Advance Access Publication Date: Day Month Year
Original Paper

Systems Biology

Sparsity of Protein-Protein Interaction Networks Hinders Function Prediction in Non-Model Species

Stavros Makrodimitris [1,2,*], Roeland C.H.J. van Ham [1,2] and Marcel J.T. Reinders [1,3*]

[1] Delft Bioinformatics Lab, Delft University of Technology, Van Mourik Broekmanweg 6, 2628XE, Delft, the Netherlands,
[2] Keygene N.V., Agro Business Park 90, 6708PW, Wageningen, the Netherlands, and
[3] Leiden Computational Biology Center, Leiden University Medical Center, Einthovenweg 20, 2333ZC, Leiden, the Netherlands.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Physical interaction between two proteins is strong evidence that the proteins are involved in the same biological process, making Protein-Protein Interaction (PPI) networks a valuable data resource for predicting the cellular functions of proteins. However, PPI networks are largely incomplete for non-model species. Here, we test whether these incomplete networks are still useful for genome-wide function prediction.

Results: We used a simple network-based classifier to predict Biological Process Gene Ontology terms from protein interaction data in three species: *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Solanum lycopersicum* (tomato). The classifier had reasonable performance in the well-studied yeast, but performed poorly in the other two species. We show that this poor performance is because many proteins are disconnected in the network and that the performance can be considerably improved by adding edges predicted from various data sources. In yeast, the addition of predicted edges did not lead to improvement. It did help when we randomly removed a large amount of edges though.

Conclusion: Our work highlights the necessity of obtaining more protein-protein interactions in non-model species, either by means of prediction or experiment.

Availability: Data and code to reproduce the results are available at github.com/stamakro/ppi-missing-data.

Contact: s.makrodimitris@tudelft.nl

Supplementary information: Supplementary data are available online.

1 Introduction

One of the main challenges of the postgenomic era is how to extract functional information from the vast amount of sequence data that are available. As the number of known protein sequences grows at a very fast pace, experimentally determining the functions of all proteins has become practically infeasible. This creates the need for accurate Automatic Function Prediction (AFP) methods, which can predict a protein's function(s) using the knowledge that has been accumulated in the past. To this end, the Gene Ontology (GO) is a very valuable resource that provides a systematic representation of function in the form of three ontologies: Biological Process (BP), Molecular Function (MF) and Cell Component (CC) (Ashburner *et al.*, 2000).

The Critical Assessment of Function Annotation (CAFA) is a community-driven benchmark study that compares a large number of available AFP methods in an independent and systematic way (Radivojac *et al.*, 2013; Jiang *et al.*, 2016; Zhou *et al.*, 2019). One of the main conclusions that one can draw from the several editions of CAFA is that top-performing methods tend to use a combination of different data sources and not only the amino acid sequence. For example, MS-kNN, one of the best methods in CAFA2, combined sequence similarity with gene co-expression and protein-protein interaction (PPI) data (Lan *et al.*, 2013). GOLabeler, which was the best in CAFA3, combined six different data sources with a powerful algorithm that predicts how suitable a GO term is for the input protein (You *et al.*, 2018). More recently, the authors of GOLabeler introduced an extension named NetGO which also uses PPI networks as an extra data source, reporting even better performance than GOLabeler on the CAFA3 dataset (You *et al.*, 2019). These observations show that PPI networks are informative data sources for AFP, which can be understood, since if two proteins physically interact, they are likely to be involved in the same biological process or pathway.

Table 1. Number of proteins and known PPI's per species in BIOGRID. (version 3.5.171)

	Yeast	Arabidopsis	Tomato
#protein-coding genes	6,000 (Engel *et al.*, 2014)	27,029 (Lamesch *et al.*, 2012)	34,727 (Suresh *et al.*, 2014)
#proteins with BPO annotations	4,962	10,460	670
#experimental edges between proteins with BPO annotations	146,434	27,049	57
#pairs of proteins with BPO annotations	12,308,241	54,700,570	224,115
% protein pairs interacting	1.18	0.05	0.03
% disconnected proteins	2.0	45.3	97.0

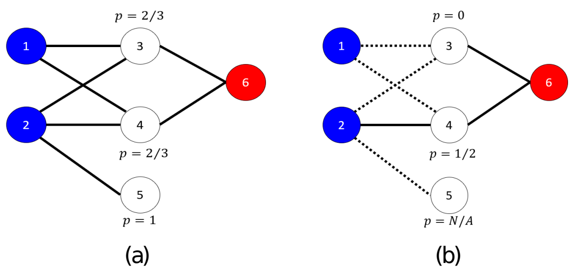


Fig. 1. Toy PPI network with 6 nodes. Nodes annotated with a GO term are shown in blue and nodes not annotated in red. Unlabeled (test) nodes are shown in white. In (a) the entire network is known and the posterior probabilities for each unlabeled node can be calculated accurately. In (b) some of the edges are missing (signified by the dashed lines), making the calculation of posterior probabilities either erroneous or even impossible (e.g. node 5).

However, almost all PPI networks are incomplete. The best-characterized model species, *Saccharomyces cerevisiae* (baker's yeast), has one of the densest PPI networks, with 115,729 experimentally-derived, physical interactions in the BIOGRID database (Oughtred *et al.*, 2019). Given the fact that *S. cerevisiae* has about 6,000 protein-coding genes (Engel *et al.*, 2014), this means that roughly 0.6% of all possible pairs of proteins are known to interact. The human interactome is also quite well characterized, with 389,300 experimental interactions in BIOGRID (about 0.2% of all possible interactions). Moreover, a recent study identified 52,569 high-quality interactions of 8,275 human proteins (Luck *et al.*, 2019). On the other hand, in *Arabidopsis thaliana*, the most well-studied plant species, there are about 27,000 protein coding genes and 48,577 experimentally-derived physical interactions in BIOGRID, i.e. only 0.01% of the possible interactions are known. This is not likely due to protein interactions being less common in *A. thaliana*, but rather because it is not as well-studied as yeast.

The number of known edges is orders of magnitude smaller in other plant species, even in very important crops. For example, in tomato (*Solanum lycopersicum*), there are only 107 interactions in BIOGRID as of June 2019 («0.01% of the total number of possible interactions). In rice (*Oryza sativa japonica*), there are 327 and in corn (*Zea mays*) 13. This phenomenon is not restricted to plants, but is true for many non-model animal species of major economic importance, such as cow (*Bos taurus*, 433) and pig (*Sus scrofa*, 80 interactions).

Most methods that employ PPI networks in AFP predict functions by propagating the GO annotations through the network (Lan *et al.*, 2013; You *et al.*, 2019). The simplest of such methods transfers the annotations of a protein to its immediate neighbors. This is also known as Guilt-By-Association (*GBA*). Figure 1a illustrates the *GBA* method in an example network with 6 proteins: Proteins 1 and 2 are annotated with a GO term, while protein 6 is not. We are asked to predict whether proteins 3-5 should be annotated with that GO term. As seen in Figure 1a, for all three of these proteins we are at least 66.6% certain that they should be assigned that GO term. Figure 1b shows the same example network, assuming that some of its edges are missing. In this case, protein 5 has no known interacting partners, so it is impossible to determine its function. Similarly, protein 1 has a known function, but is disconnected from the rest of the network, so its function cannot be propagated to other proteins. This example shows that when interactions in a PPI network are missing, function prediction cannot benefit from PPI information (as most proteins will have few or no connections to other proteins).

A way to counter the lack of edges is to predict them using other data sources. The STRING database contains a large collection of protein associations predicted using different sources, such as gene co-expression and text mining (Szklarczyk *et al.*, 2015). Moreover, the recent rise in popularity of deep learning has caused an increase in methods that attempt to predict protein-protein interactions purely from protein sequence. One of the first examples was from Sun et al. (Sun *et al.*, 2017), followed by DPPI (Hashemifar *et al.*, 2018), PIPR (Chen *et al.*, 2019) and the work of Richoux et al. (Richoux *et al.*, 2019). The advantage of predicting edges from sequence is that it is - at least in theory - not biased towards previous experiments. In contrast to, for example, predictions within the STRING database that still require other people to have previously studied a specific protein or its orthologues. Having an accurate sequence-based predictor of PPI's means that we can feed it with all possible pairs of proteins and obtain a score for how probable each interaction is. This enables us to find possible interacting partners for proteins that have not been previously studied at all, thereby improving function prediction for those proteins.

In this study, we are interested in quantifying the influence of missing edges in a PPI network on protein function prediction. This will give valuable information on whether to use this data resource for predicting protein function in organisms with a less-well measured PPI network. Moreover, we are interested in how well (deep learning based) sequence-based PPI predictors can recuperate this missing information, and how that translates in improvements of the function prediction. We show that in species with few known interactions, the performance of a network-based AFP method is significantly worse than a simple baseline and that this performance increases as we add predicted edges. We also show that in species with a dense PPI network, predicted edges do not provide any performance improvement for AFP.

2 Materials & Methods

2.1 Protein-Protein interaction networks

We compared PPI networks in *S. cerevisiae*, *A. thaliana* and *S. lycopersicum* using three types of PPI's: 1) Physical interactions that have been experimentally derived. 2) Predicted interactions based on non-experimental protein association data from the STRING database, and 3) Sequence-based predicted interactions based on the amino acid sequence of two proteins using PIPR.

Physical interactions: For the experimental interactions we used the BIOGRID (version 3.5.171) (Oughtred *et al.*, 2019) and STRING databases (Szklarczyk *et al.*, 2015). We only used physical interactions and ignored the genetic interactions. Of note, the STRING database contains a collection of experimental protein-protein interactions from different databases, including BIOGRID (marked with the "experiments" data source code) and we found edges in BIOGRID that were not present in STRING. From STRING, we only chose experimental protein-protein interactions with association scores larger than the median score over the non-zero scores for each species individually.

Predicted interactions: Besides the experimental evidence, STRING contains predicted protein associations from 12 data sources in total: neighborhood, neighborhood transferred, co-occurrence, database, database transferred, experiments transferred, fusion, homology, co-expression, co-expression transferred, text mining and text mining transferred. Table S1 (Supplementary Material 1) shows the number of interactions per species and per data type. We removed two data sources ("neighborhood" and "database transferred") because they did not add any new edges in any of the three species. We also removed "database", as it includes protein associations that were identified by using the GO annotations of proteins and these edges would cause circular reasoning if used to predict GO terms, leading to a biased evaluation. This left us with 9 data sources from which we could infer PPI's. In tomato, there were no predicted associations with the "co-expression" evidence code, so we only used 8 data sources for that species. From all these sources, we only selected the protein pairs with the 50% highest non-zero scores for each data source and species individually. Next to individually using the data sources as proxies for the protein-protein interactions, we also combined data sources. This was done by first integrating the STRING scores from different sources as described in (von Mering *et al.*, 2005) (see Supplementary Material 2 for more information) and then keeping the 50% top non-zero scores for every combination. To combine a binary STRING network with the experimental one, we applied an element-wise logical OR to the corresponding adjacency matrices, so an interaction is added to the combined network if it is present in at least one of the original networks.

Sequence-based predicted interaction: We used PIPR (Chen *et al.*, 2019) to predict PPI's from protein sequence. It uses a Siamese twin architecture with both convolutional and recurrent units and three fully connected layers at the end. PIPR also makes use of predefined amino acid embeddings, obtained from both chemical properties of amino acids and their co-occurence in protein sequences. The original PIPR model had an accuracy of about 97% in predicting yeast PPI's when trained on a large, balanced dataset from the DIP database. The originally trained PIPR model in yeast generalized poorly in Arabidopsis (accuracy of 51% on a balanced dataset). Therefore, we chose to train PIPR on Arabidopsis by using the original trained model as initial conditions. We trained using Stochastic Gradient Descent with learning rate 0.001 and early stopping based on the validation loss with patience of 40 epochs. As validation set, we randomly selected 10% of the data and as loss function the binary cross-entropy. We did not use RMSprop optimizer, which was used by the authors, as it produced unstable results. After having trained a model, we feed it all pairs of proteins. For each pair we get a score in the range [0,1] denoting the probability that these two proteins interact. We add an edge to our predicted PPI network if the score for that edge is greater than or equal to 0.5. For tomato, we did not retrain, but rather used the model trained on *A. thaliana*.

2.2 GO annotations

We obtained GO annotations from the GOA website (Huntley *et al.*, 2015) and only used the experimental annotations and curated annotations (evidence codes "EXP", "IDA", "IPI", "IMP", "IGI", "IEP", "IBA", "IBD", "IKR", "IRD", "TAS" and "NAS"). We focused on the Biological Process Ontology (BPO), as it is the most difficult ontology to predict (Jiang *et al.*, 2016) and also is the most commonly used in further analyses such as gene set enrichment. Table 1 gives an overview of the different dataset sizes for the three species.

2.3 Network-Based function prediction

We represent the protein-protein interactions as a network with the proteins as nodes and the interactions as binary, undirected edges. Let A be the network's adjacency matrix, V_{train} a set of training proteins and V_{test} a set of test proteins. Moreover, let $T(p)$ be the set of GO terms assigned to $p \in V_{train}$. Given this representation we are using three different approaches to predict protein function: 1) a random approach to be able to compare performance values to those that could be achieved by chance, 2) the naïve approach of CAFA that we consider as baseline, and 3) a Guilt-By-Association approach (*GBA*).

Random-approach: The posterior probability that a protein is associated with a GO term ($P(p_i, t)$) is drawn from a uniform distribution between $[0, 1]$. Note that this method can produce predictions that are inconsistent with the ontology graph, as it is possible that a GO term gets a higher score than its ancestors.

Baseline approach: This method follows the "naïve" method of CAFA (Radivojac *et al.*, 2013) and assigns a GO term to a protein with probability equal to the fraction of training proteins annotated with that term (equation 1).

$$P(p_i, t) = \frac{|\{p : p \in V_{train} \wedge t \in T(p)\}|}{|V_{train}|} \tag{1}$$

This means that all test proteins get the same annotation using this method (making it a quite weak baseline).

Guilt-By-Association approach (GBA): In this method functions are transferred to a protein from its direct interacting partners. For a protein $p_i \in V_{test}$, we define its neighborhood $N(p_i)$ as all its interacting partners that are in the training set:

$$N(p_i) = \{p : p \in V_{train} \wedge A[p, p_i] = 1\} \tag{2}$$

For a GO term t, the probability it is assigned to test protein p_i is equal to the fraction of its annotated neighbors that are annotated with t (equation 3):

$$P(p_i, t) = \frac{\sum_{p \in N(p_i)} I(t \in T(p))}{|N(p_i)|} \tag{3}$$

Where $I(x) = 1 \, iff \, x$ is a true statement and $|S|$ denotes the number of elements in set S.

2.4 Experimental set-up

To compare function prediction across the differently constructed protein-protein interaction networks, we followed a 5-fold cross-validation. Our main evaluation metric was the protein-centric Area Under the precision-recall Curve (*AUC*), but we also applied the F_{max} and normalized S_{min} that are extensively used in the CAFA challenges (Table S2, Supplementary Material 3). In each fold, we discarded the GO terms that had no positive examples in either the training or the test set. We compared three PPI networks: 1) the experimental PPI network (*EXP*), 2) combined experimental and predicted PPI networks using functional genomic data from the STRING database (*FG-STRING*), and 3) the sequence-based predicted PPI networks (*SEQ*).

Experimental PPI - EXP: We started from the experimental PPI network of a given species. This network includes as nodes all proteins that have at least 1 functional annotation, even if they have no interacting partners. Proteins without functional annotations were removed, even if they had known interactions.

Combined experimental and predicted PPI - FG-STRING: We added predicted edges to the experimental network from the different data sources in STRING. We evaluated all possible combinations of the 9 STRING data sources (8 for tomato): First, we added each data source individually. Then, we tested all combinations of 2 data sources (36 possibilities), all combinations of 3 (84 possibilities) and so on, until we have included all 9 data sources. So, in total, we tested $\sum_{i=1}^{9}\binom{9}{i} = 511$ combinations of data sources (255 for tomato) along with the experimental network.

Sequence-based predicted PPI - SEQ: We used edges predicted by PIPR for predicting function. We tested the performance of a network with only the PIPR edges and a network with the experimental edges combined with the PIPR predictions.

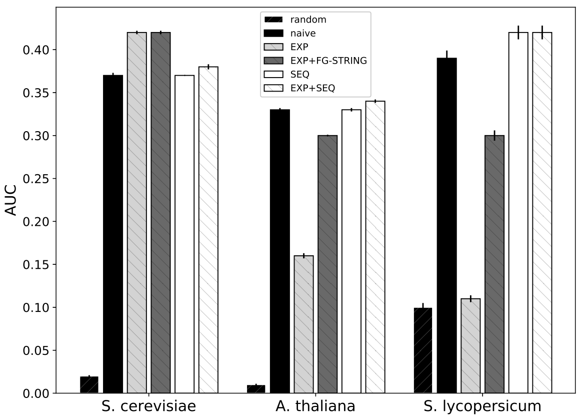


Fig. 2. AUC (y-axis) achieved by the baseline methods (black) and the GBA method using different data types for the three species (x-axis). The error bars denote the standard error over the 5 cross-validation folds.

3 Results

3.1 The naïve method outperforms *GBA* in plant experimental PPI networks

Figure 2 and Table S3 (Supplementary Material 4) show the AUC values for the three studied species. In yeast, the simple *GBA* method on the *EXP* network outperforms the naïve baseline. In Arabidopsis and tomato, the picture is quite the opposite, with the naïve method largely outperforming *GBA*. In tomato, the *EXP* network performs only slightly better than random (Figure 2). This shows that existing experimental PPI networks are insufficient for any kind of protein function prediction in plants.

When inspecting the performance of the *EXP* network for each individual protein (Figure S1, Supplementary Material 4), we can see that for most Arabidopsis and tomato proteins the AUC is close or equal to 0. This is mainly because for many proteins the interacting partners are not known, and consequently no predictions can be made. Indeed, when calculating the Spearman correlation between the node degree in the *EXP* network of Arabidopsis with protein-centric performance (AUC) we find a correlation of $\rho = 0.75$. In yeast, the performance also largely depends on node degree, but not as much as in *A. thaliana* ($\rho = 0.29$). Moreover, in yeast the AUC values are more spread in the range $[0, 1]$ compared to the other species (Figure S1, Supplementary Material 4). This implies that in yeast there are proteins with varying levels of difficulty for the classifier, while in Arabidopsis and tomato most proteins are difficult, having performance close to 0. We obtained similar results when evaluating with the F_{max} and S_{min} metrics (Table S2, Supplementary Material 3).

3.2 Adding predicted edges is beneficial for plant PPI networks, not for yeast

We tested whether predicted interactions from other data sources can improve protein function prediction performance. As we can see in Figure 2, combining the experimental PPI network of *S. cerevisae* with other protein association networks (see methods) has almost no effect on the average protein-centric performance.

On the other hand, in Arabidopsis and tomato we observed a substantial improvement in performance by adding predicted edges, although the naïve classifier still outperforms them (Figure 2). The most informative data source in Arabidopsis was "text mining" ($AUC = 0.28$) and in tomato "text mining transferred" ($AUC = 0.26$). These data sources are almost by themselves able to explain the maximum performance ($AUC = 0.31$ in Arabidopsis and 0.29 in tomato). Figure S2 (Supplementary Material 4) shows the AUC values for each combination of FG-STRING data sources for the three species and Figure S3 (Supplementary Material 4) shows the AUC of the individual data sources in tomato.

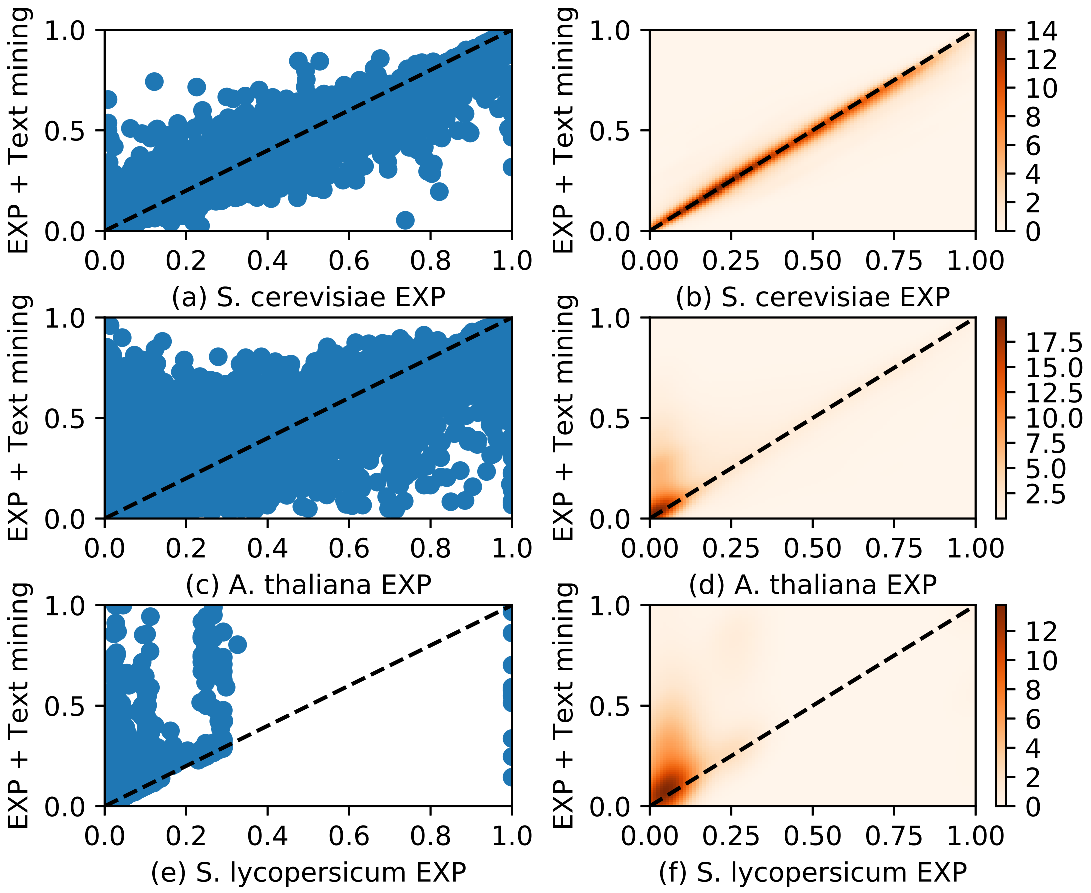
The Spearman correlation between node degree in the experimental network of Arabidopsis and AUC of the combined *EXP* and "text mining" network was reduced to 0.48, which hints at the fact that the predicted edges often manage to connect functionally related proteins that were previously disconnected. Indeed, in Figures 3c-d, we see that many proteins that have an AUC close to 0 in the EXP network get higher values when combining that network with the "text mining" network. The performance difference as a function of node degree is shown in Figure S4b (Supplementary Material 5). The same pattern is also evident for tomato (Figures 3e-f and S4c), while in yeast the functions of the vast majority of the proteins can be predicted about equally well using any of the two networks (Figures 3a-b and S4a). For all three species there are cases where adding edges actually reduces the performance, but this is much rarer according to the densities in Figures 3b, d and f.

3.3 Function Prediction with *SEQ* edges

Notably, the sequence-based predicted PPI network hampers the AFP performance in yeast as compared to the *EXP* PPI (Figure 2). This is probably due to the addition of false positive edges. In contrast, in Arabidopsis and tomato the *SEQ* PPI network seems to be useful, providing significant improvements (Figure 2) compared to the combined *EXP+FG-STRING* PPI networks, even outperforming the naïve classifier.

3.4 Removing edges from the yeast network gives similar results

We randomly removed 90% of the edges from the yeast experimental PPI network and repeated the experiment. In this case, we observed a pattern more similar to the one observed in our plant data (Figure S5, Table S4, Supplementary Material 6). The experimental network with missing edges (*EXP-10%*) is significantly outperformed by the naïve method, despite the fact that it is about 10 times denser than the plant *EXP* networks. The addition of predicted edges from FG-STRING to *EXP-10%* network leads to an improvement, however the performance remains significantly worse than that of the full *EXP* network (Tables S3, S4). As in the case of Arabidopsis and tomato, "text mining" was by far the most informative data source when added to the reduced yeast network. In fact, it was enough



(a) S. cerevisiae EXP
(b) S. cerevisiae EXP
(c) A. thaliana EXP
(d) A. thaliana EXP
(e) S. lycopersicum EXP
(f) S. lycopersicum EXP

Fig. 3. Comparison of the AUC of the experimental PPI network (x-axis) against that of the "experimental" network complemented with the "text mining" network (y-axis) for yeast (a ,b),A. thaliana (c, d) and tomato (e, f). a,c,e): Each blue dot corresponds to a protein. b,d,f): Gaussian-kernel based density of the number of proteins, with regions of low density appearing in light color and regions of high density in dark.

by itself to achieve the top performance. Not including the "text mining"-derived edges in our experiments, we managed to obtain similar maximum performance, but with more data sources, with "co-expression" being the driving force of the improvement in that case.

We further investigated how the performance of the GBA EXP classifier varies as a fraction of the amount of missing edges (Figure 4 and Supplementary Material 6). We observed that when 60% or more of the edges are removed, the performance of the experimental network starts decreasing very rapidly, eventually reaching zero when we remove all edges. The performance of the predicted network also decreases, but at a much slower rate, demonstrating the usefulness of predicted edges in cases where many experimental edges are not known.

Figure 4 also shows the performance when using the sequence-based predicted PPI. The performance of the sequence-based PPI is higher than that of the naïve method and lower than that of the complete experimental PPI network, but when the number of experimental edges drops below 30% then the sequence-based predicted PPI outperforms all other methods, including "text-mining" based networks.

4 Discussion

The aim of this work was to investigate the applicability of protein-protein interaction networks to genome-wide function prediction in not well-studied species. For that purpose, we evaluated the performance of an AFP method based on a PPI network in three species with different degrees of missing annotations.

We did not compare the network-based classifier to any state-of-the-art methods, such as GOLabeler (You *et al.*, 2018) or INGA (Piovesan

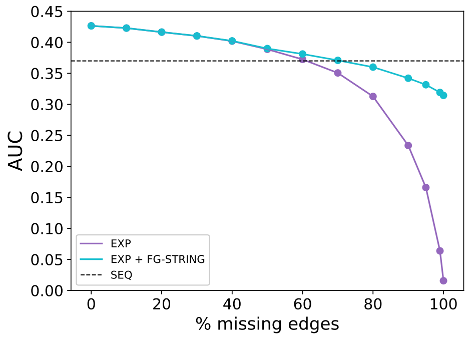


Fig. 4. AUC performance in S. cerevisiae as a function of the percentage of missing edges in the "experimental" PPI network (purple). This performance is compared to the AUC performance when the sampled experimental PPI is supplemented with the predicted PPI (light blue). The performance of the SEQ network obtained using PIPR is shown as a dashed horizontal line.

and Tosatto, 2019), but rather to the so-called “naïve” classifier from the CAFA challenges. This classifier, as its name suggests, does not use any information to relate specific proteins to GO terms, rather it only uses the frequency of each GO term in the training set. In the machine learning literature, this classifier is called the "Bayesian Marginal Predictor" (El-Yaniv *et al.*, 2017) and it corresponds to the optimal classifier when the distributions of the classes ($P(y)$) are known, but information about the relationship between the data and the classes ($P(x|y)$) is missing. This means that any classifier that uses any kind of (informative) data is expected to outperform the naïve one.

However, we clearly demonstrated the failure of the *GBA* classifier in predicting BPO terms in *A. thaliana* and tomato, as it performed considerably worse than the naïve method. In tomato, the *GBA* method barely outperformed the random one. This is probably due to the very small dataset size, which makes it easier to predict correctly by chance. This was not the case in yeast, where the *GBA* classifier outperformed the naïve one. When examining the performance for individual proteins, we found a high correlation between the number of known interacting partners and the prediction accuracy. These validate our hypothesis that a sparse PPI network is detrimental to genome-wide AFP.

The computational prediction of protein-protein interactions has been an active research area for many years (Valencia and Pazos, 2002; Jansen *et al.*, 2003). Our work is the first to evaluate the contribution of predicted edges in protein function prediction in a species-specific way. We used the STRING database as a proxy for predicting interaction using omics data such as genome features, homology, co-expression and text mining. In the very sparse experimental plant PPI networks, the *FG-STRING*-derived edges contribute a great deal, increasing the performance almost 2-fold in *A. thaliana* (from 0.16 to 0.30) and almost 3-fold in tomato (from 0.11 to 0.30). This is because these extra edges connect proteins that were previously disconnected from the rest of the graph. But, remarkably, it was still not enough to reach the performance of the naïve classifier, hinting that these information sources do not contribute knowledge about the posterior distributions ($P(y|x)$). In the case of yeast, which has a “complete” network, the *FG-STRING*-derived edges did not affect the performance, again showing that they did not bare additional information. Because yeast has a quite dense interactome, it offers us the opportunity to remove edges at random in order to obtain insights about how the AFP performance changes as a function of the number of edges (Figure 4). We observed that we can remove up to 40% of the edges (resulting in about 0.7% of all possible protein interactions) with only a minor drop in performance and up to 60% (about 0.5% of all possible interactions) while performing better than the naïve classifier. Moreover, we found that only after removing at least 60% of the edges (i.e. one order of magnitude denser than the *A. thaliana* and tomato networks), we start observing a benefit from adding *FG-STRING* edges, but this benefit is not enough to reach the performance of the full experimental network. These exact numbers are likely to differ across species, but they are indicative of the relative resilience of the PPI networks, after they become "complete enough".

Notably, text mining was the most informative STRING data source for all three species (for yeast in the case with simulated missing edges). Interestingly, from the descriptions of the methods submitted to the CAFA challenges, we know that only a small minority of them make use of text mining (Zhou *et al.*, 2019). Some of these methods are (De Bie *et al.*, 2007; Jaeger *et al.*, 2008). Our work shows that perhaps text mining is an underrated data source for functional annotation.

We also applied a sequence-based neural network model (PIPR) in PPI edge prediction. Firstly, we noticed that although PIPR was very accurate in predicting edges in yeast, it did not generalize in *A. thaliana*, performing very close to random guessing. However, starting from the PIPR model trained in yeast and re-training it with the known *A. thaliana* edges yielded a much higher accuracy in a held-out validation set. Also, applying this newly trained model on all pairs of Arabidopsis proteins significantly improved the *GBA* classifier compared to the predicted networks from STRING, reaching performance similar to the naïve classifier. Note that a performance boost was attained even without using the experimental edges, i.e. only with predicted ones.

Yeast and Arabidopsis are evolutionarily distant (about 1,200 million years (Hedges *et al.*, 2015)), which might explain why the PIPR model did not immediately generalize from the one to the other. On the contrary, our results show that a model trained in Arabidopsis can be very useful for predicting functions in tomato, a much more closely related species (about 120 million years distance (Hedges *et al.*, 2015)) with almost no known PPI’s. This suggests that it is informative to always test the generalization ability of such neural network models in other species.

A limitation of our study is that except for the variable degree of unknown PPIs among the tested species, there is also large variability in the amount of missing experimental annotations, with yeast being the most well-characterized species followed by Arabidopsis. This means that it is much more likely that a correctly predicted protein-GO term pair is flagged as a false positive in tomato than in yeast, simply because that annotation has not been discovered yet. Moreover, the GO terms have different frequencies in the three species, meaning that is virtually impossible to compare performances across species. For example, yeast contains a lot more "deeper", more specific annotations than e.g. tomato. This is not an issue in our analyses because we do not focus on the exact performance values, but rather on how the performances of different networks (i.e. networks with different edge types) compare to the performance of the naïve method within a species. Also, we have shown that the same conclusions can be drawn when using the semantic distance instead, which punishes shallow predictions.

A possible way to overcome the fact that performance values are not comparable across species, would be to report the performance for each protein relative to the performance of the naïve method for that protein. That could be done as follows by using the transformation suggested in (El-Yaniv *et al.*, 2017):

$$AUC_{corr} = \frac{1 - AUC_{observed}}{1 - AUC_{naive}} \tag{4}$$

Using this transformation, the naïve method has performance equal to 0 for each protein and each evaluated classifier has a positive score if it outperforms the naïve one (with a maximum of 1, if it makes perfect predictions) and a negative one if it is worse. This would make comparisons across species easier, although this measure is not defined if for a protein the naïve method achieves perfect performance, but such cases are extremely rare.

In conclusion, our work highlights the difficulty of applying PPI networks in AFP for non-model species. We show that predicted PPIs can partially compensate for the sparsity of the networks but cannot surpass the high-quality experimental networks that exist for model organisms. Importantly, too many predicted edges can have a negative impact when added to a good experimental network. Perhaps, that calls for a shift in the focus of the research community: At least for the aim of predicting function, it is not nearly as useful to predict edges in well-characterized model species as to non-model species. Also, for non-model species of great interest it would be beneficial to obtain more experimental edges.

Funding

This work has been supported by Keygene N.V., a crop innovation company in the Netherlands.

Conflicts of interest

None declared.

References

Ashburner, M. *et al.* (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.

Chen, M. *et al.* (2019). Multifaceted protein-protein interaction prediction based on Siamese residual RCNN. In *Bioinformatics*.

De Bie, T. *et al.* (2007). Kernel-based data fusion for gene prioritization. In *Bioinformatics*.

El-Yaniv, R. *et al.* (2017). The Prediction Advantage: A Universally Meaningful Performance Measure for Classification and Regression.

Engel, S. R. *et al.* (2014). The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3: Genes, Genomes, Genetics*.

Hashemifar, S. *et al.* (2018). Predicting protein-protein interactions through sequence-based deep learning. In *Bioinformatics*, volume 34, pages i802–i810.

Hedges, S. B. *et al.* (2015). Tree of Life Reveals Clock-Like Speciation and Diversification. *Molecular Biology and Evolution*, **32**(4), 835–845.

Huntley, R. P. *et al.* (2015). The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Research*, **43**(D1), D1057–D1063.

Jaeger, S. *et al.* (2008). Integrating protein-protein interactions and text mining for protein function prediction. In *BMC Bioinformatics*.

Jansen, R. *et al.* (2003). A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science*.

Jiang, Y. *et al.* (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, **17**(1), 184.

Lamesch, P. *et al.* (2012). The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research*.

Lan, L. *et al.* (2013). MS-kNN: protein function prediction by integrating multiple data sources. *BMC bioinformatics*, **14 Suppl 3**(Suppl 3), S8.

Luck, K. *et al.* (2019). A reference map of the human protein interactome. *bioRxiv*.

Oughtred, R. *et al.* (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Research*.

Piovesan, D. and Tosatto, S. C. E. (2019). INGA 2.0: improving protein function prediction for the dark proteome. *Nucleic Acids Research*.

Radivojac, P. *et al.* (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, **10**(3), 221–227.

Richoux, F. *et al.* (2019). Comparing two deep learning sequence-based models for protein-protein interaction prediction.

Sun, T. *et al.* (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics*, **18**(1).

Suresh, B. V. *et al.* (2014). Tomato genomic resources database: An integrated repository of useful tomato genomic information for basic and applied research. *PLoS ONE*.

Szklarczyk, D. *et al.* (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*.

Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, **12**(3), 368–373.

von Mering, C. *et al.* (2005). STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*.

You, R. *et al.* (2018). GOLabeler: Improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*.

You, R. *et al.* (2019). NetGO: improving large-scale protein function prediction with massive network information. *Nucleic Acids Research*.

Zhou, N. *et al.* (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *bioRxiv*.