories are separately shown. Fmax scores of the methods participated in CAFA2 were taken by matching the Supplementary data and Figure 4 of the CAFA2 evaluation report. Dashes (–) indicate that method did not appear among top 10 methods in Figure 4. The largest Fmax value for each GO category is highlighted in bold.

sequence hits in PSI-BLAST search were reranked by ELE. The dendrogram shows the top 75 functionally similar sequences to the query, P034223 (shown in blue). Sequence hits are shown in green if their ranks went up by ELE in comparison with their original ranks in PSI-BLAST, which include the three proteins, Q3KST4, P03200 and P68344. Shown in red are sequences whose rank went lower by ELE. As illustrated, sequences that are less similar, i.e. far from the query in the dendrogram, went lower, while those more functional similar went up in the rank. Table 3 further illustrates the amended score contribution by ELE with a few sequence examples. The three virus proteins in the table had insignificant E-values of 2.2, 0.23 and 0.55 respectively, and thus only contributed 2–10% of the scores relative to A6VQR8, a diaminopimelate epimerase sequence with a very small E-value. However, their relative contribution increased to 14–20% in ELE, which was sufficient, together with contributions of other functionally similar sequences to the query, P03423 (Fig. 4), to rank correct GO terms with the highest confidence scores (Table 2). In Supplementary Note S3, we discussed another case with sarcosine oxidase subunit β from *Corynebacterim sp. strain P-1* (UniProt ID: P40875).

3.4 Prediction on the CAFA2 target protein dataset

We further tested Phylo-PFP on the target protein sequence dataset used in CAFA2 to compare the performance with top performing methods in the assessment. In total, 56 groups submitting 126 methods participated in CAFA2.

We compared the performance of Phylo-PFP with the best performing methods in CAFA2 as well as PFP and with a baseline method, BLAST (Table 4). The top performing methods from CAFA2 were taken from Figure 4 of the CAFA2 evaluation report (Jiang *et al.*, 2016).

Remarkably, Phylo-PFP outperformed the other methods in all three categories with an Fmax score of 0.606, 0.380 and 0.506 for MF, BP and CC category, respectively. Considering the methods from CAFA2, a different method excelled for each GO category and no method showed consistent high performance among all the categories. MS-KNN scored the highest in MF with an Fmax of 0.595, Paccanaro Lab was the top among the existing methods in BP with an Fmax of 0.380, while EVEX was best in CC with an Fmax of 0.372. This is a clear contrast with Phylo-PFP, which exhibited the best performance in all the three categories.

We also compared the performance of Phylo-PFP on another metric used in CAFA2, the minimum sematic distance (S_{min}). Phylo-PFP performed fairly well placing at 3rd and 6th position in MF ad BP, respectively, but not within top 10 in CC (Supplementary Table S6).

4 Discussion

In this study, we developed a new sequence-based protein function prediction method, Phylo-PFP, which substantially improved the prediction accuracy from its predecessor, PFP, by using phylogenetics to determine the evolutionary distance of sequences retrieved from a database searches.

It has been discussed that the sequence similarity does not often accurately capture evolutionary relationship of sequences (Smith and Pease, 2017). Here we showed that there was no strong correlation between the database search scores and the phylogenetic distances for most of the sequences in a large dataset on a realistic scenario of PSI-BLAST search. Subsequently, as a practical solution for improving PFP, we implemented a distance-based phylogenetic analysis, and achieved favorable prediction accuracy improvements. Phylo-PFP takes more computational time than PFP especially when the number of PSI-BLAST hits is large (Supplementary Table S1). A practical solution for performing prediction for many sequences would be to run the method in parallel on multi-core CPUs.

Further improvement of the accuracy is expected by considering several approaches. For example, instead of the distance-based phylogenetic analysis we used in this work, a more accurate tree construction technique such as maximum likelihood (Yang, 2007) or Bayesian inference (Bouckaert *et al.*, 2014; Ronquist *et al.*, 2012) may be used. Also, functional domain (Messih *et al.*, 2012) or residue information (Gong *et al.*, 2016; Wass and Sternberg, 2008) can be explicitly considered, as currently functional transfer is performed in PFP and Phylo-PFP only by global sequence similarity.

Acknowledgements

The authors are grateful to Lyman Monroe for proofreading the manuscript.

Funding

This study is partly based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under cooperative Agreement Number W911NF-17-2-0105. DK also acknowledge supports from National Science Foundation (DBI1262189, DMS1614777) and National Institutes of Health (R01GM123055).

Conflict of Interest: none declared.

References

Attwood,T.K. (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012. *Database (Oxford)*, **2012**, bas019.

Bouckaert,R. *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.*, **10**, e1003537.

Boutet,E. *et al.* (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol. Biol.*, **1374**, 23–54.

Bru,C. *et al.* (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.*, **33**, D212–D215.

Cantarel,B.L. *et al.* (2006) Exploring the relationship between sequence similarity and accurate phylogenetic trees. *Mol. Biol. Evol.*, **23**, 2090–2100.

Chitale,M. *et al.* (2013) In-depth performance evaluation of PFP and ESG sequence-based function prediction methods in CAFA 2011 experiment. *BMC Bioinformatics*, **14**, S2.

Consortium,G.O. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.

Fabregat,A. *et al.* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Finn,R.D. *et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.

Finn,R.D. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

Gong,Q. *et al.* (2016) GoFDR: a sequence alignment based method for predicting protein functions. *Methods*, **93**, 3–14.

Haft,D.H. *et al.* (2013) TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.*, **41**, D387–3D395.

Hawkins,T. *et al.* (2009) PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins*, **74**, 566–582.

Hawkins,T. and Kihara,D. (2007) Function prediction of uncharacterized proteins. *J. Bioinform. Comput. Biol.*, **5**, 1–30.

Hawkins,T. *et al.* (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci.*, **15**, 1550–1556.

Jiang,Y. *et al.* (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.*, **17**, 184.

Letunic,I. and Bork,P. (2018) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.*, **46**, D493–D496.

Lopez,G. *et al.* (2007) Assessment of predictions submitted for the CASP7 function prediction category. *Proteins*, **69**, 165–174.

Mardis,E.R. (2013) Next-generation sequencing platforms. *Annu. Rev. Anal. Chem. (Palo Alto Calif)*, **6**, 287–303.

Messih,M.A. *et al.* (2012) Protein domain recurrence and order can enhance prediction of protein functions. *Bioinformatics*, **28**, i444–i450.

Morgat,A. *et al.* (2012) UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.*, **40**, D761–76D769.

Nikolskaya,A.N. *et al.* (2006) PIRSF family classification system for protein functional and evolutionary analysis. *Evol. Bioinform. Online*, **2**, 117693430600200–117693430600209.

Pedruzzi,I. *et al.* (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.*, **43**, D1064–D1070.

Radivojac,P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.

Remmert,M. *et al.* (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Ronquist,F. *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, **61**, 539–542.

Sahraeian,S.M. *et al.* (2015) SIFTER search: a web server for accurate phylogeny-based protein function prediction. *Nucleic Acids Res.*, **43**, W141–W147.

Sigrist,C.J. *et al.* (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–34D347.

Smith,S.A. and Pease,J.B. (2017) Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny, *Brief. Bioinform.*, **18**, 451–457.

Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.

Suzek,B.E. *et al.* (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.

Wass,M.N. and Sternberg,M.J. (2008) ConFunc – Functional Annotation in the Twilight Zone. *Bioinformatics*, **24**, 798–806.

Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.