TANA: efficient approach for predicting protein functions by transferring annotation via alignment networks

Warith Eddine DJEDDI (✉ warithsa5@gmail.com)

Universite de Tunis El Manar <https://orcid.org/0000-0001-5751-2371>

Sadok BEN YAHIA

Tallinn University of technology

Engelbert MEPHU NGUIFO

Universite Clermont Auvergne - Campus des Cezeaux

Methodology article

Keywords: PPI, networks, alignment, function prediction, neighborhood, gene ontology, computational assessment of function annotation

Posted Date: January 16th, 2020

DOI: <https://doi.org/10.21203/rs.2.21071/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

TANA: efficient approach for predicting protein functions by transferring annotation via alignment networks

Warith Eddine DJEDDI^{1*}, Sadok BEN YAHIA^{1,2} and Engelbert MEPHU NGUIFO³

*Correspondence:

warithsa5@gmail.com

¹University of Tunis El Manar,

Faculty of Sciences of Tunis,

LR11ES14, Campus

Universitaire, 2092 Tunis, Tunisia

Full list of author information is available at the end of the article

Abstract

Background: One of the challenges of the post-genomic era is to provide accurate function annotations for orphan and unannotated protein sequences. With the recent availability of huge protein-protein interactions networks for many model species, the computational methods revealed a great requirement to elucidate protein function based on many strategies. In this respect, most computational approaches integrate diverse kinds of functional interactions to unveil protein functions by transferring annotations across different species by relying on similar sequence, structure 2D/3D, amino acid motifs or phylogenetic profiles.

Results: In this work, we introduce a new approach called TANA for inferring protein functions. The main originality of the introduced approach stands on the function prediction for the unannotated protein by transferring annotation via a network alignment as well as from the direct interaction neighborhood within their PPI networks. Doing so, we are able to discover the functions of proteins that could not to be easily described by sequence homology. We assess the performance of our method using the standard metrics established by the CAFA and highlight a sharp significant improvement over other competitive methods, in particular for predicting molecular functions.

Conclusions: This research is one of the first attempts that combine sequence and networks-multiple-alignment-based function prediction approaches. We have been able to assess the accuracy of the prediction using pairwise and multiple alignment of the PPI networks for the compared species. Therefore, we recommend using different strategies (i.e pairwise, multiple, with/without neighborhood networks) especially in situations where the functions of the protein are not known in advance.

Keywords: PPI; networks; alignment; function prediction; neighborhood; gene ontology; computational assessment of function annotation

Background

The past decade has witnessed an increasing in genomic and proteomic data, leading to a large variety of sequenced proteins. A fundamental challenge is the interpretation of this overwhelming data in order to elucidate more accurate protein functions. The manual annotation of protein function is a daunting task which paves the way to the emergence of successful computational predictive methods. The latter have been applied starting from incorporating gene expression patterns [1], phylogenetic profiles [2, 3], protein sequences [4, 5], protein structures [6, 7], and protein interactions [8]. A wealthy number of com-

putational approaches for predicting function from networks have been proposed can be organized into two major classes: (i) those using a direct network-context: to wit the direct annotation of proteins infer functions based on its connections in the network; (ii) those assisted by a prediction module. The latter first identifies clusters, or modules, of related proteins and then annotates each protein based on the known functions of its members [9].

There are many strategies that have been developed to assign a function to an unknown protein:

- **Gene expression pattern:** protein function prediction by analyzing gene expression patterns [10].
- **Phylogenetic profile:** analyzing phylogenetic profile, i.e. evolutionary history of proteins [11].
- **Protein sequence:** protein function prediction using protein sequence, sequence similarity measures, homologies are primarily used [12, 13].
- **Protein structure:** the function prediction using protein structure, by using some approaches to analyze the secondary [14, 15, 16] and tertiary structures [17, 6] of proteins.
- **Protein-protein interactions (PPIs):** protein function prediction using protein-protein interactions [18, 19, 20, 21, 22, 23, 24, 25] can be deduced from the interaction of the neighborhood. Chua et al. [26] demonstrated the useful strategies using the PPIs as a complementary approach to sequence homology by specifying the maximum additional coverage for the protein-protein interactions.
- **Network alignment:** the function prediction using annotation transfers via network alignment by conferring the annotations of a protein in an aligned cluster to the unannotated member of the identical cluster [27, 28]. However, a thorough analysis highlights that such automated transfers may not always be adequate to feed correct function predictions. Indeed, integrating the results of global alignment in order to predict function, gives more trusty predictions [29].

All of these factors have provided an increase in a varied number of automated approaches based on a number of features (i.e. Direct or module assisted approaches) [9]. The "FFPRed 3" [30] is an approach for protein function prediction based on the scanning of the input protein sequences accross an array of SVMs considering the tie between protein function and alternative motifs. GoFDR [31] is an alignment-based method for protein function prediction from the query sequence-based MSA produced by BLAST or PSI-BLAST search. After that, it induces the FDRs for a target GO term and builds up a PSSM for the FDRs. Finally, it scores the protein target using the PSSM, and tuning the raw score into probability.

DeepText2GO [32] applies a 'consensus' strategies to integrating text-based with sequence-based classifiers. Furthermore, the approach uses a myriad of protein information, such as sequence homology, families, citations, domains, and motifs. In the same trend, GOLabeler [33] integrates five component classifiers, trained from different features, such as GO term frequency, sequence alignment, amino acid trigram, domains and motifs, and biophysical properties, in the LTR machine learning framework. Meanwhile, the authors developped the NetGO web server [34], which incorporates protein-protein network information from STRING Database into the GOLabeler framework (i.e. LTR). NetGO can use a network information to annotate a protein at a large-scale level by using the massive network information of all species in STRING.

DeepGO [35] is an approach for predicting protein functions from protein sequences and PPI networks. It applies a neural networks to learn sequence and PPI network elements and hierarchically classifies it with GO classes. [36] is also an other approach to predict protein functions from a combination of "sequence similarity" (by using BLAST [12]), "domain architecture searches" (by using PFAM [37]) and "PPI networks data" (by using STRING [38]) into a consensus prediction for each of the three GO sub-ontologies, MF, BP and CC.

Additionally, there are some additional approaches, which have been applied to predict protein function based on the guilty by-association rules, e.g. the neighbor-counting method [39] and the Chi-square method [40]. Worthy of mention, a thorough review on methods in automated protein function prediction is provided in [41]. Although many computational approaches have been developed in recent years to predict protein function, for the CAFA^[1], most of these traditional algorithms do not use the PPI alignment to predict the function for unknown proteins.

In this paper, we introduce a new strategy to predict protein function by exploiting PPI networks, sequence similarity and functional similarity. The new strategy predicts the functional annotation of proteins through the comparison of multiple protein-protein interaction networks from different species. We provide a global network alignment with k-networks, by identifying modules of related proteins and then annotate each module based on the most frequent annotations. Our method aligns PPI from many species to discover functionally similar or conserved protein modules between them. Two major steps are involved:

- 1 Discover the modules or clusters which are functionally coherent using a largely revised version of MAPPIN method [42];
- 2 Discover the neighbors of a given unannotated protein that share a common functionality, especially by finding densely connected regions in a PPI network. Indeed, this neighborhood may lead to enhance the prediction results by adding the set of proteins interacting with the unannotated protein. The approach uses the *SCHLICKER* functionality similarity [43] to compare the Gene Ontology (GO) terms assigned to each neighbor of each unannotated protein (belonging to each discovered aligned nodes) in order to deduce the most frequent annotations. Adding to that, it is worth noting that this strategy leads to discover a functional module. The latter corresponds to an association of proteins involved in a common biological process or a molecular function and binding among themselves at different times and places.

It is worth mentioning that our approach is the first one that predicts functionality of unannotated proteins from transferring annotations via multiple GNA by integrating the GOA during the alignment process. Indeed, it gives us an advantage to predict useful and accurate function for unknown proteins. Our approach aims to find a multiple global network alignment, with k-networks, in order to find out clusters of proteins across the k-compared networks such that these clusters depict a conserved biological function. Here, we explore the possibility of using the GOA, i.e., functional similarity of protein between the compared networks to extract modules that correspond to specific biological processes by increasing the number of conserved interactions.

Results

Test Datasets

As a dataset for our prediction process, we used :

^[1]<https://www.biofunctionprediction.org/cafa/>

- The GO released in 2016_05;
- In addition to protein sequences similarity computed from the BLAST, we use PPI networks for multiple species from the STRING database [44] (cf. Table 1);
- The datasets from UniProtKB-GOA released in 2016_05 for the compared species from the CAFA3 challenge. We applied two datasets (i.e. D1 and D2) to align the different species (cf. Table 2), since it gives the best prediction quality, respectively in terms of protein-centric, term-centric and time ratio;
- The protein targets released on November 15th 2017 that had no function annotations at the time of training. It is available for download at <https://biofunctionprediction.org/cafa/>;
- The different alignment strategies (i.e pairwise or multiple alignment) for eukaryotes and prokaryotes species are described respectively in Table 2 and Additional file 1: Table S1.

Experimental Setup

TANA provides different options of protein function prediction in different modes including:

- TANA-1: the prediction process is only based on transferring shared annotation via multiple PPI networks alignment of D1 and D2 (cf., Table 2).
- TANA-2: the prediction process is based on the transfer of annotations by considering the shared GOA and the network neighborhood via multiple PPI network alignments of D1 and D2 (cf., Table 2).
- TANA-3: the prediction process is only based on transferring shared annotations via pairwise PPI networks alignment of D3 and D4, as shown respectively in Additional file 1: Table S1.
- TANA-4: the prediction process is based on the transfer of annotations by considering the shared GOA and the network neighbours via pairwise PPI networks alignments of D3 and D4, as shown respectively in Additional file 1: Table S1.

Evaluation Metrics

To evaluate the quality of protein function prediction for each algorithm, we consider two scenarios^[2]: (i) protein centric and (ii) term centric.

1) Protein-centric metrics:

The protein centric maximum F-measure is used in the CAFA3 challenge [45]. Here, we compute the F-measure, for a threshold $t \in [0, 1]$, using the average precision for proteins for which we predict at least one term and average recall for all proteins. Then, we select the maximum F-measure value of all thresholds. We compute the F_{max} measure using the following formulas:

$$pr_i(t) = \frac{\sum_f I(f \in P_i(t) \wedge f \in T_i)}{\sum_f I(f \in P_i(t))} \quad (1)$$

^[2]The python code to evaluate our approach and its competitors are available at <https://drive.google.com/open?id=1zx976F-IDrx-YiUn47yfV705VCMdQMds>

$$rc_i(t) = \frac{\sum_f I(f \in P_i(t) \wedge f \in T_i)}{\sum_f I(f \in T_i)} \quad (2)$$

$$AvgrPr(t) = \frac{1}{m(t)} \cdot \sum_{i=1}^{m(t)} pr_i(t) \quad (3)$$

$$AvgrRc(t) = \frac{1}{n} \cdot \sum_{i=1}^n rc_i(t) \quad (4)$$

$$F_{max} = \max_t \left\{ \frac{2 \cdot AvgrPr(t) \cdot AvgrRc(t)}{AvgrPr(t) + AvgrRc(t)} \right\} \quad (5)$$

In these measures, f stands for GO term, $P_i(t)$ denotes the set of terms for a protein i applying a threshold t , and T_i denotes the corresponding ground-truth set of terms for the protein i . Precision is averaged over the proteins with at least one predicted score greater than or equal to t and $m(t)$ is the number of such proteins. The parameter n stands for the number of targets used in such evaluation. Moreover, we have used the remaining uncertainty (ru), misinformation (mi) and the resulting minimum semantic distance (S_{min}) to evaluate the performance of our approaches. The latter metrics are defined as follows:

$$ru(t) = \frac{1}{n} \sum_{i=1}^n \sum_f ic(f) \cdot \mathbb{1}(f \notin P_i(t) \wedge f \in T_i), \quad (6)$$

$$mi(t) = \frac{1}{n} \sum_{i=1}^n \sum_f ic(f) \cdot \mathbb{1}(f \in P_i(t) \wedge f \notin T_i), \quad (7)$$

$$S_{min} = \min_t \left\{ \sqrt{ru(t)^2 + mi(t)^2} \right\}, \quad (8)$$

where $ic(f)$ stands for the information content of the ontology term.

2) Term-centric metrics:

We have also evaluated our results on the term-centric evaluation. To do so, for each GO term f , we compute the AUC for the ROC of a sensitivity (or a recall) for a given false positive rate (or 1-specificity). We compute the sensitivity and the specificity for a functional GO term f and threshold t as follows:

$$sn_f(t) = \frac{\sum_i I(f \in P_i(t) \wedge f \in T_i)}{\sum_i I(f \in T_i)} \quad (9)$$

$$sp_f(t) = \frac{\sum_i I(f \notin P_i(t) \wedge f \notin T_i)}{\sum_i I(f \notin T_i)} \quad (10)$$

where $P_i(t)$ denoting the set of predicted GO terms for a protein i by applying a threshold t , and T_i denotes the corresponding ground-truth set of terms for a protein i . Then, we define a term-centric F_{max} measure where for each GO term f we compute the F-measure using threshold t and all proteins in our test set. After that, we fix the maximum for all the thresholds.

$$pr_f(t) = \frac{\sum_i I(f \in P_i(t) \wedge f \in T_i)}{\sum_i I(f \in P_i(t))} \quad (11)$$

$$rc_f(t) = \frac{\sum_i I(f \in P_i(t) \wedge f \in T_i)}{\sum_i I(f \in T_i)} \quad (12)$$

$$F_{maxf} = \max_t \left\{ \frac{2 \cdot pr_f(t) \cdot rc_f(t)}{pr_f(t) + rc_f(t)} \right\} \quad (13)$$

Moreover, we compute the global ROC AUC for all predictions scores and Mathews Correlation Coefficient (MCC) for a threshold. Doing so, we compute a maximum protein-centric F-measure. Therefore, the ROC AUC metric is computed by applying the following formulas for a given threshold t :

$$AUC = \int_{-\infty}^n TPRT(\cdot)(-FPRT'(t))dt \quad (14)$$

$$TPR(t) = \frac{TP(t)}{TP(t) + FN(t)}, \quad FPR(t) = \frac{FP(t)}{FP(t) + TN(t)} \quad (15)$$

TP is the cardinality of true positive elements, FN is the cardinality of false negatives, FP is the cardinality of false positive elements and TN is the cardinality of true negatives elements. By doing so, the MCC metric is computed as follows:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (16)$$

Moreover, we plotted precision-recall curves for each approach for BP, CC and MF ontologies. The AUPR metric is applied in our evaluation in order to punish false positive prediction especially in highly imbalanced data [46].

Discussion

Application to protein function prediction

Table 1 glances an idea about the number and characteristics of species used by the CAFA3 challenge. The BP, CC and MF respective columns, in Table 2, show the number of proteins that have no experimental annotation in the current ontology category, but having experimental annotations in at least another ontology. We compared our approach versus

DEEPGO, GOFDR approaches as well as three baselines: DiamondBLAST, DiamondScore and Naive approach thoroughly described in [47]. Table 4 depicts the results obtained by the four versions of TANA^[3] (i.e. TANA-1, TANA-2, TANA-3 and TANA-4) and its competitors. The GOFDR prediction results for CAFA3 targets are retrieved from the web service available at <http://gofdr.tianlab.cn/>.

Protein-centric evaluation

The six approaches, as shown in Table 4, produce different results to unannotated proteins for the compared species. Indeed, the TANA-2 algorithm outperforms, in terms of F_{max} , all of the competing approaches, specifically for predicting MF, BP and CC terms. Whereas, DeepGo, DiamondBLAST and DiamondScore outperform TANA in BP and CC terms for F_{max} , AUC and MCC metrics. The influence of mixing the shared gene ontology annotations with network neighborhood is clearly noticeable by the alignment of multiple PPI performed by TANA-1 and TANA-2. Indeed, the accuracy of the prediction of TANA-1 on the MF category in terms of F_{max} evaluation was 0.63, whereas for the TANA-2 was 0.54 (cf. Table 4).

As respectively shown in Table 4 and in the Additional file 2: Table S2, predicting the BP terms is more critical process than predicting MF terms. Indeed, BP terms illustrate the relations between proteins, whereas those of MF terms illustrate the properties of a protein. Therefore, we can conclude that the feature or property of a given protein is determined by itself, whereas the relations of a protein with its neighborhood is not determined by itself, however also by other proteins. Thus, applying the sequence of a protein and other motifs (PPIs interaction) during the prediction process helps us to correctly identify the biological process for each target protein. The protein-centric performance assesses the accuracy of the approaches in assigning functional GO terms to an unannotated protein. Therefore, and as protein-centric evaluation shows (cf. Table 4), our approach leads to a good performance in predicting the BP terms to each unannotated proteins. Indeed our approach does not only rely on the homologs during the alignment process but it also takes the strength of the functional similarity from the GOA for each protein. Moreover, the out performance of TANA over its competitors for the MF prediction is owe to its ability to predict the functionality of unannotated proteins from different species even if the target proteins are not related by an homology sequence.

Nevertheless, the reason behind the fair performance of the TANA version in predicting CC terms category stands on the fact that the different versions of TANA cannot predict the "interlog" between the proteins for a given specie. Indeed, TANA doesn't leverage the network neighborhood during the process of the prediction. Therefore, the alignment of a PPI networks from a myriad of species leads to more "noisy" prediction coming from different types of intra-cellular location. Moreover, predicting these specific terms yields to a high number of false positives and thus hampers to get a good performance in terms of F_{max} metric.

Term-centric evaluation

We can come to the conclusion that, from Table 4, using the different versions of TANA for predicting MF, BP and CC terms (i.e., TANA-1, TANA-3) based only on the transfer of

^[3]Details about the prediction results in different modes of TANA are visible at: https://drive.google.com/open?id=1JmSBBbhW1bJbQ_H00nfzsTCjIUBIraAI

annotation by considering the shared GOA yields us a good performance in terms of term-centric metric than do the TANA-2 and TANA-4. Moreover, to have a good evaluation at the term-centric side, the TANA-1 would be the appropriate version since it discerns well and poorly predictable GO terms. The quality of the prediction for each individual GO terms was evaluated by computing the area under the AUC method. Computing the AUCs for an individual GO term associated to the MF gives us a different insight on the performance evaluation of the GO term belonging to the binding and those with transporter and catalytic activity. In general, the association of GO terms with displayed binding function decreases the AUC values.

A precision-recall curve is a plot of the precision (y-axis) and the recall (x-axis) for different thresholds, much like the ROC curve. To evaluate the accuracy of the prediction model, we plot the ROC curve for each approach. ROC curve represents the worst case with a strong effect of the hazard (AUC of 0.5 associated with a very weak discrimination capacity) and the best case with an AUC of 1 associated with a very strong discrimination capacity. In a ROC curve, true positive rates (y-axis) are plotted and false positive rates (x-axis). The AUC is the area covered by the curve surrounded by a given line generated by the competing methods and the axis. The bigger the area covered, the most suitable the approaches are at differentiating the given classes. As Figure 1 depicts, DEEPGO and Naive approach achieved the best score in BP, MF and CC ontologies. The scores are promising for the different versions of TANA. TANA-1 is roughly above the baseline DiamondBLAST for predicting MF, BP and CC terms.

Figure 2 depicts that TANA-1 outperformed the other approaches in the MF ontology by achieving the highest AUPR in the MF ontology, while both DEEPGO and the three baselines outperform TANA version in predicting BP terms. In fact, TANA-1 achieved 29% improvements over DEEPGO in terms of AUPR in the MF ontology. Moreover, TANA-1 achieved 18% improvements over the DiamondBLAST baseline in terms of AUPR in CC ontology. The TANA-1 version obtained better performance on both recall and precision, which indicates the influence of associating network neighborhood information with the shared annotation over BP, CC and MF ontologies. Therefore, the accurate term-centric evaluation of BP depend on data that is not sequence-based or network neighborhood-based, and need expansive strategies like using the Interpro-domain, 2D/3D structures and phylogenetic tree for each target protein.

As shown in the Additional file 2: Table S2 and the Additional file 3: Table S3, most of the F_{max} of GO terms in MF or BP obtained by TANA-1 are better than TANA-3. Whereas the performance of the two approaches are the same by predicting the GO terms in CC (cf. Additional file 4: Table S4). Moreover, among the BP terms, we uncover a null F_{max} value related to less specific terms such as "cellular process", "response to drug" and "response to insulin". Therefore, the annotation of the function biological macromolecules is a critical task caused by the dominance of poorly predictable GO terms. We also unveil that annotation of the proteins with GO terms in MF, e.g., "antigen binding", "polysaccharide binding", "drug binding" and "macromolecular complex binding" gave high performance. We can also conclude that the prediction of the MF and CC terms performed well than the BP category. In addition, we notice that the more specific function (GO terms from BP, MF or CC) perform well than the high-level functions. The reason behind this performance can be explained by the influence of the accurate transfer annotation on the quality of the prediction from the GNA and from the direct interaction neighborhood.

Our proposed approach is based on the idea to replace sequence similarities between proteins across different species with their functional similarities, in order to improve the prediction accuracy. Nevertheless, we didn't obtain an optimal prediction quality across the bacterial species, since their proteins are not well annotated (cf. Table 1). It is worth mentioning that some coverage (i.e. number of target proteins) of TANA for some species are low because our proposed approach is basically based on a homology transfer of function prediction. Therefore, this will influence negatively on the quality of the term centric evaluation applied on the NK and LK, and on the "partial-mode" which is based on averaging over the predicted subset. The proteins having no annotation for the BP, MF and CC ontologies belong to the NK category. Whereas, proteins with LK are those that had been experimentally annotated in one or three GO (BP, MF or CC ontology). Meanwhile, the goal is to try to increase the discovering of the number of UA, NMF, NBP, NCC without increasing the value of ME and MNE for each multiple or pairwise alignment (cf. Table 3).

Running time

All the experiments carried out on a 2.6 GHz Intel(R) Xeon(R) 48 CPUs, machine running GNU/Linux with 503GB system memory. For the sake of brevity, in this subsection, we only report the running time for the TANA-1 and TANA-2 approaches. So, the required time of the predictions tasks performed by TANA-1 for the datasets "D1" and "D2" was roughly 4 hours and 12 minutes, respectively. Whereas the time required of TANA-2 for the datasets "D1" and "D2" was roughly 6 hours and 12 minutes, respectively (cf. Table 3).

The reason behind this required time for the dataset "D1" is the given huge number of sequences for each compared species (about roughly 100.000 proteins used during the alignment process), released by the CAFA3 project. Moreover, there are many proteins that have a degree in their PPI network greater than 2000 interactions (e.g. Q5RKN4, Q00420) which should be integrated during the prediction process from the direct interaction neighborhood for a given unannotated protein.

Conclusions

In this paper, we introduced a new approach for protein function prediction by transferring annotation via PPI networks alignment and from the direct interaction neighbours of the target protein. The approach considers that annotation from BP, MF or CC ontologies shared by annotated protein, can be predicted from its interacting partners belonging at the same consistent cluster. The results of the alignment and the prediction of the functionalities of proteins from different species using both GOA and sequence homology are very encouraging and flexible in terms of computational runtime.

As a future work, we plan to integrate the metabolic pathway for each specie during the prediction process. We expect that doing so would give us insight on the different types of reactions involved by each compared protein. Moreover, in terms of quality scores, there is still significant improvement in all ontologies, and particularly in BP and CC terms using different strategies. We also plan to assess the ability of our approach to associate proteins with disease terms from disease gene prediction tasks using the HPO from CAFA. Finally, we will put the focus on increasing the coverage of the prediction for each species in order to evaluate our approach on the partial mode.

Methods and Algorithms

We start this section by providing a thorough description of the TANA algorithm.

Shared function associated with the cluster generated by GNA

In the following, we describe our approach for specifying function to an unannotated protein based on the frequency of each GO term within a GNA-aligned cluster of proteins. A functional cluster illustrates a subnetwork of proteins that shares a common function. The driving idea of our approach stands on the fact that the functionalities shown by more proteins member, within their cluster, is eligible to prediction.

The algorithm implemented in our approach operates in six major steps (cf., Figure 3):

- **Step 1:** The function prediction of the annotated protein can be inferred by finding enriched annotations within the cluster by taking into account that each protein from the cluster may be engaged in multiple roles and functions. We extract the GOA induced by a node and its member in the same cluster resulting from GNA.
- **Step 2:** Following, each annotated protein within its cluster is considered suitable for a possible annotation transfer. Among the annotated proteins in the cluster, we consider the proteins that contain at least two GOA overlaps. When modules or clusters are explored, every shared function associated with the module, is used for transferring annotations to the unannotated protein. For this reason, the applied clustering method is mandatory for enhancing the quality of the functional predictions. Interestingly enough, instead of predicting functions for individual proteins, our approach tries, at first, to discover consistent clusters of proteins and then assign functions to all the unannotated proteins in each cluster.
- **Step 3:** Afterward, TANA finds the shared GOA based on the majority functions of the neighbors in the PPI of an annotated protein and then assigns it these functions.
- **Step 4:** Next, each function shared by the majority of the clusters' proteins is assigned to all the proteins in the module or in the cluster by putting them into the set of GO terms.
- **Step 5 :** Then, if the annotation of proteins is not shared by the cluster, then we try to compute the semantic similarity in the context of GO [48] using the Relevance [49] metrics.
- **Step 6:** After that, we select the compared annotations having the highest confidence score. Therefore, if the given "Relevance metric" between both of compared GO terms is greater than the score 0.5, then we add them to the list of the predicted terms for each unannotated protein.

To avoid a circular argument in our methodology, we use proteins functional knowledge from the inter-network links (by linking proteins that are functionally similar) that are different from the prediction task:

i) Firstly, during the alignment process, we exclude the GOA with evidence code IEA. ii) Secondly, during the prediction task, we only take in consideration the GOA with evidence code EXP, IDA, IEP, IPI, IMP, IGI, TAC and IC.

TANA generates a global alignment from each discovered cluster *VertexCluster* belonging to the set $V^* = \bigcup \{VertexCluster(v) : \forall v \in V\}$. We denote by *VertexClusterTOPredict*(*up*) the set of all annotated proteins aligned to the unannotated protein *up* $\in V$. Here, *GOT*(*p_i*) stands for the set of GO terms annotating a protein *p_i* $\in VertexClusterToPredict$ (*up*), i.e.,

$$F_{global} = \bigcup \{GOT(p_i = \{t_1, t_2, \dots, t_k\} : \forall p_i \in VertexClusterToPredict(up) \quad (17)$$

For each function $t_i \in GOT(p_i)$ of a given annotated protein p_i , we assign a score based on the frequency of its occurrence in the F_{global} list, in order to emerge the collection of the shared functionalities (i.e., GO annotation term) shown by the entire annotated protein member in the cluster or module.

$$Score_{t_i}(up) = \frac{\sum_{p_i \in VertexClusterToPredict(up)} \delta(p_i, t_i)}{|VertexClusterToPredict(up)|} : \\ \forall t_i \in F_{global} \quad \text{and} \quad \forall p_i \in VertexClusterToPredict(up) \quad (18)$$

- 1: (p_i, fct) , if the annotated protein p_i has the function $fct \in GOT(p_i)$;
- 0: otherwise.

Where $|VertexClusterToPredict(up)|$ stands for the number of annotated protein for each predicted cluster.

Afterwards, if the score based on the frequency for each predicted GO term t_i is lower than the value 1.0 (i.e., $Score_{t_i}(up) < 1.0$), then we try to compute a given semantic simi-

```

Input: Global network alignment (GNA), Gene Ontology annotation (GOA) for each
species.
Output: A set of predicted functions for the unannotated protein
for all  $V^* \in GNA$  do
  for all  $up \in VertexCluster(V^*)$  do
    for all  $\{a_k\} : a_k \in N(up)$  do
      4:  $s_{funct}(up, a_k) \leftarrow s_{Schlicker}(up, a_k);$ 
        if  $s_{funct}(up, a_k) > 0.7$  then
           $VertexClusterToPredict(up) = VertexClusterToPredict(up) \cup \{a_k\}$ 
        end if
      8: end for
    end for
  for all  $up \in VertexCluster(V^*)$  do
     $F_{global} \leftarrow \emptyset;$ 
    12:  $St_i \leftarrow 0$ 
    for all  $p_i \in VertexClusterToPredict(up)$  do
       $F_{global} = \cup \{GOT(p_i) = \{t_1, t_2, \dots, t_k\}\};$ 
      for all  $t_i \in GOT(p_i)$  do
        16:  $St_i \leftarrow St_i + 1;$ 
      end for
    end for
    for all  $t_i \in F_{global}$  do
      20:  $Score_{t_i} \leftarrow \emptyset$ 
      if  $St_i \geq 1.0$  then
         $Score_{t_i} \leftarrow \frac{St_i}{|VertexClusterToPredict|}$ 
         $St_i \leftarrow 0$ 
      24: else
         $Score_{relevance} \leftarrow RelevanceSimilarity(t_i, t_{i+1})$ 
        if  $Score_{relevance} \geq 0.5$  then
           $Score_{t_i} \leftarrow Score_{relevance}$ 
        28: end if
      end if
    end for
  end for
32: end for

```

Algorithm 1: *FunctionPrediction*(GNA, GOA)

larity (i.e., Resnik, Lin, Jiang, Relevance metric) between the GO term t_i against the other terms $t_i \in F_{global}$. Therefore, if the given "Relevance metric" between both of compared GO terms is greater than the score 0.5, then we add them to the list of the predicted terms for each unannotated protein up . We set the value 0.5 as the threshold, since it yields us a confidence to ascertain the degree of similarity between the compared GO terms. We applied this method in order to stress on the importance of GO terms that gives a good score using the semantic similarities, even if the score of the two compared terms is low in terms of frequency (i.e., $Score_{t_i}(up)$) (cf., Algorithm 1).

We perform the prediction task by assuming that the GNA results is with a higher evaluation in terms of total coverage and consistency between clusters [42]. Therefore, to assess the biological relevance of the clusters, we consider the clusters having at least two annotated proteins (using the biological process or molecular function annotations) as well as the cluster that yields good scores on two key dimensions: coverage and consistency. Then, the evaluation is of paramount importance in order to ascertain the biological relevance to be used to transfer annotations, after the evaluation of the all PPI network alignment paradigms. Figure 4 depicts the three cases encountered by our approach to predict the function of the unannotated protein;

- 1 **Case 1:** When TANA encounters a cluster with 6 proteins, only two of them are with one known shared function (i.e., yellow ovals). Then, our approach blindly assigns the shared function to the four unannotated proteins (i.e., white ovals);
- 2 **Case 2:** When TANA encounters a cluster with 6 proteins, five of them are annotated with two known shared functions (i.e., yellow and blue ovals). Then, our approach blindly assigns both of the shared functions to the unannotated protein;
- 3 **Case 3:** When TANA encounters a cluster with seven proteins, six of them are annotated with two known shared functions (i.e., yellow and blue ovals) and two different functions (i.e., green and red ovals) but they are related semantically with the other shared functions by applying the different functional similarities (i.e., Resnik, Lin, Jiang or Relevance similarity). Then, our approach blindly assigns the two shared functions to the unannotated protein and semantically the two other functions (i.e., green and red ovals).

The shared function associated with network neighbors of the unannotated protein

The degree of a protein is the amount of interactions in the PPI network. If a given protein has a great degree then means that form two types of functional module: a) protein complexes (simultaneous interactions); b) metabolic pathways (consecutive interactions). The proteins do not accomplish their biological functions by themselves but achieve their functions by interrelating with other proteins or with copies of themselves in PPIs networks. The idea of shared network neighbors has been used: proteins that share many common neighbors also have close functional associations (cf., Figure 5). For each given protein, we computed how many times each of the shared functions occur as annotation in the direct interaction neighbourhood of the protein.

For a given unannotated protein $up \in V$, we designate the PPI neighborhood of up , $N(up)$, to be the subgraph of G whose vertex set consists of all of interaction neighbors passed over up . Formally, we set $N(up) = (V_{up}, E_{up}, W_{up})$, where $V_{up} = \{v \in V | (up, v) \in E\}$, $E_{up} = \{(v, v') \in E | (up, v) \in E \text{ and } (up, v') \in E\}$, and W_{up} are the weight on the edges in E_{up} . Then, we try to compute the functional similarity using the

SCHLICKER metric in order to select the compared protein having the highest confidence score (greater than 0.7), then we add the annotated protein from the direct network neighborhood to the list *VertexClusterToPredict(up)* of the cluster for each *up* unannotated protein. Finally, the added annotated protein from the neighbour of *up* will share its GO terms within the cluster or the module which *up* belong (cf. Algorithm 1). This technique will enhance the results of the CC GO terms, because this process is based on the discovering of the direct interaction neighbours inner the same specie of *up*.

Theoretical Complexity Study

The complexity of TANA depends on the number of k network, which should be aligned. Adding to that, TANA applies the strategy of the shared GO terms with the comparison of the network neighbours for each unknown proteins using the *SCHLICKER* functional similarity. So, for each k networks, where the maximum network size is $n = \max_i |V_i|$, the maximum number of interactions in a network is $m = \max_i |E_i|$. For each bipartite graph, $B_s = (V_{s1} \cup V_{s2}, E_s)$ the running time complexity on B_s is about $O(|V_{s1} \cup V_{s2}| \cdot \log |E_s|)$. Adding to that, the collection of candidate edge $\binom{k}{2} O(n \log(n))$ time. Finally, after the running of the Simulated Annealing [42], the function prediction task depends on the density of the network neighbour which is about $O(n)$.

Abbreviations

TANA: Transferring Annotation via Network Alignment; PPI: Protein-Protein Interaction; CAFA: Computational Assessment of Function Annotation; SVMs: Support Vector Machines; MSA: Multiple sequence alignment; FDRs: Functionally discriminating residues; PSSM: Position specific scoring matrix; LTR: Learning to rank; MF: Molecular function; BP: Biological process; CC: Cellular component; GNA: Global network Alignment; GOA: Gene ontology annotation; GO: Gene ontology; STRING: Search Tool for the Retrieval of Interacting Genes/Proteins; UniProtKB: UniProt Knowledgebase; ROC: Receiver Operating Characteristic; AUC: Area Under Curve; AUPR: Area under the precision-recall curve; NK: No-knowledge; LK: Limited-knowledge; HPO: Human Phenotype Ontology; IEA: Inferred from Electronic Annotation; EXP: Inferred from Experiment; IDA: Inferred from Direct Assay; IEP: Inferred from Expression Pattern; IPI: Inferred from Physical Interaction; IMP: Inferred from Mutant Phenotype; IGI: Inferred from Genetic Interaction; TAC: Traceable Author Statement; IC: Inferred by Curator.

Acknowledgments

Not applicable.

Authors' contributions

EMN and SBY initiated the study. WDJ developed the code, conducted the experiments and wrote the manuscript. All authors read and approved the final version.

Funding

Not applicable.

Availability of data and materials

The implementation of TANA and the datasets are freely available at https://drive.google.com/open?id=1oAkYAX2MHmMFXajI6l1niGZ_pZdQrhoa3

Additional Files

Additional file 1 Table S1 — The different pairwise alignment strategies on the Eukaryotes and Prokaryotes species used by TANA

Additional file 2 Table S2 — Evaluation of TANA for each GO term (Biological Process), on the different organisms of CAFA3 evaluation set (full evaluation)

Additional file 3 Table S3 — Evaluation of TANA for each GO term (Molecular Function), on the different organisms of CAFA3 evaluation set (full evaluation)

Additional file 4 Table S4 — Evaluation of TANA for each GO term (Cellular Component), on the different organisms of CAFA3 evaluation set (full evaluation)

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹University of Tunis El Manar, Faculty of Sciences of Tunis, LR11ES14, Capmus Universitaire, 2092 Tunis, Tunisia. ²Tallinn University of Technology, Department of Software Science, Akadeemia tee 15a, 12618, Tallinn, Estonia. ³University Clermont Auvergne, CNRS, LIMOS, F-63000 CLERMONT-FERRAND, FRANCE.

References

- Tran, L.: Hypergraph and protein function prediction with gene expression data. arXiv preprint arXiv:1212.0388 (2012)
- Kotaru, A.R., Joshi, R.C.: Classification of phylogenetic profiles for protein function prediction: An svm approach. In: International Conference on Contemporary Computing, pp. 510–520 (2009). Springer
- Morin, M.M.: Phylogenetic Networks: Simulation, Characterization, and Reconstruction, (2007)
- Sael, L., Chitale, M., Kihara, D.: Structure-and sequence-based function prediction for non-homologous proteins. *Journal of structural and functional genomics* **13**(2), 111–123 (2012)
- Wang, Z., Cao, R., Cheng, J.: Three-level prediction of protein function by combining profile-sequence search, profile-profile search, and domain co-occurrence networks. In: *BMC Bioinformatics*, vol. 14, p. 3 (2013). BioMed Central
- Laskowski, R.A., Watson, J.D., Thornton, J.M.: Protein function prediction using local 3d templates. *Journal of molecular biology* **351**(3), 614–626 (2005)
- Glazer, D.S., Radmer, R.J., Altman, R.B.: Improving structure-based function prediction using molecular dynamics. *Structure* **17**(7), 919–929 (2009)
- Zhao, X.-M., Chen, L., Aihara, K.: Protein function prediction with the shortest path in functional linkage graph and boosting. *International journal of bioinformatics research and applications* **4**(4), 375–384 (2008)
- Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. *Molecular systems biology* **3**(1), 88 (2007)
- Zhou, X., Kao, M.-C.J., Wong, W.H.: Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences* **99**(20), 12783–12788 (2002)
- Wu, J., Kasif, S., DeLisi, C.: Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* **19**(12), 1524–1530 (2003)
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. *Journal of molecular biology* **215**(3), 403–410 (1990)
- Benso, A., Di Carlo, S., ur Rehman, H., Politano, G., Savino, A., Suravajhala, P.: A combined approach for genome wide protein function annotation/prediction. *Proteome science* **11**(1), 1 (2013)
- Wang, K., Samudrala, R.: Fssa: a novel method for identifying functional signatures from structural alignments. *Bioinformatics* **21**(13), 2969–2977 (2005)
- Ferré, S., King, R.D.: Finding motifs in protein secondary structure for use in function prediction. *Journal of Computational Biology* **13**(3), 719–731 (2006)
- Erdin, S., Venner, E., Lisewski, A.M., Lichtarge, O.: Function prediction from networks of local evolutionary similarity in protein structure. In: *BMC Bioinformatics*, vol. 14, p. 6 (2013). BioMed Central
- Laskowski, R.A., Watson, J.D., Thornton, J.M.: Profunc: a server for predicting protein function from 3d structure. *Nucleic acids research* **33**(suppl_2), 89–93 (2005)
- Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F.: Prediction of protein function using protein–protein interaction data. *Journal of computational biology* **10**(6), 947–960 (2003)
- Letovsky, S., Kasif, S.: Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* **19**(suppl_1), 197–204 (2003)
- Jaeger, S., Gaudan, S., Leser, U., Rebholz-Schuhmann, D.: Integrating protein-protein interactions and text mining for protein function prediction. In: *BMC Bioinformatics*, vol. 9, p. 2 (2008). BioMed Central
- Ahmed, K.S., Saloma, N.H., Kadam, Y.M.: Improving the prediction of yeast protein function using weighted protein-protein interactions. *Theoretical Biology and Medical Modelling* **8**(1), 11 (2011)
- Xiong, W., Liu, H., Guan, J., Zhou, S.: Protein function prediction by collective classification with explicit and implicit edges in protein-protein interaction networks. *BMC bioinformatics* **14**(12), 4 (2013)
- Peng, W., Wang, J., Cai, J., Chen, L., Li, M., Wu, F.-X.: Improving protein function prediction using domain and protein complexes in ppi networks. *BMC systems biology* **8**(1), 35 (2014)
- Wu, Q., Ye, Y., Ng, M.K., Ho, S.-S., Shi, R.: Collective prediction of protein functions from protein-protein interaction networks. In: *BMC Bioinformatics*, vol. 15, p. 9 (2014). BioMed Central
- Yu, G., Zhu, H., Domeniconi, C., Guo, M.: Integrating multiple networks for protein function prediction. In: *BMC Systems Biology*, vol. 9, p. 3 (2015). BioMed Central
- Chua, H.N., Sung, W.-K., Wong, L.: An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics* **23**(24), 3364–3373 (2007)
- Kuchaiev, O., Pržulj, N.: Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics* **27**(10), 1390–1396 (2011)
- Singh, R., Xu, J., Berger, B.: Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences* (2008)
- Sharan, R., Ideker, T.: Modeling cellular machinery through biological network comparison. *Nature biotechnology* **24**(4), 427 (2006)
- Cozzetto, D., Minneci, F., Currant, H., Jones, D.T.: Ffpred 3: feature-based function prediction for all gene ontology domains. *Scientific reports* **6**, 31865 (2016)
- Gong, Q., Ning, W., Tian, W.: Gofdr: a sequence alignment based method for predicting protein functions. *Methods* **93**, 3–14 (2016)

32. You, R., Huang, X., Zhu, S.: Deeptext2go: Improving large-scale protein function prediction with deep semantic text representation. *Methods* **145**, 82–90 (2018)
33. You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H., Zhu, S.: Golabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* **34**(14), 2465–2473 (2018)
34. You, R., Yao, S., Xiong, Y., Huang, X., Sun, F., Mamitsuka, H., Zhu, S.: Netgo: improving large-scale protein function prediction with massive network information. *Nucleic acids research* **47**(W1), 379–387 (2019)
35. Kulmanov, M., Khan, M.A., Hoehndorf, R.: Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* **34**(4), 660–668 (2017)
36. Piovesan, D., Giollo, M., Leonardi, E., Ferrari, C., Tosatto, S.C.E.: Inga: protein function prediction combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Research* **43**(W1), 134–140 (2015). doi:10.1093/nar/gkv523
37. Finn, R.D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.*: Pfam: the protein families database. *Nucleic acids research* **42**(D1), 222–230 (2013)
38. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Von Mering, C., *et al.*: String v9. 1: protein–protein interaction networks, with increased coverage and integration. *Nucleic acids research* **41**(D1), 808–815 (2012)
39. Fellenberg, M., Albermann, K., Zollner, A., Mewes, H.-W., Hani, J., *et al.*: Integrative analysis of protein interaction data. In: *Ismb*, vol. 8, pp. 152–161 (2000)
40. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T.: Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast* **18**(6), 523–531 (2001)
41. Hawkins, T., Kihara, D.: Function prediction of uncharacterized proteins. *Journal of bioinformatics and computational biology* **5**(01), 1–30 (2007)
42. W. E., D., Ben-Yahia, S., Mephu-Nguifo, E.: A novel computational approach for global alignment for multiple biological networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **15**(6), 2060–2066 (2018)
43. Schlicker, A., *et al.*: A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics* **7**(1), 302 (2006). doi:10.1186/1471-2105-7-302
44. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., *et al.*: String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research* **43**(D1), 447–452 (2014)
45. Clark, W.T., Radivojac, P.: Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics* **29**(13), 53–61 (2013)
46. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240 (2006). ACM
47. Kulmanov, M., Hoehndorf, R.: Deepgoplus: Improved protein function prediction from sequence. *bioRxiv*, 615260 (2019)
48. Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E., Falcão, A.O., Couto, F.M.: Metrics for go based protein semantic similarity: a systematic evaluation. In: *BMC Bioinformatics*, vol. 9, p. 4 (2008). BioMed Central
49. Schlicker, A., Domingues, F.S., Rahnenführer, J., Lengauer, T.: A new measure for functional similarity of gene products based on gene ontology. *BMC bioinformatics* **7**(1), 302 (2006)

Figures

[width=120mm]rocBP.eps [width=120mm]rocMF.eps [width=120mm]rocCC.eps
(a) (b) (c)

Figure 1: The ROC curves of TANA and other predictors over BPO, MFO and CCO in full evaluation mode.

[width=120mm]auprBP.eps [width=120mm]auprMF.eps [width=120mm]auprCC.eps
(a) (b) (c)

Figure 2: The precision-recall plot of TANA and other predictors over BPO, MFO and CCO in full evaluation mode.

[width=7cm, height=7cm]PredictionProcess.eps

Figure 3: The different steps of TANA for scoring the prediction of GO terms for each unannotated protein.

[width=9cm, height=7cm]img22.eps

Figure 4: Functional annotation yield by TANA. This shows clusters of proteins composed from unannotated proteins as well as annotated ones

[width=15cm, height=12cm]img23.eps

Figure 5: Two different strategies of protein function prediction are yield by TANA. In the PPI network on the left, proteins A, E, and F are annotated with "green" function, proteins C and D are annotated with "blue" function. The node H is annotated with the "yellow" function and G and B are unannotated proteins (as denoted by white node color). With the "direct neighborhood" approach (the middle panel), node B gets annotated with "green" function, since the majority of its neighbors (E and F) have the "green" function. Similarly, G gets annotated with "blue" function, since all of its neighbors (C and D) have the "blue" function. Adding to that, by using the other strategy based on transferring annotation from the alignment of PPI network, the protein B is also annotated with two shared functions (i.e, orange and black colors) coming from the "Cluster 1". Moreover, protein G also gets annotated with one shared function (i.e., Black color) from the shared function of the "Cluster 2". It is worth mentioning that: a) the proteins with a shape different from the circle node are a proteins coming from different species; b) The proteins B and G are the proteins with limited Knowledge which had been experimentally annotated in one or two GO ontologies (BP, MF or CC).

Tables

Table 1: Characteristics of the PPI Networks from 18 Species.

Domain	Taxonomy	Name	Number of proteins	Number of interactions	MF	BP	CC	Largest degree
Eukaryota	227	Drosophila melanogaster	12792	1942502	2776	5815	3963	2873
Eukaryota	6239	Caenorhabditis elegans	17522	3640543	1740	2575	2197	3708
Eukaryota	9606	Homo sapiens	19080	5867186	12202	11036	12581	7613
Eukaryota	10090	MOUSE	16049	5598454	3765	4604	6886	8512
Eukaryota	10116	RAT	17139	6444811	2534	3485	3612	15261
Eukaryota	702	ARATH	19080	5787443	4947	6979	8380	6753
Eukaryota	44689	DICDI	9737	1387080	325	450	777	2597
Eukaryota	4932	YEAST	6492	919449	3933	4632	4035	2188
Eukaryota	7955	DANRE	22291	6945346	670	833	570	6072
Archaea	243232	Methanocaldococcus jannaschii	1751	178286	46	25	3	697
Archaea	273057	Sulfolobus solfataricus P2	2858	299521	22	10	0	1027
Bacteria	160488	Pseudomonas putida KT2440	5163	745882	19	13	0	1815
Bacteria	170187	Streptococcus pneumoniae serotype 4	2083	153530	0	0	0	835
Bacteria	223283	Pseudomonas syringae pv. tomato	5318	653393	15	15	17	1742
Bacteria	224308	Bacillus subtilis	4161	509967	209	145	32	2012
Bacteria	243273	Mycoplasma genitalium	474	30419	1	2	0	318
Bacteria	316407	Escherichia coli str. K-12	4140	386303	2264	2201	2079	1453
Bacteria	85962	Helicobacter pylori	1461	82068	89	16	13	624

Table 2: Statistics on the species used by CAFA3

Domain	Taxonomy	Name	Counts	BP (LM)	CC (LM)	MF (LM)	FULL	NO	D1	D2
Eukaryota	7227	Drosophila melanogaster	13907	-	-	-	-	-	×	
Bacteria	208963	Pseudomonas aeruginosa UCBPP PA14	5892	-	-	-	-	-		*
Eukaryota	237561	Candida albicans SC5314	12421	-	-	-	-	-	*	
Eukaryota	9606	Homo sapiens	20197	4282	1905	5021	4647	8266	×	
Eukaryota	10090	MOUSE	16806	2296	2994	5221	3005	6850	×	
Eukaryota	10116	RAT	7963	824	1112	1354	1920	3781	×	
Eukaryota	284812	SCHPO	5120	2939	67	3745	679	582	*	
Eukaryota	3702	ARATH	14754	2727	2136	5084	1899	6464	×	
Eukaryota	44689	DICDI	4131	198	315	587	218	3203	×	
Eukaryota	559292	YEAST	6721	430	983	1704	2426	1801	×	
Eukaryota	7955	DANRE	2967	42	709	672	49	2113	×	
Eukaryota	8355	Xenopus laevis	3402	72	230	201	84	2996	*	
Archaea	243232	Methanocaldococcus jannaschii	1787	22	45	5	3	1739		×
Archaea	273057	Sulfolobus solfataricus P2	469	4	14	0	0	455		×
Bacteria	160488	Pseudomonas putida KT2440	705	3	16	0	0	689		×
Bacteria	170187	Streptococcus pneumoniae serotype 4	501	5	4	1	0	496		×
Bacteria	223283	Pseudomonas syringae pv. tomato	678	0	1	1	0	677		×
Bacteria	224308	Bacillus subtilis	4185	51	166	83	8	3987		×
Bacteria	243273	Mycoplasma genitalium	483	0	2	1	0	481		×
Bacteria	321314	Salmonella choleraesuis	882	0	0	0	0	882		*
Bacteria	83333	Escherichia coli	4434	978	1105	1130	1056	1242		×
Bacteria	85962	Helicobacter pylori	593	7	17	12	0	573		×
Bacteria	99287	Salmonella typhimurium	1789	11	36	22	7	1733		*

Note:

FULL : Number of proteins that have experimental annotations in all three ontologies (BP, MF or CC ontology);

NO: Number of proteins that have no experimental annotations in any ontology.

* : The specie mentioned by '*' indicates that it do not have a PPI network in the STRINGS database

- : The number of annotations for the proteins belonging to a given specie is not supplied by the CAFA3 LM : refers to limited

Table 3: Comparative Consistency on the Different PPI Networks.

Datasets	CV(%)	ME	MNE	Time	UA	NMF	NBP	NCC
TANA-1								
D1	38.60	0.938	0.373	≈4d	15349	7788	9420	8658
D2	7.36	0.330	0.230	5min	373	202	156	176
TANA-2								
D1	38.50	0.929	0.370	≈6h	15251	7703	9479	8680
D2	7.38	0.337	0.237	3.1min	373	202	156	176

Note: the algorithms TANA-1 and TANA-2 were tested on the two datasets D1 and D2. Each row lists the coverage (CV), mean entropy (ME), mean-normalized entropy (MNE), the running time, unknown alignment (UA) and the number of proteins annotated with NMF, NBP and NCC in alignment graph respectively. "Sec", "min", "h" and "d" in the row of time represent seconds, minutes, hours and days.

Table 4: Evaluation of TANA, DeepGO, DiamondScores, DiamondBLAST, Naive and GoFDR methods on the CAFA3 evaluation set (Full evaluation)

Methods	BP					MF					CC				
	F_{max}	S_{min}	AUPR	AUC	MCC	F_{max}	S_{min}	AUPR	AUC	MCC	F_{max}	S_{min}	AUPR	AUC	MCC
DiamondBLAST	0.399	25.320	0.116	0.738	0.156	0.431	10.233	0.178	0.608	0.074	0.506	8.800	0.142	0.731	0.146
DiamondScore	0.427	22.860	0.267	0.828	0.314	0.509	9.031	0.340	0.774	0.239	0.557	8.198	0.335	0.837	0.260
Naive	0.357	25.028	0.254	0.966	0.245	0.290	10.733	0.130	0.963	0.132	0.562	8.465	0.456	0.979	0.321
TANA-1	0.362	33.294	0.219	0.776	0.228	0.630	9.271	0.491	0.656	0.120	0.477	7.486	0.320	0.710	0.000
TANA-2	0.354	34.049	0.160	0.767	0.270	0.549	12.331	0.371	0.723	0.317	0.445	9.321	0.277	0.694	0.280
TANA-3	0.339	35.526	0.170	0.743	0.190	0.533	10.923	0.356	0.690	0.277	0.487	9.376	0.317	0.675	0.189
TANA-4	0.333	36.339	0.151	0.752	0.244	0.493	12.618	0.321	0.731	0.098	0.470	9.358	0.265	0.702	0.272
GoFDR	0.195	37.451	0.048	0.674	0.000	0.190	10.129	0.024	0.803	0.471	0.465	11.567	0.225	0.730	0.372
DeepGO	0.375	33.487	0.312	0.887	0.338	0.303	9.153	0.198	0.889	0.361	0.561	9.706	0.548	0.955	0.531

Note: Best results are indicated in bold with respect to each column.

Figures

Figure 1

The ROC curves of TANA and other predictors over BPO, MFO and CCO in full evaluation mode.

Figure 2

The precision-recall plot of TANA and other predictors over BPO, MFO and CCO in full evaluation mode.

Figure 3

The different steps of TANA for scoring the prediction of GO terms for each unannotated protein.

Figure 4

Functional annotation yield by TANA. This shows clusters of proteins composed from unannotated proteins as well as annotated ones

Strategy 2: Transferring annotation from the GNA

Strategy 1: Transferring annotation from the "Direct neighborhood"

Figure 5

Two different strategies of protein function prediction are yield by TANA. In the PPI network on the left, proteins A, E, and F are annotated with "green " function, proteins C and D are annotated with "blue" function. The node H is annotated with the "yellow" function and G and B are unannotated proteins (as denoted by white node color). With the "direct neighborhood" approach (the middle panel), node B gets annotated with "green" function, since the majority of its neighbors (E and F) have the "green" function. Similarly, G gets annotated with "blue" function, since all of its neighbors (C and D) have the "blue" function. Adding to that, by using the other strategy based on transferring annotation from the alignment of PPI network, the protein B is also annotated with two shared functions (i.e, orange and black colors) coming from the "Cluster 1". Moreover, protein G also gets annotated with one shared function (i.e., Black color) from the shared function of the "Cluster 2". It is worth mentioning that: a) the proteins with a shape different from the circle node are a proteins coming from different species; b) The proteins B and G are

the proteins with limited Knowledge which had been experimentally annotated in one or two GO ontologies (BP, MF or CC).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [additionalfile1.1.pdf](#)
- [additionalfile1.2.pdf](#)
- [additionalfile1.3.pdf](#)
- [additionalfile1.4.pdf](#)