A novel approach for predicting protein functions by transferring annotation via alignment networks

Warith Eddine Djeddi, Sadok Ben Yahia, Engelbert Mephu Nguifo

A novel approach for predicting protein functions by transferring annotation via alignment networks

Warith Eddine Djeddi[1], Sadok Ben Yahia[1,2*] and Engelbert Mephu Nguifo[3*]

[1]University of Tunis El Manar, Faculty of Sciences of Tunis, LR11ES14, Capmus Universitaire 2092, Tunis, Tunisia

[2]Tallinn University of Technology, Department of Software Science, Akadeemia tee 15a, 12618 Tallinn, Estonia and

[3] University Clermont Auvergne, CNRS, LIMOS, F-63000 CLERMONT-FERRAND, FRANCE

*Corresponding author: sadok.ben@taltech.ee, engelbert.mephu_nguifo@uca.fr

Abstract

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 for many model species, it becomes an opportunity to computational methods to elucidate protein function based on many strategies. In this respect, most automated computational approaches integrate diverse kinds of functional interactions to deduce protein functions by transferring annotations across different species by relying on similar sequence, structure 2D/3D, amino acid motifs or phylogenetic profiles. In this work, we introduce a new approach called TANA (Transferring Annotation via Network Alignment) for inferring protein function which is based on our approach MAPPIN for GNA (Global Network Alignment). The main originality of the introduced approach stands on discovering functional modules within the PPI network by transferring annotation via network alignment. 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 standards established by the Computational Assessment of Function Annotation (CAFA) and highlight a sharp significant improvement over other competitive methods, in particular for predicting molecular functions.

1 Introduction

The past decade has witnessed a rising in genomic and proteomic data, leading to a large variety of sequenced genomes and proteomes. A fundamental challenge is the interpretation of this overwhelming of data 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, 2], phylogenetic profiles [3, 4], protein sequences [5, 6], protein structures [7, 8], and protein interactions. A wealthy number of computational 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].

Combining both Gene Ontology Annotation (GOA) and protein-protein interaction (PPI) data allows the discovery of function for unknown proteins based on three general categories of molecular function, biological process and cellular component specified in all three Gene Ontology (GO) types. Prediction of the protein function based on the annotation transfers via the network alignment of multiple networks poses many thriving issues such as:

- The computational complexity, i.e., the number of proposed correspondences increases potentially as far as the number of compared networks grows;
- The size of genomes related to the varied networks to be aligned may change widely, e.g., because of differing degrees of gene duplication;
- The genomes or proteomes available are noisy, biased and incomplete;
- The GO only carries out positive terms, i.e., there is no data on functions that proteins do not have which diffuse mis-annotations when homology-based approaches are applied. In addition, not all unknown proteins have homologous proteins in databases which could give putative fake functions to unknown proteins, e.g. the chromosomal proximity method [10], the Rosetta stone method [10, 11, 12], the phylogenetic method [13] and the combined method [14, 15, 16].

In this paper, we introduce a new strategy to predict 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 our method based on MAPPIN for aligning PPI networks [17]. Nevertheless, we also introduce with some variations in order to align the different species of the CAFA3 challenge;
2. Predict the function of unannotated proteins in a cluster using our novel strategy thoroughly described in remainder.

To evaluate the predictions for the unannotated proteins, we compare our results versus those of pioneering approaches dedicated to function prediction.

The amount of large scale PPI networks have emerged quickly. Simultaneously, collaborative attempts to annotate proteins and genes using GO annotations. Knowledge bases using GO annotations, such as the UniProt Knowledgebase (UniProtKB), provide a rich annotation data on PPI networks and afford relevant information for discerning the biological processes that preserve cellular structure and function. The alignment of PPI networks is a convenient strategy for comparing the networks of different species. This comparison helps identifying functional modules that are conserved across the PPI networks. This alignment is performed by first establishing a mapping between the nodes of the compared PPI networks relying on biological information, commonly sequence homology.

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

- **Gene expression pattern:** protein function prediction by analyzing gene expression pattern [18].
- **Phylogenetic profile:** analyzing phylogenetic profile, i.e., evolutionary history of proteins [13, 19].

- **Protein sequence:** protein function prediction using protein sequence, sequence similarity measures, homologies are primarily used. Applying program such as the Basic Local Alignment Search Tool (BLAST) [20], PSI-BLAST [21] and FASTA [22] to find possible homologs in sequence databases such as TrEMBL [23] and UniProt [24].

- **Protein structure:** the function prediction using protein structure, by using some approaches to analyze the secondary [25, 26] and tertiary structures [27, 28, 7] of proteins.

- **Protein-protein interactions (PPIs):** protein function prediction using protein-protein interactions [29, 30, 31, 32, 33, 34, 35, 36] can be deduced from the interaction of the neighborhood. Chua et al. [36] 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. Whereas, other methods analyze a single specie's protein network to distinguich functional modules (as reviewed in [37]. A typical single-species appproch applying connectivity strategy to cluster a protein network into highly connected modules, e.g., MCODE [38]. Moreover, PPI networks of single species have been used to extract biological pathways. The reader is referred to [9] for a survey on the topic.

- **Network alignment:** the function prediction using annotation transfers via network alignment by confering the annotations of a protein in an aligned cluster to the unannotated member of the identical cluster [39, 40]. However, a thorough analysis highlights that such automated transfers may not always be adequate to feed correct function predictions. Integrating the global alignment results into the function prediction strategies, using network analysis techniques, that gives more trusty predictions [41].

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 [42, 29] and the Chi-square method [43]. Worthy of the mention, a thorough review on methods in automated protein function prediction is provided in [44].

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]. We take an example of tools for protein function prediction as Predictprotein [45], DEEPGO [46], PFAM [47], SIFTER [48], INTERPRO [49]. The ffpred3 [50] is an approach for protein function prediction based on the scanning of the input protein sequences accross an array of Support Vector Machines (SVMs) considering the tie between protein function and alternative motifs. GOFDR [51] is an alignment-based method for protein function prediction from the query sequence-based multiple sequence alignment (MSA) produced by BLAST or PSI-BLAST search. After that, it induces the functionally discriminating residues (FDRs) for a target GO term and builds up a position specific scoring matrix (PSSM) for the FDRs. Finally, it scores the protein target using the PSSM, and tuning the raw score into probability.

DeepGO [46] is an approach for predicting protein functions from protein sequences and PPI networks. It applies a deep neural networks to learn sequence and PPI network elements and hierarchically classifies it with GO classes. [52] is also an other approach to predict protein functions from a combination of "Sequence similarity" (by using BLAST [20]), "domain architecture searches" (by using PFAM [47]) and PPI networks data (by using STRING [53]) into a consensus prediction for each of the three GO sub-ontologies (i.e., MF, BP and CC).

Although many computational approaches have been developed in recent years to predict protein function, most of these traditional algorithms do not take functional similarity during protein function prediction process except the PINALOG approach [54]. However, the latter is only used for pairwise alignment.

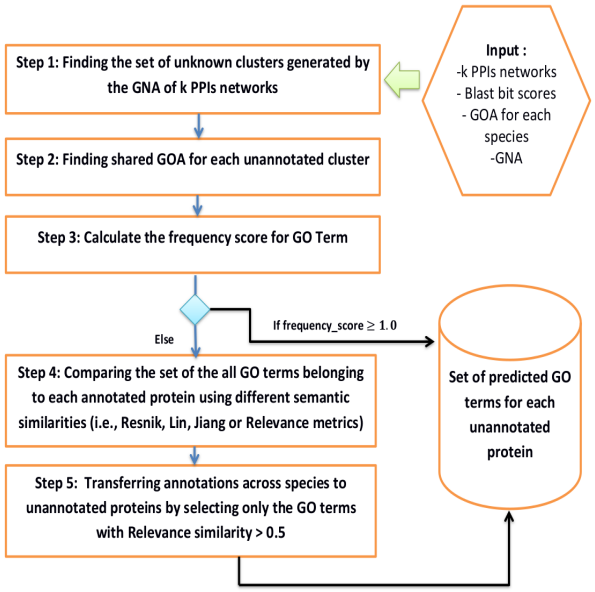

Figure 1: The different steps of TANA for scoring the prediction of GO Terms for each unannotated protein.

In the remainder, we introduce TANA, an approach that predicts protein function exploiting PPI networks, sequence similarity and functional similarity. TANA doesn't only rely on homology inference to assign function, since it is very difficult to infer homology for highly divergent proteins. We evaluated our approach according to their ability to predict terms in the Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) ontologies established by CAFA3.

It is worth mentioning that our approach is the first one that predicts functionality of unannotated proteins from transferring annotations via GNA with low computational complexity. Indeed, it gives us an advantage to predict a batch of 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.

The remainder of this paper is organized as follows. Section 2 depicts the architecture of the introduced algorithm and presents our method for protein function prediction. Section 3 describes our evaluation methodology and discusses experimental results. Finally, section 4 concludes with an outline of future work.

2 Methods and Algorithms

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

2.1 The TANA Algorithm

In the following, we describe our approach for specifying function to an unannotated protein based on its cluster's functional annotation's frequency i.e., annotation transfer across PPI networks. 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 has five major steps:

- **Step 1**: The function prediction of the annotated protein can be inferred by finding enriched annotations within the cluster by taking in account that each protein from the cluster may be engaged in multiple roles and functions. We extract GOA induced by a node and its member in the same cluster resulting from GNA.

- **Step 2**: Each annotated protein within its cluster is considered suitable for a possible annotation transfer. Among the top nodes in the list, 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 proteins in each cluster.

- **Step 3**: Therefore, 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 inti the set of GO terms.

Input: Global network alignment (GNA), Gene Ontology annotation for each species (GOA).
Output: A set of predicted function for unannotated protein
for all $V^* \in GNA$ **do**
  for all $up \in VertexCluster(V^*)$ **do**
    $F_{global} \longleftarrow \emptyset$;
    $St_i \longleftarrow 0$
    for all $p_i \in VertexClusterToPredict(up)$ **do**
      $F_{global} = \cup\{GOT(p_i) = \{t_1, t_2, ..., t_k\}\}$;
      for all $t_i \in GOT(p_i)$ **do**
        $St_i \longleftarrow St_i + 1$;
      end for
    end for
    for all $t_i \in F_{global}$ **do**
      $Score_{t_i} \longleftarrow \emptyset$
      if $St_i \geq 1.0$ **then**
        $Score_{t_i} \longleftarrow \frac{St_i}{|VertexClusterToPredict|}$
        $St_i \longleftarrow 0$
      else
        $Score_{relevance} \longleftarrow RelevanceSimilarity(t_i, t_{i+1})$
        if $Score_{relevance} \geq 0.5$ **then**
$Score_{t_i} \longleftarrow Score_{relevance}$
        end if
      end if
    end for
  end for
end for

Algorithm 1: $FunctionPrediction(GNA, GOA)$

- **Step 4** : Moreover, if the annotation of proteins is not shared by the cluster, then we try to compute the semantic similarity in the context of GO [55] using Resnik [56], Lin [57], Jiang [58] or the Relevance [59] metrics.

- **Step 5**: By doing so, we select the compared annotations having the highest confidence score (cf., Figure 1). 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.

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

$$F_{global} = \cup\{GOT(p_i = \{t_1, t_2, ..., t_k\} : \forall p_i \in VertexClusterToPredict(up) \quad (1)$$

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} set, in order to emerge the set of the shared functionalities (i.e., GO annotation term) shown by the entire annotated protein member in the cluster or module.

$$Score_{t_i}(up) = \frac{\Sigma_{p_i \in VertexClusterToPredict(up)} \delta(p_i, t_i)}{|VertexClusterToPredict(up)|} : \forall t_i \in F_{global} \quad and \quad \forall p_i \in VertexClusterToPredict(up) \quad (2)$$

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

Where $|VertexClusterToPredict(up)|$ is denoting 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 the semantic similarity (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, because 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).

Table 1 provides an example for the prediction process, by our approach, for the two unannotated proteins "O97121" and "A5JYW2". As Table 1 depicts, the shared function "GO:0030170 (*pyridoxal phosphate binding*)" from the biological process (mentioned with orange color) is assigned to the unannotated proteins "*O97121" "A5JYW2". Moreover, our approach assigns four probable functions (i.e., GO:0019343 (*cysteine biosynthetic process via cystathionine*), GO:0019346 (*transsulfuration*), GO:0071266 (*'de novo' L-methionine biosynthetic process*) and GO:0019450 (*L-cysteine catabolic process to pyruvate*)) from the biological process, and three functions (i.e., GO:0004121 (*cystathionine beta-lyase activity*), GO:0004123 (*cystathionine gamma-lyase activity*) and GO:0080146 (*L-cysteine desulfhydrase activity*)) from the molecular function since their respective scores are greater than 0.5 (the seven functions are

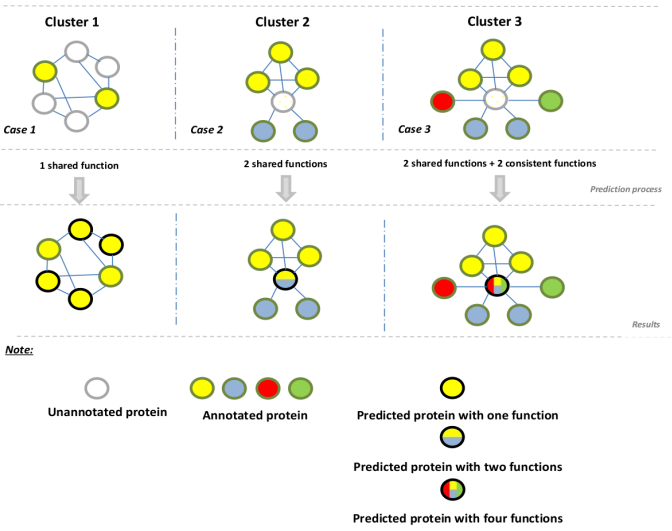

Figure 2: Functional annotation yield by TANA. This shows clusters of proteins composed from unannotated proteins as well as annotated ones

mentioned with blue color in Table 1). It is important to mention that, the GO terms with a value less than 0.5 are omitted. A sample of these omitted functions is indicated in Table 1 with a barred text highlighted with gray color.

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 [17]. Therefore, to assess the biologically relevance of the clusters, we consider the clusters having at least two annotated proteins (using the GO biological process or molecular function annotations) as well as the cluster that yields a good scores on two key dimensions: coverage and consistency. Then, the evaluation is of paramount importance in order to ascertain the biologically relevance to be used to transfer annotations, after the evaluation of the all PPI network alignment paradigms. Figure 2 depicts the three cases encountered by our approach to predict the function of the unannotated protein; **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). **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. **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).

3 Results and Discussion

3.1 Test Datasets

As a dataset for our prediction process, we tried to use :

- The Gene Ontology (GO) released in 2016_05;

Table 1: The prediction process by TANA for the two unannotated proteins 'O97121' and 'A5JYW2'

Step 1: Finding Cluster

Number of the selected cluster from the alignment is 430: Protein marked with *, is an Unannotated protein.
The Number of annotated proteins in the cluster is equal to two proteins

P55216 *A5JYW2 P06721 *O97121

Step 2: Finding shared Gene Ontology Annotation
430->P55216|GO:0003962|IBA(spec=1)(MF) (score=0.5)(frequency GOT=1)
430->P55216|GO:0019343|IBA(spec=1)(BP) (score= 0.5)(frequency GOT=1)
430->P55216|GO:0030170|IBA(spec=1)(MF) (score= 1)(frequency GOT=2)
430->P55216|GO:0071266|IBA(spec=1)(BP) (score= 0.5)(frequency GOT=1)
430->P55216|GO:0019346|IBA(spec=1)(BP) (score= 0.5)(frequency GOT=1)
430->P06721|GO:0004121|IDA(spec=3)(MF) (score= 0.5)(frequency GOT=1)
430->P55216|GO:0004123|IBA(spec=1)(MF) (score= 0.5)(frequency GOT=1)
430->P06721|GO:0019450|IBA(spec=3)(BP) (score= 0.5)(frequency GOT=1)
430->P06721|GO:0030170|IDA(spec=3)(MF) (score= 1)(frequency GOT=2)
430->P06721|GO:0080146|IMP(spec=3)(MF) (score= 0.5)(frequency GOT=1)
430-> *O97121(spec=2)
430-> *A5JYW2(spec=1)

Step 3: Comparing the set of the all GO terms belonging to each annotated protein using different semantic metrics
>Prediction of GO terms by transferring shared annotation to unannotated protein (marked with *) in cluster 430
MF:
~~GO:0080146 vs GO:0003962(Resnik =0.955594, Lin =0.115966, Jiang =0.0642288, Relevance =0.0713676)~~
GO:0080146 vs GO:0004123(Resnik =6.61275, Lin =0.785493, Jiang =0.21684, Relevance =0.784438)
~~GO:0080146 vs GO:0030170(Resnik = 0, Lin =0, Jiang =0, Relevance = 0)~~
GO:0080146 vs GO:0004121(Resnik =6.61275, Lin =0.770104, Jiang =0.202095, Relevance =0.769069)
~~GO:0003962 vs GO:0004123(Resnik =0.955594, Lin =0.118383, Jiang =0.0656476, Relevance =0.0728551)~~
~~GO:0003962 vs GO:0030170(Resnik = 0, Lin =0, Jiang =0, Relevance = 0)~~
~~GO:0003962 vs GO:0004121(Resnik =0.955594, Lin =0.115966, Jiang =0.0642288, Relevance =0.0713676)~~
~~GO:0004123 vs GO:0030170(Resnik = 0, Lin =0, Jiang =0, Relevance = 0)~~
GO:0004123 vs GO:0004121(Resnik =6.61275, Lin =0.785493, Jiang =0.21684, Relevance =0.784438)
~~GO:0030170 vs GO:0004121(Resnik = 0, Lin =0, Jiang =0, Relevance = 0)~~
BP:
GO:0019450 vs GO:0071266(Resnik =5.48696, Lin =0.619381, Jiang =0.129138, Relevance =0.616817)
GO:0019450 vs GO:0019346(Resnik =6.46474, Lin =0.748859, Jiang =0.187403, Relevance =0.747692)
GO:0019450 vs GO:0019343(Resnik =6.46474, Lin =0.729755, Jiang =0.172769, Relevance =0.728618)
GO:0071266 vs GO:0019346(Resnik =5.48696, Lin =0.685321, Jiang =0.165593, Relevance =0.682484)
GO:0071266 vs GO:0019343(Resnik =6.13526, Lin =0.745258, Jiang =0.192519, Relevance =0.743644))
GO:0019346 vs GO:0019343(Resnik =6.46474, Lin =0.807446, Jiang =0.244898, Relevance =0.806188)

Step 4: Transferring annotations across species to unannotated proteins
The process is performed by selecting only the GO terms with Relevance similarity > 0.5 or with frequency score >=1.0

*O97121	*A5JYW2
GO:0019343 [BP] 0.728618	GO:0019343 [BP] 0.728618
GO:0019346 [BP] 0.747692	GO:0019346 [BP] 0.747692
GO:0071266 [BP] 0.616817	GO:0071266 [BP] 0.616817
GO:0019450 [BP] 0.616817	GO:0019450 [BP] 0.616817
GO:0004121 [MF] 0.769069	GO:0004121 [MF] 0.769069
GO:0004123 [MF] 0.784438	GO:0004123 [MF] 0.784438
GO:0080146 [MF] 0.784438	GO:0080146 [MF] 0.784438
GO:0030170 [MF] 0.80	GO:0030170 [MF] 0.80

Note:
(spec=1): Arabidopsis PPIs network has 26337 proteins and 8311584 interactions.
(spec=3): Drosophila melanogaster PPIs network has 13471 proteins and 3901815 interactions.

- In addition to protein sequences similarity computed from the Blast, we use protein-protein interaction (PPI) networks for multiple species from the STRING database [60];
- The datasets from UniProtKB-GOA released in 2016_05 for the compared species from the CAFA3 challenge. Moreover, we select the proteins with annotations with experimental evidence code (EXP, IDA, IPI, IMP, IGI, IEP, TAS and IC);
- The protein targets released on 05 June 2017 that had no function annotations at the time of training. The dataset contains 1367 proteins and 3619 annotations. It is available for download at https://biofunctionprediction.org/cafa/.

3.2 Experimental Setup

We applied a TANA version which excludes sequence similarity with low similarity, since they lead to an uncoherent prediction. Furthermore, we replace the low sequence similarity for each compared protein with the functional similarity between them, in the case where the value of the functional similarity is high.

Therefore, to get the prediction from the alignment of the target species, the approach computes the semantic similarity between two GO terms using the functional similarity proposed by Schlicker et al. [59]. Moreover, to avoid the unreliability of mis-annotation in the Uniprot database, we exclude the GOA with evidence code IEA (inferred from electronic annotation) and GO annotations derived from Cellular Component.

We have set to 0.3 the value of the Alpha parameter, since it gives the best biological alignment quality in terms of CV, ME, MNE and time ratio [17].

3.3 Evaluation Metrics

To evaluate the quality of protein function prediction, we apply the protein centric maximum F-measure which are used in the CAFA3 challenge [61]. 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:

$$pr_i(t) = \frac{\sum_f I(f \in P_i(t) \wedge f \in T_i)}{\sum_f I(f \in P_i(t))} \tag{3}$$

$$rc_i(t) = \frac{\sum_f I(f \in P_i(t) \wedge f \in T_i)}{\sum_f I(f \in T_i(t))} \tag{4}$$

$$AvgrPr(t) = \frac{1}{m(t)} . \sum_{i=1}^{m(t)} pr_i(t) \tag{5}$$

$$AvgrRc(t) = \frac{1}{n} . \sum_{i=1}^{n} rc_i(t) \tag{6}$$

$$F_{max} = \max_t \{ \frac{2 . AvgrPr(t) . AvgrRc(t)}{AvgrPr(t) + AvgrRc(t)} \} \tag{7}$$

In these measures, f is GO term, $P_i(t)$ denotes the set of terms for a protein i applying a threshold t, and T_i

Table 2: Statistics on the species used by CAFA3

Domain	Taxonomy	Name	Counts	BP (LIMITED)	CC (LIMITED)	MF (LIMITED)	FULL	NO
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	Bacilus 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 annotation in all three ontologies (BP, MF or CC ontology);
NO: Number of proteins that have no experimental annotation in any ontology.

Table 3: Performance of TANA on the Human specie (Partial evaluation)

Organisms	BP				MF				CC			
	BC	F_{max}	S_{min}	$n-smin$	BC	F_{max}	S_{min}	$n-smin$	BC	F_{max}	S_{min}	$n-smin$
No-knowledge (NK)												
Human	68	0.40	21.12	0.57	73	0.53	5.52	0.46	53	0.46	6.50	0.56
Limited-Knowledge (LK)												
Human	163	0.30	31.07	0.58	93	0.53	6.64	0.49	68	0.24	4.6	0.69

Note:
BC: refers to the benchmark count used to test the accuracy of the prediction for each ontology category;
n-smin: refers to minimum normalized semantic distance.

Table 4: Evaluation of TANA, DeepGO, FFPred3 and GoFDR methods on a CAFA3 preliminary evaluation set (Full evaluation)

Methods	BP			MF			CC		
	F_{max}	$AvgPr$	$AvgRc$	F_{max}	$AvgPr$	$AvgRc$	F_{max}	$AvgPr$	$AvgRc$
TANA	**0.42**	**0.45**	**0.40**	**0.54**	0.60	**0.49**	0.26	0.39	0.19
FFPRed3	0.26	0.30	0.23	0.38	0.35	0.40	0.44	0.46	0.43
GoFDR	0.20	0.27	0.15	0.52	**0.89**	0.36	0.40	0.40	0.41
DeepGO	0.34	0.31	0.37	0.47	0.61	0.39	**0.52**	**0.55**	**0.49**

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

denotes the corresponding ground-truth set of terms for a 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 applied the remaining uncertainty (ru), misinformation (mi) and the resulting minimum semantic distance (S_{min}) to evaluate the performance of our approach. The latter metrics are defined as follows

$$ru(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{f} ic(f).\mathbb{1}(f \notin P_i(t) \wedge f \in T_i), \tag{8}$$

$$mi(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{f} ic(f).\mathbb{1}(f \in P_i(t) \wedge f \notin T_i), \tag{9}$$

$$S_{min} = \min_{t}\left\{\sqrt{ru(t)^2 + mi(t)^2}\right\}, \tag{10}$$

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

3.4 Application to protein function prediction

CAFA3 provided two types of benchmarks, no-knowledge (NK) and limited-knowledge (LK), and two modes of evaluation, full-mode (FM) by averaging over the entire benchmark sets and partial-mode (PM) by 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 two GO ontologies (BP, MF or CC). Table 2 gives an idea about the number and characteristics of species used by the CAFA3 challenge. The BP, CC and MF column in Table 2 indicate the number of proteins that have no experimental annotation in the current ontology category, but have experimental annotations in at least one other ontology.

A glance to Table 2 shows that the LK evaluation provided by TANA yielded improvement in terms of performance accuracy. Therefore, our algorithm have exploited the correlations between experimental annotations across the three ontologies in order to enhance the quality of the function prediction. The prediction of the function applied to human proteome sequences are encouraging (cf. Table 3), we confirm that TANA's alignments can be used to predict biological characteristics, i.e., GO molecular function (MF) and biological process (BP), of unannotated proteins based on their alignments with annotated ones. The human proteome sequences include 18,380 human protein sequences with 5,746, 5,850 and 9,684 human proteins annotated with experimental evidence code (EXP, IDA, IPI, IMP, IGI, IEP, TAS and IC) in BP, MF and CC categories, respectively. For the NK type, the quality for predictions of GO terms in the MF, BP, and CC category in terms of F_{max} evaluation was 0.40, 0.53, and 0.46, respectively (cf. Table 3). Whereas for the LK type, the quality for predictions of GO terms in the MF, BP, and CC category in terms of F_{max} evaluation was 0.30, 0.53, and 0.24, respectively (cf. Table 3).

Table 4 depicts the results using different metrics to evaluate the prediction quality yielded by TANA[1], DeepGO, FFPred3 [50] and GoFDR [51] algorithms, on a dataset released as part of the CAFA3 challenge. The FFPred3 [50] prediction results for CAFA3 targets are retreived from http://bioinfadmin.cs.ucl.ac.uk/downloads/ffpred/cafa3/, the GoFDR results by the web service available at http://gofdr.tianlab.cn/.

The four approaches produce different results to unannotated proteins for the compared species. Indeed, the TANA algorithm outperforms, in terms of F_{max}, its competitors, specially for predicting molecular function and biological process for the F_{max} evaluation. Whereas, DeepGo, FFPred3 and GoFDR outperform TANA in CC GO terms for F_{max}, $AvgPr$ and $AvgRc$ metrics.

As respectively shown, in Tables 3 and 4, predicting the BP GO terms is a critical process than predicting MF GO terms. Indeed, BP GO terms illustrate the relations between proteins, whereas those of MF GO 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 neighborhoud is not determined by itself, however also by other proteins. Thus, applying the sequence of a protein and other motifs (alignment of metabolic pathway) during the prediction process helps us to correctly identify the biological process for each target protein. The protein-centric performance measures the accuracy of the approaches in assigning functional GO terms to an unannotated protein. The reason behind the low performance of our method in predicting CC GO Terms category is that our approach cannot predict the "interlog" between the proteins for a given specie. Therefore the alignment of a PPI from a myriad of species lead to more noise prediction coming from different types of intra cellular location. Moreover, predicting these specific terms yields to a great number of false positives and thus hampers to get a good performance in terms of F_{max} metric. Another solution to get a good evaluation, is to try to assign annotation with more general CC GO terms like using the annotation "organelle" (GO:0043226, level 2), "intracellular part" (GO:0044424, level 3), and "cytoplasm" (GO:0005737, level 4). TANA flags out a good performance in BP, MF GO terms, by relying on the transfer of annotation by only considering the experimental annotation derived from the Uniprot-GOA. Moreover, the reason, behind the superiority of TANA over its competitors consists in predicting the functionality of unannotated proteins from different

[1]Details about the prediction results of TANA are visible at: https://github.com/waritheddine/TANA/blob/master/TANA-Prediction-CAFA.txt

species even if the target proteins are not related by homology sequence (i.e., Difficult target).

TANA relies on the MAPPIN algorithm to generate clusters from the alignment of PPI. Thus, our approach for the protein function prediction takes roughly 8 hours to generate the alignment and performing the function prediction process. The reason behind this required time 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.

3.5 Validation on the non-IEA annotation of proteins

Indeed, there are many predictions performed by TANA, that have been added to the current release as a non-IEA annotation. Therfore, we assessed the prediction accuracy of TANA by validating on the non-IEA annotation of proteins included in the current release from UniProtKB released in 2018_07. We tried to use the anterior datasets from UniProtKB released in 2016_11 for the compared species. After that, we tried to validate the prediction of unannotated protein against the last release datasets from UniProtKB released in 2018_07. It is worth mentioning that many unannotated proteins from the anterior became in the meanwhile annotated proteins in the last release. The predictions performed by our approach are more accurate and roughly are the same when we compared them against the annotations of the last release. To illustrate that, let us consider one of the unannotated proteins, to wit "Q9VRX7", which is with no function from UniProtKB released in 2016_11. So, after using the anterior release by TANA during the alignment process in order to predict its function, our approach predicts three GO terms in MF (i.e., GO:0000175 (*3'-5'-exoribonuclease activity*), GO:0004535 (*poly(A)-specific ribonuclease activity*) and GO:0005515 (*protein binding*)) and one GO terms in BP (i.e., GO:0000289 (*nuclear-transcribed mRNA poly(A) tail shortening*)). The unannotated protein "Q9VRX7" became an annotated protein, and the curator of the database Uniprot-GOA (release 2018_07) assigned one GO terms in MF with evidence code "IDA" (i.e., GO:0000175 (*3'-5'-exoribonuclease activity*)) and two GO terms in BP with evidence code "IMP" (i.e., GO:0031125 (*rRNA 3'-end processing*) and GO:0031126 (*snoRNA 3'-end processing*)). Indeed, there are many predictions performed by TANA, that have been added to the current release as a non-IEA annotation.

4 Conclusion and Future Works

In this paper, we introduced a new approach for protein function prediction by transferring annotation via PPI networks alignment. 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 promising and flexible in terms of computational runtime.

As a future work, we plan to integrate the metabolic pathway for each species during the prediction process which gives us insight on the different type of reaction 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 GO 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 Human Phenotype Ontology (HPO) [8] from CAFA.

References

[1] Xing-Ming Zhao, Luonan Chen, and Kazuyuki Aihara. 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.

[2] Loc Tran. Hypergraph and protein function prediction with gene expression data. *arXiv preprint arXiv:1212.0388*, 2012.

[3] Appala Raju Kotaru and Ramesh C Joshi. Classification of phylogenetic profiles for protein function prediction: An svm approach. In *International Conference on Contemporary Computing*, pages 510–520. Springer, 2009.

[4] Monique Marlene Morin. *Phylogenetic networks: simulation, characterization, and reconstruction*. University of New Mexico, 2007.

[5] Lee Sael, Meghana Chitale, and Daisuke Kihara. Structure-and sequence-based function prediction for non-homologous proteins. *Journal of structural and functional genomics*, 13(2):111–123, 2012.

[6] Zheng Wang, Renzhi Cao, and Jianlin Cheng. Three-level prediction of protein function by combining profile-sequence search, profile-profile search, and domain co-occurrence networks. In *BMC bioinformatics*, volume 14, page S3. BioMed Central, 2013.

[7] Roman A Laskowski, James D Watson, and Janet M Thornton. Protein function prediction using local 3d templates. *Journal of molecular biology*, 351(3):614–626, 2005.

[8] Dariya S Glazer, Randall J Radmer, and Russ B Altman. Improving structure-based function prediction using molecular dynamics. *Structure*, 17(7):919–929, 2009.

[9] Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Molecular systems biology*, 3(1):88, 2007.

[10] Ross Overbeek, Michael Fonstein, Mark DâĂŹsouza, Gordon D Pusch, and Natalia Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, 96(6):2896–2901, 1999.

[11] Edward M Marcotte, Matteo Pellegrini, Ho-Leung Ng, Danny W Rice, Todd O Yeates, and David Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

[12] Anton J Enright, Ioannis Iliopoulos, Nikos C Kyrpides, and Christos A Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86, 1999.

[13] Matteo Pellegrini, Edward M Marcotte, Michael J Thompson, David Eisenberg, and Todd O Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

[14] Edward M Marcotte, Matteo Pellegrini, Michael J Thompson, Todd O Yeates, and David Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83, 1999.

[15] Yu Zheng, Richard J Roberts, and Simon Kasif. Genomic functional annotation using co-evolution profiles of gene clusters. *Genome biology*, 3(11):research0060–1, 2002.

[16] Paul Pavlidis, Jason Weston, Jinsong Cai, and William Noble Grundy. Gene functional classification from heterogeneous data. In *Proceedings of the fifth annual international conference on Computational biology*, pages 249–255. ACM, 2001.

[17] W. E. Djeddi, S. B. Yahia, and E. M. Nguifo. A novel computational approach for global alignment for multiple biological networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(6):2060–2066, 2018.

[18] Xianghong Zhou, Ming-Chih J Kao, and Wing Hung Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences*, 99(20):12783–12788, 2002.

[19] Jie Wu, Simon Kasif, and Charles DeLisi. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, 19(12):1524–1530, 2003.

[20] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[21] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[22] William R Pearson and David J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

[23] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'donovan, Isabelle Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365–370, 2003.

[24] Cathy H Wu, Rolf Apweiler, Amos Bairoch, Darren A Natale, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, et al. The universal protein resource (uniprot): an expanding universe of protein information. *Nucleic acids research*, 34(suppl_1):D187–D191, 2006.

[25] Kai Wang and Ram Samudrala. Fssa: a novel method for identifying functional signatures from structural alignments. *Bioinformatics*, 21(13):2969–2977, 2005.

[26] Sébastien Ferré and Ross D King. Finding motifs in protein secondary structure for use in function prediction. *Journal of Computational Biology*, 13(3):719–731, 2006.

[27] Florencio Pazos and Michael JE Sternberg. Automated prediction of protein function and detection of functional sites from structure. *Proceedings of the National Academy of Sciences*, 101(41):14754–14759, 2004.

[28] Roman A Laskowski, James D Watson, and Janet M Thornton. Profunc: a server for predicting protein function from 3d structure. *Nucleic acids research*, 33(suppl_2):W89–W93, 2005.

[29] Benno Schwikowski, Peter Uetz, and Stanley Fields. A network of protein–protein interactions in yeast. *Nature biotechnology*, 18(12):1257, 2000.

[30] Christine Brun, François Chevenet, David Martin, Jérôme Wojcik, Alain Guénoche, and Bernard Jacq. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome biology*, 5(1):R6, 2003.

[31] Minghua Deng, Kui Zhang, Shipra Mehta, Ting Chen, and Fengzhu Sun. Prediction of protein function using protein–protein interaction data. *Journal of computational biology*, 10(6):947–960, 2003.

[32] Stanley Letovsky and Simon Kasif. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(suppl_1):i197–i204, 2003.

[33] Manoj Pratim Samanta and Shoudan Liang. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proceedings of the National Academy of Sciences*, 100(22):12579–12583, 2003.

[34] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, 21(6):697, 2003.

[35] Hon Nian Chua, Wing-Kin Sung, and Limsoon Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.

[36] Hon Nian Chua, Wing-Kin Sung, and Limsoon Wong. An efficient strategy for extensive integration of diverse biological data for protein function prediction. *Bioinformatics*, 23(24):3364–3373, 2007.

[37] Peer Bork, Lars J Jensen, Christian von Mering, Arun K Ramani, Insuk Lee, and Edward M Marcotte. Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, 14(3):292 – 299, 2004.

[38] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):2, 2003.

[39] Oleksii Kuchaiev and Nataša Pržulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396, 2011.

[40] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 2008.

[41] Roded Sharan and Trey Ideker. Modeling cellular machinery through biological network comparison. *Nature biotechnology*, 24(4):427, 2006.

[42] Matthias Fellenberg, Kaj Albermann, Alfred Zollner, Hans-Werner Mewes, Jean Hani, et al. Integrative analysis of protein interaction data. In *Ismb*, volume 8, pages 152–161, 2000.

[43] Haretsugu Hishigaki, Kenta Nakai, Toshihide Ono, Akira Tanigami, and Toshihisa Takagi. Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, 18(6):523–531, 2001.

[44] Troy Hawkins and Daisuke Kihara. Function prediction of uncharacterized proteins. *Journal of bioinformatics and computational biology*, 5(01):1–30, 2007.

[45] Guy Yachdav, Edda Kloppmann, Laszlo Kajan, Maximilian Hecht, Tatyana Goldberg, Tobias Hamp, Peter HÃűnigschmid, Andrea Schafferhans, Manfred Roos, Michael Bernhofer, Lothar Richter, Haim Ashkenazy, Marco Punta, Avner Schlessinger, Yana Bromberg, Reinhard Schneider, Gerrit Vriend, Chris Sander, Nir Ben-Tal, and Burkhard Rost. PredictproteinâĂŤan open resource for online prediction of protein structural and functional features. *Nucleic Acids Research*, 42(W1):W337–W343, 2014.

[46] Maxat Kulmanov, Mohammed Asif Khan, and Robert Hoehndorf. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668, 2017.

[47] Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, et al. Pfam: the protein families database. *Nucleic acids research*, 42(D1):D222–D230, 2013.

[48] Sayed M Sahraeian, Kevin R Luo, and Steven E Brenner. Sifter search: a web server for accurate phylogeny-based protein function prediction. *Nucleic acids research*, 43(W1):W141–W147, 2015.

[49] Alex Mitchell, Hsin-Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, Conor McMenamin, Gift Nuka, Sebastien Pesseat, et al. The interpro protein families database: the classification resource after 15 years. *Nucleic acids research*, 43(D1):D213–D221, 2014.

[50] Domenico Cozzetto, Federico Minneci, Hannah Currant, and David T Jones. Ffpred 3: feature-based function prediction for all gene ontology domains. *Scientific reports*, 6:31865, 2016.

[51] Qingtian Gong, Wei Ning, and Weidong Tian. Gofdr: a sequence alignment based method for predicting protein functions. *Methods*, 93:3–14, 2016.

[52] Damiano Piovesan, Manuel Giollo, Emanuela Leonardi, Carlo Ferrari, and Silvio C.E. Tosatto. Inga: protein function prediction combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Research*, 43(W1):W134–W140, 2015.

[53] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian Von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815, 2012.

[54] Hang T. T. Phan and Michael J. E. Sternberg. Pinalog: a novel approach to align protein interaction networksâĂŤimplications for complex detection and function prediction. *Bioinformatics*, 28(9):1239–1245, 2012.

[55] Catia Pesquita, Daniel Faria, Hugo Bastos, António EN Ferreira, André O Falcão, and Francisco M Couto. Metrics for go based protein semantic similarity: a systematic evaluation. In *BMC bioinformatics*, volume 9, page S4. BioMed Central, 2008.

[56] Philip Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the XI International Joint Conferences on Artificial*, pages 448–453, 1995.

[57] Dekang Lin et al. An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304, 1998.

[58] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008, 1997.

[59] Andreas Schlicker, Francisco S Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC bioinformatics*, 7(1):302, 2006.

[60] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2014.

[61] Wyatt T Clark and Predrag Radivojac. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, 29(13):i53–i61, 2013.