Amino Acids (2015) 47:2583–2592
DOI 10.1007/s00726-015-2049-3

ORIGINAL ARTICLE

Protein function prediction using guilty by association from interaction networks

Damiano Piovesan[1] · Manuel Giollo[1,2] · Carlo Ferrari[2] · Silvio C. E. Tosatto[1,3]

Received: 15 May 2015 / Accepted: 10 July 2015 / Published online: 28 July 2015

Abstract Protein function prediction from sequence using the Gene Ontology (GO) classification is useful in many biological problems. It has recently attracted increasing interest, thanks in part to the Critical Assessment of Function Annotation (CAFA) challenge. In this paper, we introduce Guilty by Association on STRING (GAS), a tool to predict protein function exploiting protein–protein interaction networks without sequence similarity. The assumption is that whenever a protein interacts with other proteins, it is part of the same biological process and located in the same cellular compartment. GAS retrieves interaction partners of a query protein from the STRING database and measures enrichment of the associated functional annotations to generate a sorted list of putative functions. A performance evaluation based on CAFA metrics and a fair comparison with optimized BLAST similarity searches is provided. The consensus of GAS and BLAST is shown to improve overall performance. The PPI approach is shown to outperform similarity searches for biological process and cellular compartment GO predictions. Moreover, an analysis of the best practices to exploit protein–protein interaction networks is also provided.

Keywords Protein function · Protein interaction network · Gene ontology · CAFA · Protein sequence

✉ Silvio C. E. Tosatto
silvio.tosatto@unipd.it

[1] Department of Biomedical Sciences, University of Padua, Viale G. Colombo 3, 35131 Padua, Italy

[2] Department of Information Engineering, University of Padua, Via Gradenigo 6, 35121 Padua, Italy

[3] CNR Institute of Neuroscience, Viale G. Colombo 3, 35131 Padua, Italy

Introduction

The large amount of available protein sequences requires usage of in silico methods for automatic large-scale function prediction. The annotation process, assigning functions to target proteins, is generally based on the transfer by homology principle (Pellegrini et al. 1999). Protein space can be partitioned in subsets (families) that groups proteins with a common ancestor and, possibly, the same function. Whenever evolutionary relationships between two different proteins are available, all features from one protein are transferred to the other. Sequence comparison is used to infer homology and collect evidence about membership in a given family. However, it requires to properly choose similarity measures and related cutoff values in order to avoid false positives (and, conversely, false negatives). As each family has its own story and is the result of different and complex evolutionary phenomena, available data are usually not sufficient to trace an unambiguous phylogenetic tree (Engelhardt et al. 2011). Any time two sequences appear to greatly diverge, it becomes impossible to find annotated homologs. On the other hand, the same protein can perform different functions when placed in a different organism, and sequence information alone cannot distinguish such situations. Within the Critical Assessment of protein Function Annotation (CAFA) experiment (Radivojac et al. 2013), it has been stated that the currently best methods to predict protein function rely on sequence similarity searches for conserved regions or homologous proteins (Piovesan et al. 2011; Cozzetto et al. 2013). Moreover, it has been recommended to extend standard homology search with new methods that use different sources of information on protein function (Clark and Radivojac 2011; Minneci et al. 2013; Piovesan et al. 2015). The CAFA experiment also provided standard criteria for the evaluation of the predictions, e.g., the dataset used for the blind test and the definition of function space through Gene Ontology (GO) terms (Ashburner et al. 2000). The scoring metrics for comparing function predictions in CAFA are mainly based on precision-recall curves.

New effective experimental techniques to find genome-wide interactions make protein–protein interaction data widely available and ready to be used for functional annotation (Ho et al. 2002; Zhu and Snyder 2003; Johnson et al. 2007). Approaches exploiting interaction networks have been widely used for annotation of the Yeast genome (Hishigaki et al. 2001; Brun et al. 2003; Deng et al. 2003; Nabieva et al. 2005; Chua et al. 2006). At the same time, many tools which analyze biological network properties are already available. Some of them use interaction networks to prioritize genes that are part of disease pathways. These applications use enriched functional terms to describe clusters of interacting proteins or genes. The STRING interaction database (Franceschini et al. 2013) itself provides tools to compute GO term enrichment in selected sub-networks. To the best of our knowledge, functional enrichment in protein–protein interaction networks has never been used effectively as a tool for predicting function of unknown proteins.

For example, the χ^2 test has been used to rank the functional terms associated to a group of interacting partners by comparing the frequency of the terms within the group and with the expected distribution in the whole network (Hishigaki et al. 2001). Another work, PRODISTIN (Brun et al. 2003), focuses on the clusterization of the entire Yeast interaction graph by means of a distance measure to define groups associated with the same functional class. A Bayesian approach (Deng et al. 2003) has been applied to calculate the posterior probability that a given protein has the function of interest. This method takes into account the prior probability of the entire network but it does not consider the dependencies among terms. Another method, FunctionalFlow (Nabieva et al. 2005), treats annotated nodes as "sources" and propagates the associated annotation through the connecting edges following some simple rules. These rules take into account the distance between two nodes and the number of alternative paths connecting them to produce a score.

All of these methods are based on a single model organism and cannot easily be compared with other state-of-the-art methods like those participating in CAFA. Moreover, they used a very small ad hoc ontology for Yeast which is two orders of magnitude smaller than the full GO. It is also difficult to evaluate their impact on the coverage of genome annotation, as the number of interactions available today is not comparable with networks available a few years ago.

In this paper, we introduce GAS (Guilty by Association on STRING) to predict protein function exploiting protein–protein interaction networks without sequence similarity measures. GAS is part of the algorithmic core of the INGA server (Piovesan et al. 2015), which performed well as group "Tosatto-UniPD" at the most recent CAFA experiment (2014; URL: http://biofunctionprediction.org/). Here, we provide a hitherto unpublished analysis on the GAS implementation details and parameters necessary to maximize accuracy as well as important considerations about best practices to exploit protein–protein interaction networks.

Methods

GAS

Protein–protein interaction (PPI) networks provide relevant information about protein function. The aim of GAS is to exploit the annotation of the neighborhood of a protein to transfer the function. The choice of the network, the definition of the set of interacting partners, the strategy to transfer annotation, and the method to build the consensus represent key factors to improve accuracy and implement an effective prediction tool. The idea at the basis of GAS arises from the analogy with the "Guilty-by-association" principle. This concept asserts that qualities of one object are inherently qualities of another, merely by an independent association. In our case it means that if a protein physically interacts (association) with other proteins it should share a similar function (quality). Proteins in a living cell have many physical interactors, each group of interacting proteins is expected to participate in the same biological process and to operate in the same sub-cellular compartment. This hypothesis is supported by the evidence that proteins in the same pathway are more interconnected (Barabási et al. 2011). Given a protein with unknown function, GAS uses the STRING (Franceschini et al. 2013) network to collect the set N of directly interacting nodes. All experimental GO terms are then associated to the annotated proteins retrieved from SwissProt (Dimmer et al. 2011) and ranked by a measure representing their specificity in the collected set. We estimated this specificity, by measuring enrichment with respect to the entire training set (i.e., the remaining STRING nodes). The P-value associated to the enrichment is computed according to Fisher's exact test, which represents the probability that a specific term, $\mathrm{GO_i}$, is associated to a given set by chance (null hypothesis). For each collected term $\mathrm{GO_i}$, the contingency table is shown in Table 1. The P-value is generated with the following standard formula:

$$P\text{-value}\ (\mathrm{GO_i}) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \qquad (1)$$

Table 1 Contingency table

	Node	Set
	Cluster	DB
Categories		
GOi	a	b
Not GOi	c	d

Cluster the set *N* of directly interacting nodes, *DB* the rest of nodes in STRING associated to experimental GO terms

GAS was evaluated on two different protein interaction networks. In the first case, we focused on highly confident STRING interactions (edge score ≥900). In the second one, we selected all STRING interactions with edge score ≥500. All nodes in the STRING network were mapped to the UniRef90 database to extend the number of interacting nodes and increase the chance of collecting experimental GO terms as well as to make GAS comparable with our version of BLAST (see "GAS-C").

GAS considers only direct interactors even in the case of poorly connected proteins. This is a sensible choice as we found that performance decreases when including second level interactors (data not shown). The main reason is that if a hub of the network is present among the direct interactors, it is expanded including lot of unrelated proteins. Most dangerous cases are protein hubs interacting widely with many other proteins without functional specificity, like chaperones and ubiquitin. A solution could be to exclude hub proteins, however it is very difficult to define a cutoff based on the number of interactors. The degree distribution of experimentally annotated proteins in STRING does not follow a power law decay as for scale-free networks, instead resulting in a broader bell shaped function with many proteins having up to 500 interactions (data not shown).

GAS-C

GAS-C, where "C" indicates Consensus, is an extended version of the algorithm that merges GAS with BLAST predictions. For each input protein, GAS-C first computes GAS and BLAST predictions independently and then combines them. BLAST hits are retrieved running the program with default parameters discarding hits with e-value higher than 1×10^{-3} sorted by Bit-score (default output), since it maximizes performances (Radivojac et al. 2013). The presence of large groups of homologous proteins with high sequence similarity in the sequence database may affect a BLAST prediction. The UniRef90 (Suzek et al. 2007) database was used to address the redundancy issue. For each hit corresponding to the representative sequence of a UniRef90 cluster, all experimental GO terms associated to all cluster members are transferred. This strategy increases sensitivity, allowing to retrieve hits with lower sequence identity but possibly richer annotations. To make GAS comparable with BLAST, we mapped UniRef90 clusters to the interacting nodes and transferred functional annotation from all members belonging to these clusters.

Since the *F*-score was found to be poorly correlated with the native output score (Bit-score and *P*-value), the *F*-score computed on the rank position was considered instead. For BLAST, the rank corresponds to the hit position in the output list, e.g., at rank 1 we find GO terms (plus ancestors) transferred from the first hit, the one with best Bit-score. For GAS, the rank is given by the *P*-value, e.g., at rank 1, we find terms (plus ancestors) with the lowest *P*-value. The values f_{tr}, converted to the *F*-measure of the *r*th (or higher) ranked terms for target *t*. From these data, the expected rank-dependent performance was evaluated through an exponential curve emphasizing the correlation between the ranking *r* of the predicted term and the *F*-measure:

$$E[F(r)] = e^{a+b\times r} + C \qquad (2)$$

where *a*, *b*, and *c* were estimated through a nonlinear least square on the predictions and corresponding *F*-measures. Next, the corresponding rank-dependent score S_I and S_B were assigned to each GO term predicted by GAS and BLAST, respectively. Whenever the same GO_i was predicted by both approaches, its score was updated as follows:

$$S_{\text{combined}}(\text{GO}_\text{i}) = 1 - (1 - S_\text{B}(\text{GO}_\text{i}) \times (1 - S_\text{I}(\text{GO}_\text{i}) \qquad (3)$$

Finally, all scores are propagated to the root of the ontology guaranteeing that each ancestor node always inherits the maximum probability from its children.

Training and test sets

The evaluation set for the prediction models is made of protein sequences with experimental annotation from SwissProt. It includes previously unannotated proteins that accumulated GO terms annotation in 1 year, accounting for 8976 proteins from 283 organisms. 4432, 3931, and 3194 sequences were counted for the MF, BP, and CC subontologies, respectively. It was obtained as the difference between the SwissProt releases v2012_07 and v2013_07, applying a filtering criterion for automatically predicted terms. Experimental ("trusted") annotation was considered as those terms which are associated to the evidence codes EXP, IDA, IMP, IGI, IEP, TAS, and IC. Figure 1 shows that the organism distribution of new annotated sequences differs strongly. 1940 new sequences (22 % of the entire test set) come from "other" organisms. The training set was obtained by randomly sampling 10,000 targets from the experimentally annotated sequences in SwissProt

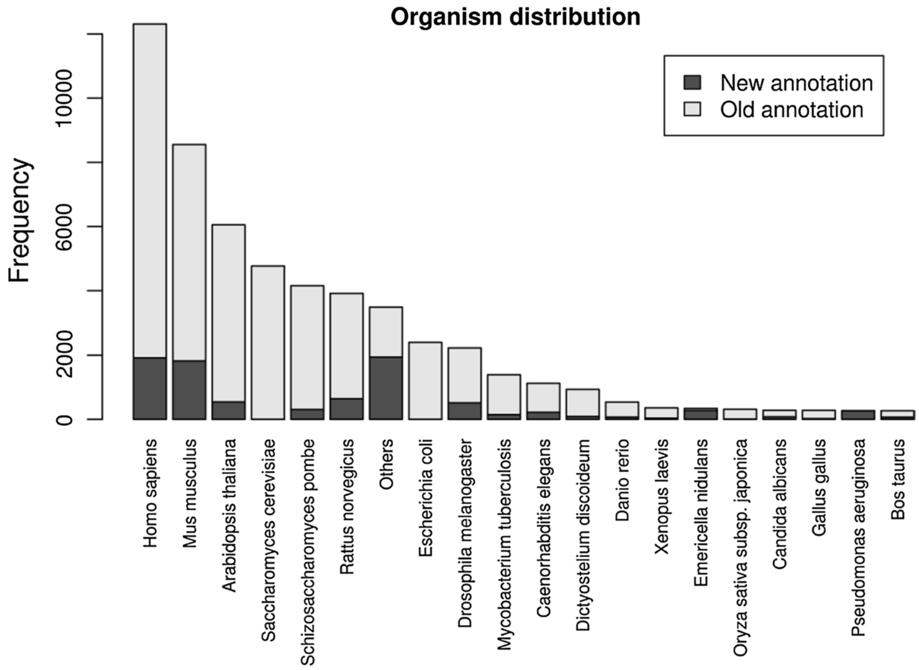

Fig. 1 Distribution of SwissProt entries annotated experimental GO terms and categorized by organism. *Dark bars* ("new annotation") represent the number of sequences that accumulated experimental annotation in 1 year and that were used as test set

v2012_07. The datasets, all predictions and GAS annotations for the Yeast genome are available from URL: http://protein.bio.unipd.it/inga/gas_dataset.tar.gz.

Performance evaluation

Two different strategies to evaluate GAS and GAS-C models, one based on a target-by-target comparison and the other based on the whole dataset were adopted. In the former approach, for each target protein, predicted GO terms by their ranking position were evaluated (as described in the GAS-C paragraph) and then the mean on the entire test set for all possible ranking r computed. In the whole dataset strategy, all targets were considered together and performance was calculated for all possible score thresholds t. The scores in this case correspond to the P-value, Bit-score, and S_{combined}, respectively, for GAS, BLAST, and GAS-C predictors. For FANN-GO (Clark and Radivojac 2011), we used the score as it is provided by the tool and for the Naïve method the frequency in the SwissProt database. We used the following well-established measures adopted in CAFA to evaluate performance:

$$\text{Precision } (r) = \frac{\left|\text{GO}_{\text{t}} \cap \text{GO}_{\text{p}}(r)\right|}{\left|\text{GO}_{\text{p}}(r)\right|} \tag{4}$$

$$\text{Recall } (r) = \frac{\left|\text{GO}_{\text{t}} \cap \text{GO}_{\text{p}}(r)\right|}{\left|\text{GO}_{\text{t}}\right|} \tag{5}$$

where GO_{t} represents the set of true terms associated to a protein in the test set, while GO_{p} is the set of predicted terms. Precision and recall are measures of correctness and completeness for a method, respectively. They both depend on r, which corresponds to the ranking in the target-by-target approach and to the score threshold in the whole dataset strategy. A third useful metric is the F-measure, which is obtained by calculating the harmonic mean of precision and recall:

$$F(r) = 2 \times \frac{\text{Precision } (r) \times \text{Recall } (r)}{\text{Precision } (r) + \text{Recall } (r)} \tag{6}$$

Results

We introduce Guilty by Association on STRING (GAS), a tool to predict protein function exploiting protein–protein interaction networks without sequence similarity measures. The assumption is that whenever a protein interacts with other proteins, it is part of the same biological process and located in the same cellular compartment. Two proteins exhibiting the same interaction partners can reasonably be inferred to have the same function. Given the sequence of an unknown target protein, GAS is able to retrieve its interacting partners from the STRING network and measures the enrichment of the associated functional annotations to generate a sorted list of putative functions. In the following, we will present some experiments that explain how protein interaction networks can contribute to solve the problem of protein function prediction. We will start with an analysis of the STRING network and then we will provide a comparison with some methods. The list of evaluated tools includes BLAST (Altschul 1990), which is known as the standard baseline tool for function prediction

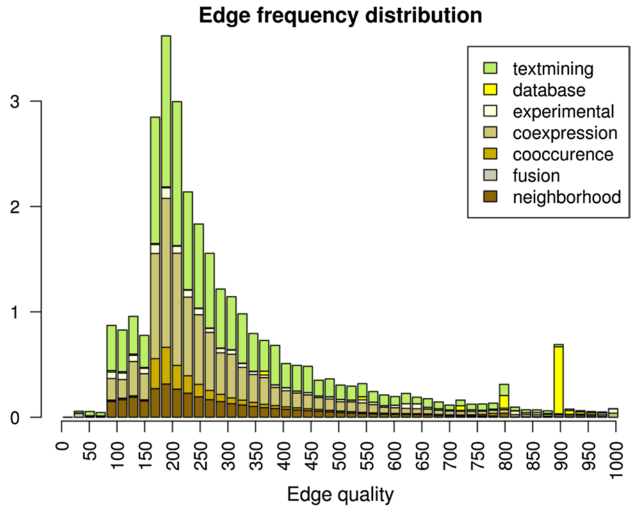

Fig. 2 Distribution of STRING edge types by edge quality. Frequencies are calculated using entries with experimental GO terms in SwissProt. Text mining and co-expression edges are the most common among the low qualities, while interactions from database are the most reliable

Table 2 Edge types distribution of experimentally annotated proteins in STRING

Edge type	Edges (%)
Text mining	53.3
Co-expression	40.1
Neighborhood	14.3
Co-occurrence	7.4
Experimental	5.7
Database	4.3
Fusion	0.3

The percentage refers to the total number of edges. Note that one interaction between two proteins may be confirmed by multiple source of information increasing the overall confidence for that edge

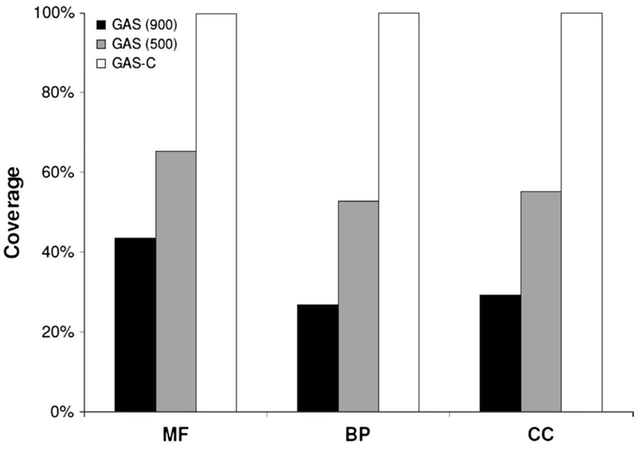

Fig. 3 Prediction coverage on the dataset for GAS at STRING edge weight cutoffs 900 and 500, as well as GAS-C

based on homology inference, the Naïve method implemented as described in CAFA and FANN-GO that was the only one available as stand-alone software, and trained on an old dataset. Finally, we show that the GAS-C consensus model can increase performance accuracy using both GAS and BLAST predictions.

Experimental annotation in STRING

We implemented GAS using STRING as the reference interaction network. STRING is the largest database of protein–protein interactions including experimental derived data, third party information coming from other databases, and predicted interactions (Franceschini et al. 2013). However, GAS does not use the entire network but only a portion composed by only those nodes that can be mapped to SwissProt entries and annotated with experimental terms. Exploitation of functional information coming from protein–protein interaction networks requires minimization of false-positive interactions. STRING provides a score representing an edge quality estimate that also tracks the information source. Figure 2 shows the distribution of STRING edge scores for different interaction types coming from different sources.

Most of the STRING edges connecting SwissProt entries have low quality values and come from text mining and co-expression data, 53 and 40 % of the total interactions, respectively, while only 5.7 % are confirmed experimentally (Table 2). When multiple sources of information support the existence of an interaction, they result in a higher global score. We evaluated GAS performances by filtering the STRING network for different edge confidence values. An edge cutoff of 900 on one hand not only guarantees the selection of reliable protein interactions, often confirmed in third party databases, but also reduces the amount of available interacting partners and therefore the annotation that can be transferred. On the contrary, a relaxed threshold yields a higher chance of collecting experimental GO terms useful for the prediction.

One of the major limitations of function prediction from interactome data is coverage (see Fig. 3). In fact, when no restriction in terms of alignment coverage and identity is applied, BLAST is capable of generating new GO terms in almost the totality of targets. For GAS, we are able to find experimentally annotated interacting nodes for our target protein in 29–47 % of the cases, depending on the ontology. We tested GAS performance by filtering edges for different cutoff values. All tables and figures in the paper refer to the GAS predictions coming from a high confidence STRING sub-network with a cutoff of 900. In Fig. 3, we reported the same comparison relaxing the edge filtering at

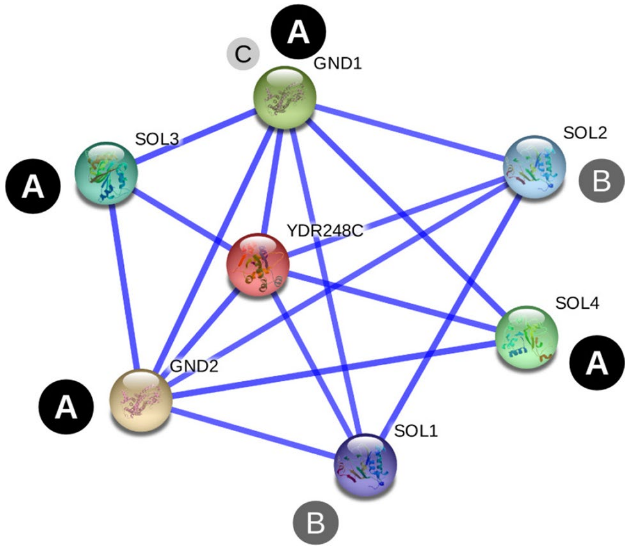

Fig. 4 The STRING sub-network for YDR248C (edge cutoff 900). *Circles* next to nodes represent experimental GO terms. Their size is proportional to the enrichment measure (*P*-value) provided by GAS. See main text for details

Table 3 Yeast genome annotation

Ontology	Method	Depth		Coverage (%)
		Average	SD	
MF	GAS	2.95	2.17	13.8
	GAS-C	**3.15**	1.83	**37.6**
	BLAST	3.11	1.80	32.1
BP	GAS	3.48	1.74	15.4
	GAS-C	**3.53**	1.67	**40.0**
	BLAST	3.37	1.69	34.4
CC	GAS	2.09	1.24	15.1
	GAS-C	**2.10**	1.21	**42.2**
	BLAST	2.03	1.21	37.5

Values computed for Yeast proteins of unknown function, i.e., missing experimental SwissProt annotation. The maximum *F*-score is used to measure average term depth, standard deviation and dataset coverage. The best performance for each ontology is highlighted in bold

a cutoff of 500. The ability to predict new potential functions increases greatly, ranging between 57 and 68 % coverage. Moreover, filtering out low confidence edges significantly decreases false-positive interactions resulting in a slightly greater accuracy (data not shown).

GAS prediction

To clarify the GAS the prediction procedure we provide an example. Figure 4 shows the GAS prediction for a Yeast "Probable gluconokinase" (UniProt: Q03786, gene name: YDR248C) missing experimental annotation. According to STRING, it interacts with 6 experimentally annotated proteins involved in the glucose metabolism. Four of them (marked as "A" in Fig. 4) are annotated with "pentose-phosphate shunt, oxidative branch" (GO:0009051), two ("B") with "tRNA export from nucleus" (GO:0006409), and only one ("C") with "cellular response to oxidative stress" (GO:0034599). GO terms in the figure are represented by circles. Their size is inversely proportional to the *P*-value (8.3×10^{-13}, 4.2×10^{-6} and 6.5×10^{-2}, respectively) and reflects the ranking in the output. GAS-C prioritizes different terms by combining the prediction of GAS and BLAST. In this case, our optimized BLAST predicts the "D-gluconate metabolic process" (GO:0019521) term, transferred from a "probable gluconokinase" (UniProt accession Q10242, sequence identity 43.2 %, Bit-score 122). BLAST and GAS predict different terms belonging to the same ontology branch. GAS-C prioritizes the common ancestors "single-organism carbohydrate metabolic process" (GO:0044723), "monocarboxylic acid metabolic process" (GO:0032787), and "carbohydrate catabolic process" (GO:0016052). The original leaf terms are placed in lower positions since they are not supported by both methods and less reliable.

In Table 3, we provide an overview of the GAS predictions for Yeast (NCBI tax. id. 559292) proteins missing experimental annotation in SwissProt (1748 sequences, 26 % of the entire genome). GAS-C for all the three ontologies provides always more specific terms (average depth), i.e., more distant from the ontology root. Moreover, it always provides better coverage (number of predicted sequences).

Target-by-target performance

We compared GAS, GAS-C, and the other tools by evaluating their performance on the test set using the same approach adopted by the CAFA assessors. Table 4 reports the target-by-target maximum *F*-score (see "Performance evaluation"), computed on the test set targets, where all listed methods are able to make a prediction. The first consideration is that different methods behave differently for the three ontologies. Protein–protein interaction networks contain useful information about the biological process (BP) and the cellular compartment (CC) of a target. GAS does not produce good results for molecular function (MF). This is not surprising, since interacting proteins, even if they participate in the same biological process, usually carry out different biochemical reactions. For example, two proteins may be involved in the regulation of the cell cycle, but the first can be a regulatory protein performing phosphorylation and the second a transcription factor with a completely different biochemical attitude.

Table 4 Target-by-target performance

Ontology	Method	Precision	Recall	*F*-score	Rank cutoff
MF	GAS	0.342	0.261	0.296	1
	GAS-C	0.544	0.336	0.416	2
	BLAST	0.378	**0.395**	0.387	1
	NAÏVE	0.646	0.330	0.437	3
	FANN-GO	**0.901**	0.342	**0.496**	1
BP	GAS	0.314	**0.345**	0.329	3
	GAS-C	**0.405**	0.315	**0.355**	4
	BLAST	0.302	0.288	0.295	1
	NAÏVE	0.375	0.214	0.273	11
	FANN-GO	0.334	0.274	0.301	13
CC	GAS	0.487	**0.598**	0.537	2
	GAS-C	0.663	0.539	**0.595**	2
	BLAST	0.484	0.573	0.525	1
	NAÏVE	**0.776**	0.421	0.545	4
	FANN-GO	*	*	*	*

Performances are computed for entries where methods can make a prediction. The maximum *F*-score is used to select the corresponding precision and recall. The cutoff explains the number of top rank scores that should be considered to achieve the best *F*-score. The best performance for each ontology is highlighted in bold. * FANN-GO does not predict cellular component

The second observation is about the difference in terms of performance observed for the three ontologies in general. The BP terms are definitely the hardest to predict due to the more complex structure of the sub-ontology. Focusing the attention to the BP ontology is possible to observe the effectiveness of combining GAS and BLAST in the GAS-C consensus that rewards those terms predicted by both methods (see "GAS-C"). GAS-C obtains the maximum *F*-score over all methods even if the recall is penalized compared to GAS itself. To better appreciate the predictor performance, we plotted the precision-recall curves for all methods (Fig. 5). Both GAS and GAS-C are also good at predicting membrane proteins, ca. 1/3 of the entire validation set corresponding to 2870 proteins (highlighted in the provided prediction files).

Another important observation is the good performance obtained by the Naïve method for the MF terms in Table 4. This behavior was already observed during the CAFA experiment and is due to the very high frequency of proteins annotated with some shallow leaf terms very close to the root of the ontology. Naïve reaches a very high accuracy since it always predicts two ancestors of these leaf terms in the first positions ("protein binding" and "catalytic activity"). FANN-GO is subjected to the same phenomenon but achieves better results since the machine learning approach overcomes Naïve limitations.

In order to characterize different cases of predicted proteins, we measured the correlation between GAS *F*-score and the number of interactions available for a given target. We found that the Pearson correlation coefficient is very close to zero for all the three sub-ontologies (0.010 MF, −0.051 BP, −0.021 CC). When plotting the data (not shown), a slight decrease is observable when the number of interaction becomes larger than 10. This is a natural consequence of the enrichment procedure that fails to prioritize specific terms when the functional diversity of interacting partners is relevant and not specific, e.g., ubiquitin and chaperones.

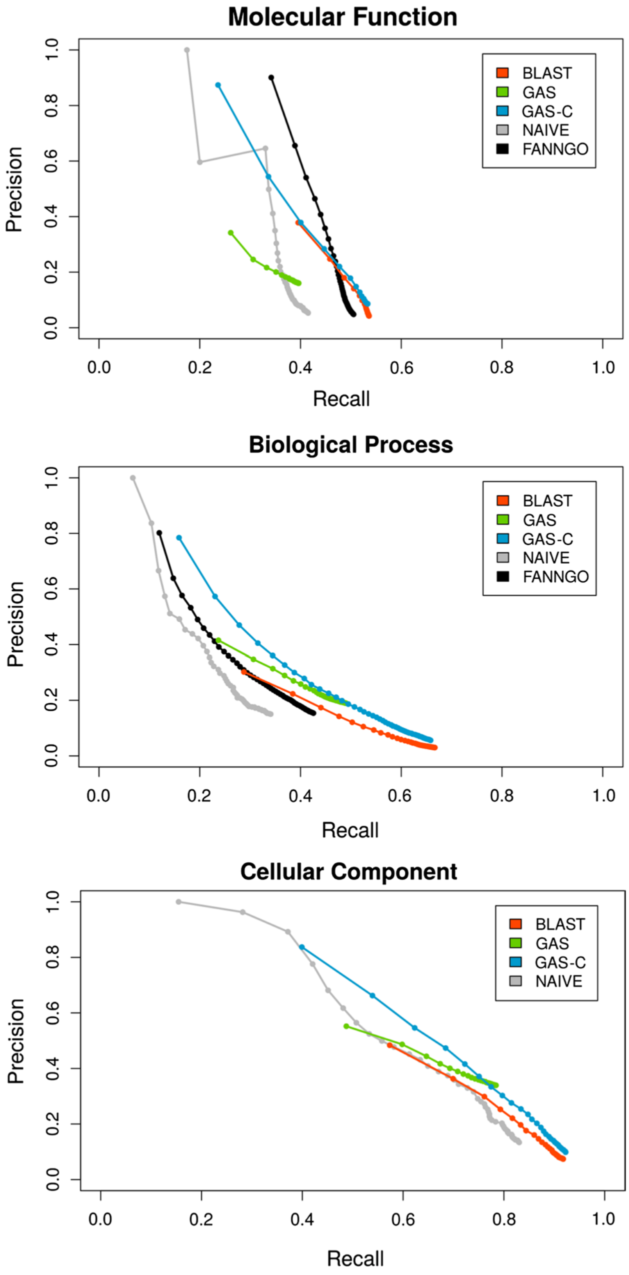

Fig. 5 Precision-recall curves. FANN-GO is missing in the cellular component chart because the tool does not provide prediction for that ontology

Table 5 Whole dataset performance

Ontology	Method	Precision	Recall	*F*-score	Score cutoff
MF	GAS	0.228	0.269	0.247	0.002
	GAS-C	0.637	0.320	0.426	0.455
	BLAST	0.300	0.327	0.313	85.1
	NAÏVE	0.646	0.330	0.437	0.362
	FANN-GO	**0.801**	**0.427**	**0.557**	0.214
BP	GAS	0.291	0.317	0.303	0.0006
	GAS-C	**0.450**	0.319	**0.373**	0.419
	BLAST	0.195	**0.368**	0.254	125.0
	NAÏVE	0.375	0.214	0.273	0.217
	FANN-GO	0.372	0.282	0.321	0.270
CC	GAS	0.339	**0.785**	0.474	0.998
	GAS-C	0.689	0.538	**0.604**	0.706
	BLAST	0.318	0.612	0.418	140
	NAÏVE	**0.776**	0.421	0.545	0.495
	FANN-GO	*	*	*	*

The performance is calculated over the same target set of Table 4 but for all possible thresholds for the score provided by the tools themselves. Score cutoff indicates the score threshold where the tool gets the best *F*-score. For every tool, the scores are *GAS P*-value, *GAS-C* Tool score, *BLAST* Bit-score, *NAÏVE* frequency, *FANN-GO* Tool score. The best performance for each ontology is highlighted in bold.
* FANN-GO does not predict cellular component

Whole dataset performance

One important aspect about the different predictors can be highlighted by a correlation analysis. For BLAST, we observed a limited relationship between the Bit-score and the *F*-measure for each target, with values below 0.290 for the three ontologies. Surprisingly, the same low correlation is also observed when considering sequence identity (not shown), suggesting that is very difficult to find a specific identity threshold useful for discriminating a good source of annotation.

For GAS, the same result holds for the enrichment *P*-value (correlation below 0.225). Such a limited correlation between *F*-score and the predictor confidence score suggested the use of ranks to improve results. We observed that BLAST generally achieved best results by just picking the GO terms associated to the first hit, i.e., sequence with the highest Bit-score. GAS ranks GO terms rather than sequences and has to consider up to the first three predicted terms, depending on the sub-ontology, to achieve optimal performance (see Table 4, column rank cutoff). This is likely to be the reason why GAS shows a higher maximum precision in general, while BLAST has a higher maximum recall (Table 4). Interestingly, the GAS-C score is strongly correlated with the expected *F*-score (correlation higher than 0.406), and outperforms the rank-based strategy as shown in Table 5. This is likely to be a consequence of a good fitting procedure. The increased predictive power shows that GAS and BLAST generate different knowledge. The consensus enables a better prioritization of predicted terms by using two orthogonal sources of information jointly and can truly guide a user to select GO terms depending on the expected annotation quality.

Discussion

In this paper, we presented a novel strategy to predict protein function exploiting protein–protein interaction (PPI) networks, developing a statistical significance estimation to rank GO terms. To the best of our knowledge, this is the first attempt to fairly evaluate the contribution of network interaction data to predict protein function.

GAS is based on the “Guilty-by-association” principle applied in the context of PPI networks. If a protein physically interacts with other proteins, it should share a similar function. For example, when all interacting partners operate inside the nucleus, it is reasonable to believe that the sub-cellular localization of a given target will be the nucleus itself.

However, even if the principle is very simple, some details need to be considered to implement an effective tool. Some aspects are related to the PPI network and others to the scoring function applied for ranking the predicted terms. The size of the PPI network and the reliability of the interactions affect the prediction in two different ways. A big network increases the probability of finding interacting partners endowed with GO annotation, while filtering low quality interactions correspond to a gain in the precision of the prediction. The number of interactors appears not relevant for performance, even if a slight precision loss is observable when the number of interaction partners becomes too large. The other key factor is the method used to sort and prioritize the transferred GO terms. We found that the *P*-value generated by measuring the enrichment of each collected annotation can be conveniently used to sort terms, but there is not a linear relationship between the *P*-value and the *F*-score that measures the quality of a prediction. In other words, it means that is not possible to say which could be an optimal *P*-value threshold that guarantees a good annotation. This is also true for BLAST where the Bit-score provided by the tool correlates very poorly with the *F*-score (Table 5). Conversely, both the Bit-score and the *P*-value provide a good sorting of GO terms and we found a good correlation between the *F*-score and the position (ranking) in the output list (compare GAS and BLAST *F*-score in Tables 4, 5). The comparison between GAS and BLAST highlighted important differences among the three GO sub-ontologies. As expected, PPI data are very effective for the CC and BP cases. On the other hand, evolutionary inference from sequence similarity represents a better discriminative approach for MF. This fact is consistent with the idea that network prediction can infer knowledge from the local neighborhood. Conversely, the molecular function cannot be directly inferred from the interactome, since the interacting proteins participating in a given biological process contribute themselves with different specific activities and biochemical reactions.

As shown in the first CAFA experiment, the performance of consensus methods is generally higher than standard tools. We implemented GAS-C that is able to generate a consensus prediction by combining both BLAST and GAS results. The implemented consensus strategy is extremely fast and simple, consisting in a score transformation, which can be generated in linear time with respect to the number of predictions. GAS-C achieves better results for all the three ontologies compared to BLAST and GAS themselves (Tables 4, 5) also for difficult cases like membrane proteins. It also outperforms FANN-GO for the BP ontology. A particular discussion has to be done for the Naïve performance in the MF ontology. The good *F*-score was already observed during the CAFA experiment and is due to a bias in SwissProt of some shallow leaf terms very close to the root of the ontology (see "Results").

In general, all presented results in terms of *F*-score, precision and recall are slightly underestimated compared to the numbers provided by the first CAFA experiment. This happened because we evaluated all the predictions without filtering those terms in the test set not yet available 1 year before in 2012. However, this does not change the validity of this work, since all methods were affected equally by this problem. At the moment, sequence similarity approaches outperform GAS in terms of target coverage, but we believe that good quality interaction data are going to increase consistently, resulting in a better capacity to generate new hypotheses. Moreover, PPI networks represent a complementary source of knowledge compared to evolutionary information, and will be even more effective in the future, when entire organism interactomes will become available. Future GAS extensions may leverage the presence of disordered regions (Potenza et al. 2015) or repetitive units (Di Domenico et al. 2014) to improve the background distribution for enrichment calculation, thereby increasing term specificity.

Acknowledgments The authors are grateful to members of the BioComputing UP lab for insightful discussions. This project was funded by FIRB Futuro in Ricerca grant RBFR08ZSXY, University of Padua grant CPDR123473, and AIRC grant MFAG12740 to S.T. D.P. is funded by FIRC Fondazione Italiana per la Ricerca sul Cancro project no. 16621.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Altschul S (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi:10.1006/jmbi.1990.9999

Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29. doi:10.1038/75556

Barabási A-L, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12:56–68. doi:10.1038/nrg2918

Brun C, Chevenet F, Martin D et al (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol 5:R6. doi:10.1186/gb-2003-5-1-r6

Chua HN, Sung W-K, Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinforma Oxf Engl 22:1623–1630. doi:10.1093/bioinformatics/btl145

Clark WT, Radivojac P (2011) Analysis of protein function and its prediction from amino acid sequence. Proteins Struct Funct Bioinforma 79:2086–2096. doi:10.1002/prot.23029

Cozzetto D, Buchan DWA, Bryson K, Jones DT (2013) Protein function prediction by massive integration of evolutionary analyses and multiple data sources. BMC Bioinformatics 14:S1. doi:10.1186/1471-2105-14-S3-S1

Deng M, Zhang K, Mehta S et al (2003) Prediction of protein function using protein-protein interaction data. J Comput Biol J Comput Mol Cell Biol 10:947–960. doi:10.1089/106652703322756168

Di Domenico T, Potenza E, Walsh I et al (2014) RepeatsDB: a database of tandem repeat protein structures. Nucleic Acids Res 42:D352–D357. doi:10.1093/nar/gkt1175

Dimmer EC, Huntley RP, Alam-Faruque Y et al (2011) The UniProt-GO annotation database in 2011. Nucleic Acids Res 40:D565–D570. doi:10.1093/nar/gkr1048

Engelhardt BE, Jordan MI, Srouji JR, Brenner SE (2011) Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res 21:1969–1980. doi:10.1101/gr.104687.109

Franceschini A, Szklarczyk D, Frankild S et al (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41:D808–D815. doi:10.1093/nar/gks1094

Hishigaki H, Nakai K, Ono T et al (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast Chichester Engl 18:523–531. doi:10.1002/yea.706

Ho Y, Gruhler A, Heilbut A et al (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415:180–183. doi:10.1038/415180a

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316:1497–1502. doi:10.1126/science.1141319

Minneci F, Piovesan D, Cozzetto D, Jones DT (2013) FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences. PLoS One 8:e63754. doi:10.1371/journal.pone.0063754

Nabieva E, Jim K, Agarwal A et al (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinforma Oxf Engl 21(Suppl 1):i302–i310. doi:10.1093/bioinformatics/bti1054

Pellegrini M, Marcotte EM, Thompson MJ et al (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci 96:4285–4288. doi:10.1073/pnas.96.8.4285

Piovesan D, Luigi Martelli P, Fariselli P et al (2011) BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences. Nucleic Acids Res. doi:10.1093/nar/gkr292

Piovesan D, Giollo M, Leonardi E et al (2015) INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. Nucleic Acids Res 43:1–5. doi:10.1093/nar/gkv523

Potenza E, Di Domenico T, Walsh I, Tosatto SCE (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res 43:D315–D320. doi:10.1093/nar/gku982

Radivojac P, Clark WT, Oron TR et al (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10:221–227. doi:10.1038/nmeth.2340

Suzek BE, Huang H, McGarvey P et al (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinforma Oxf Engl 23:1282–1288. doi:10.1093/bioinformatics/btm098

Zhu H, Snyder M (2003) Protein chip technology. Curr Opin Chem Biol 7:55–63. doi:10.1016/S1367-5931(02)00005-4