BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 16 2012, pages 2137–2145
doi:10.1093/bioinformatics/bts351

Systems biology

Advance Access publication June 19, 2012

Topology of functional networks predicts physical binding of proteins

Ömer Sinan Saraç[1,*], Vera Pancaldi[2], Jürg Bähler[2] and Andreas Beyer[1,3]

[1]Biotechnology Center, Technische Universität Dresden, D-01062 Dresden, Germany, [2]Department of Genetics, Evolution & Environment and UCL Cancer Institute, University College London, London WC1E 6BT, UK and [3]Center for Regenerative Therapies Dresden, Technische Universität Dresden, D-01062 Dresden, Germany

Associate Editor: Mario Albrecht

ABSTRACT

Motivation: It has been recognized that the topology of molecular networks provides information about the certainty and nature of individual interactions. Thus, network motifs have been used for predicting missing links in biological networks and for removing false positives. However, various different measures can be inferred from the structure of a given network and their predictive power varies depending on the task at hand.

Results: Herein, we present a systematic assessment of seven different network features extracted from the topology of functional genetic networks and we quantify their ability to classify interactions into different types of physical protein associations. Using machine learning, we combine features based on network topology with non-network features and compare their importance of the classification of interactions. We demonstrate the utility of network features based on human and budding yeast networks; we show that network features can distinguish different sub-types of physical protein associations and we apply the framework to fission yeast, which has a much sparser known physical interactome than the other two species. Our analysis shows that network features are at least as predictive for the tasks we tested as non-network features. However, feature importance varies between species owing to different topological characteristics of the networks. The application to fission yeast shows that small maps of physical interactomes can be extended based on functional networks, which are often more readily available.

Availability and implementation: The R-code for computing the network features is available from www.cellularnetworks.org

Contacts: andreas.beyer@biotec.tu-dresden.de

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

Received on August 8, 2011; revised on June 8, 2012; accepted on June 11, 2012

1 INTRODUCTION

Information about physical protein–protein interactions (PPIs) is essential for a system-level understanding of complex biological mechanisms. Two of the major areas of research in that respect are the discovery and interpretation of PPIs. A set of PPIs can be represented as a network (or graph) where nodes represent proteins and edges represent physical binding proteins. The topology of molecular networks contains an extra layer of information beyond the mere representation of binary interactions, thus, topological properties and graph theoretical algorithms can be used for understanding and interpreting these networks (Przulj, 2011). Topological information is mainly used for the interpretation of PPIs, i.e. for extracting biological information. Based on network topology, proteins have been assessed for their association with a given disease (Aragues *et al.*, 2008; Vanunu *et al.*, 2010) or functionally related modules have been extracted (Hua Li and Liang 2009; Jingchun Chen and Yuan, 2006; Milenković and Przulj 2008; Przulj *et al.*, 2004).

In addition to interpretation, PPI network topology is also used for discovering new interactions. Features derived from network topology have been used for improving the quality of PPIs obtained by high-throughput screens such as yeast two-hybrid (Y2H) or co-purification methods followed by affinity-purification mass spectrometry (AP-MS). PPI networks are improved either by eliminating false-positive edges or by complementing the network by adding undetected interactions. For example, Friedel and Zimmer (2009) used minimum spanning trees (MSTs) to infer the topology of direct binding in protein complexes obtained from affinity-purification assays. Kuchaiev *et al.* (2009) took advantage of the fact that high-quality PPIs are well modeled by geometric graphs for the de-noising of PPIs. They embedded the given PPI network into a Euclidian space and predicted interactions between proteins that are mapped closer than a threshold distance in the Euclidian space. Edges in the PPI that link proteins that are distant in the embedded space were classified as false positives. Others have identified protein complexes in networks by searching for densely connected regions (Bu, 2003; Qi *et al.*, 2008; Spirin and Mirny, 2003; Vazquez *et al.*, 2003). All of these studies require physical PPI networks to be available as input, which is often not the case for many less-studied organisms. Direct measurement of physical protein binding is a resource-intensive task even though high-throughput technologies such as AP-MS and Y2H exist. However, there is a large amount of data derived from other techniques (such as co-expression or genetic interactions) that hint at functional associations between genes or proteins rather than physical protein–protein binding. As a result, predicted functional interaction networks (where 'functional interaction' is defined as pathway co-membership) are readily available for many organisms (Szklarczyk *et al.*, 2011). In this study, we investigate the possibility of using functional interaction networks to predict physical binding.

Given a functional network, it is likely that only a subset of the edges will also correspond to physical association of the proteins.

*To whom correspondence should be addressed.

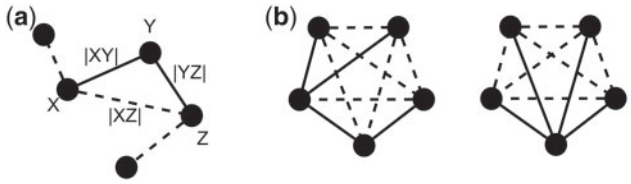

Fig. 1. Sample subgraphs from a functional network. Solid edges represent physical interactions and dashed edges represent functional-only interactions. **(a)** Proteins X, Y and Z are all participating in the same pathway (i.e. are functionally linked), but only XY and YZ are directly interacting. We observed in such situations that if we convert strength of the functional interactions to edge weights such that stronger interactions have smaller weights, the weight, |XZ| between X and Z in the functional network is often higher than |XY| and |YZ|. This motivates the shortest path ratio, which is (|XY|+|YZ|)/|XZ|. This ratio is less than 1 if the edges XY and YZ are strong (small weight) compared to XZ. **(b)** Both proteins participating in a common signaling pathway (left) as well as a protein complex represented by the spoke model (right) may lead to densely connected subgraph in the functional network. (See Supplementary analysis for a detailed explanation of matrix and spoke models for representing results of co-purification experiments)

Non-physical functional interactions in turn are often indirectly mediated through a sequence of molecular interactions, e.g. because the two proteins participate in the same pathway. Our conjecture is that the functional network is by and large shaped by the physical interactions (Fig. 1). For example, it is likely to detect a functional link between two genes that are physically interacting with a common third gene. We noticed that these indirect interactions often have lower confidence scores (in a functional network) than the direct physical interactions. This notion can be exploited for predicting the subset of functional interactions that are also physical, e.g. based on the ratio of the shortest distance between two proteins and the edge weight directly connecting the two (Fig. 1a). Likewise, densely connected sub-networks may be indicative of protein complexes, i.e. physical association. Accordingly, the density of a given sub-network has been used for predicting missing links in a physical protein interaction network (Albert and Albert, 2004; Kuchaiev *et al.*, 2009; Qi *et al.*, 2008). This situation is, however, more complicated in functional networks: closely related genes operating in a common pathway may also be densely connected in a functional network (Fig. 1b). Hence, topological features quantifying the density of edges (such as 'clustering coefficient') may have poor predictive power when used in isolation in functional networks. Herein, we show that combining topological features with other lines of evidence improves the specificity when predicting physical protein binding from functional networks.

We set out to test the effectiveness of combining various network topological features for predicting the interaction type represented by a given edge in a functional network. Specifically, the goal is to infer the subset of interactions from a functional network that also corresponds to physical associations of the respective proteins. To achieve that we combined network topology-based features with features that are used to predict the functional association in the first place. We extracted seven topological features for each edge in an existing functional interaction network. These features describe the topological characteristics of an edge such as its centrality, node neighborhoods and connection strength relative to alternative paths. We complemented these network features with the features that were used to predict that functional edge and trained a machine learning classifier to distinguish physical from purely functional interactions. Using such a machine learning approach, we can assess the importance of network features compared to other types of features, and we can quantify the relative importance of individual network features for predicting specific types of physical binding events.

We defined three classes of interactions in a functional network, namely, 'binary', 'co-complex' and 'functional-only'. Given a scaffold network of functional interactions, the task is to assign each edge to one of these three classes. The first two classes represent physical association of interacting proteins, while the last one denotes functionally related proteins (or genes) that do not bind. The distinction between binary and co-complex interactions is operational and based on the available experimental data (see Section 2 for more details). The binary class is mainly composed of directly binding protein pairs, while the co-complex class contains sets of proteins that are purified together and likely are part of a common complex. However, we emphasize that these are not strict classifications, because the experimental data do not allow to rigorously and unambiguously separate these two classes. For each edge in the functional network, we calculated a set of network features extracted from the topology of the network. Then we used a supervised machine learning approach to predict and distinguish the three types of interactions. We showed the effectiveness of our approach on human and budding yeast networks. The analysis of these networks showed the value of individual network features, revealing that the predictive power of the features is dependent on the species and the available data. Finally, we applied our method to predict 2853 novel physical PPIs in fission yeast (*Schizosaccharomyces pombe*), which is an important model organism with so far few endogenously measured protein interactions.

2 METHODS

2.1 Datasets

For each of the three organisms, the functional interaction network is acquired from the STRING database (version 8.3) (Jensen *et al.*, 2009). STRING combines seven different feature types for predicting functional interactions between genes. The features used by STRING are 'genomic neighborhood, gene fusion, phylogenetic profile (co-occurence), co-expression, text mining, experimental evidence and database' evidence. Using these features, STRING calculates a combined score that represents the confidence that a functional association exists. We re-calculated the STRING combined score without using the experimental and database features in order to avoid circular reasoning (since these lines of evidence might be based on the same data as our training set). A similar concern may be raised with respect to text mining since it may rest on the same publications used for generating the training sets. Therefore, we re-computed the performance scoring without text mining and showed that text mining does not inflate the performance scores due to circular reasoning (see Supplementary Fig. S1). Re-calculation of the combined STRING score was done according to the original scheme used for STRING:

$$S = 1 - \prod (1 - S_i),$$

where S_i is the individual evidence expressed as probability of being true and S is the integrated score. STRING only reports interactions with confidence scores S above 0.15. We followed the same scheme and converted scores for the remaining interactions into edge weights $E = -\log(S)$. Thus, when computing network distances, nodes connected by a high-confidence edge (large S) will have a small weight (small E).

We defined three different classes of interactions: binary, co-complex and functional-only. In the binary class, we tried to capture more transient interactions such as phosphorylation or ubiquitination. The co-complex class contains protein pairs from stable complexes. Finally, the functional-only class represents functionally related proteins that are not physically interacting. Binary and co-complex classes together constitute the physical interactions. The numbers of interactions in each class and in the modified STRING network are listed in Table 1.

For budding yeast, we used physical interaction data from BioGRID (Stark *et al.*, 2011). Interactions annotated as 'Biochemical Activity' or 'Two-hybrid' are used as binary interactions, where interactions captured by 'Co-purification', 'Affinity Capture-Luminescence', 'Affinity Capture-MS' and 'Affinity Capture-Western' were classified as co-complex interactions. If an interaction was detected by both types of methods (e.g. 'Two-hybrid' and 'Co-purification'), we treated it as a co-complex interaction. We also included binary interactions defined in KEGG pathways (Kanehisa *et al.*, 2010). We collected physical interactions from IntAct (Aranda *et al.*, 2010) as an external validation set. We removed all interactions overlapping with the training data from this validation set. For human, binary and co-complex interactions were generated similarly with the exception that complexes reported in CORUM (Ruepp *et al.*, 2010) are also incorporated into the co-complex class using the matrix model (i.e. all pairwise interactions). We used high-confidence interactions reported in a previous analysis (Bossi and Lehner, 2009) (termed CRGhigh) as an independent validation set for human after removing all interactions overlapping with our training data.

In fission yeast, intersection of the BioGRID with STRING resulted in only 659 interactions that could be used as training data. In order to increase this number, we incorporated interactions from other databases, namely, DIP, MINT, IntAct and KEGG (Ceol *et al.*, 2010; Xenarios *et al.*, 2000). This approach resulted in 1161 distinct interactions in total that could be used for training. Information about the experimental system or the nature of the interactions is not available for many of these interactions. Thus, we only performed a two-class (physical versus non-physical) classification for fission yeast. Additionally, we mapped a set of interactions from budding yeast using orthology, i.e. 'interologs' (Matthews *et al.*, 2001). We only mapped interologs for budding yeast proteins with unique orthologs in fission yeast. We used these interactions as an external validation set for fission yeast.

For all three organisms, interactions in the functional-only class were generated using KEGG pathways. This functional-only set is composed of gene pairs that are in the same KEGG pathway but whose protein products are not known to be physically interacting. This set forms the negative training and test set for the prediction of physical interactions in the STRING functional network. A general approach for generating a negative set is to generate random pairs of genes and sampling from these pairs. As the number of physical interactions is small compared to all possible pairings in a genome, this sampled negative set is expected to be highly enriched for non-interacting pairs thus containing a small number of false negatives. This approach, however, fails in our case, because we need a negative reference set of links in the functional network. All proteins connected in a functional network are *per se* more likely to also physically interact. We therefore chose to define the negative (functional-only) set using well-studied pathways. Since these proteins are better studied, it is less likely that there are unknown physical interactions (false negatives) among those protein pairs.

2.2 Network features

For each edge of a given graph, we defined seven features using topological properties of the functional interaction graph.

2.2.1 Minimum spanning tree A spanning tree of a graph is defined as a subgraph, which is a tree that connects all vertices of the graph. A MST is the spanning tree where the sum of its edge weights is smaller than or equal to the sum of the weights of all other spanning trees. After calculating a MST, we defined this binary feature to be 1 if the corresponding edge is on the MST and 0 if not (Friedel and Zimmer, 2009). It is possible to have more than one MST in a given network but we only consider the first one for simplicity. We compared the union of all possible MSTs with the first MST. In case of the human network, we found that the union contained only 486 edges more than the first MST (the human network has 14 063 nodes (genes) and 924 616 edges). Hence, the MSTs are very similar. Anyway, the extended MST (see Section 2.2.2) is a superset of all possible MSTs.

2.2.2 Extended minimum spanning tree For a graph with n vertices, a MST may have at most $n-1$ edges. This is sparser than expected in protein interaction networks (Friedel and Zimmer, 2009). To extend the number of edges in the MST, we considered each edge of the graph that was not already in the MST. We took the nodes connected by the edge in consideration and calculated the shortest path (a path between two vertices where the sum of the weights of the edges on the path is minimum) between them in the MST. If the sum of the weights on this shortest path (the distance of the nodes) in the MST is larger than the weight of the direct edge itself, then the edge is added to the extended MST. After calculating the extended MST, we again generated a binary feature as we did for the MST. Note that this extending strategy guarantees that any edge on any possible MST will be added to the extended MST.

Since we used the STRING combined score to obtain the edge weights of the graph, edges in the MST and extended MST represent a backbone of the functional network that is connected with high scoring functional interactions. MST-based features have previously been used to identify direct physical binding in protein complexes identified by affinity purification methods (Friedel and Zimmer, 2009).

2.2.3 Clustering coefficient In graph theory, a clustering coefficient is a measure of the edge density within a subgraph. First, we calculated the local clustering coefficient for each vertex, which is defined as the ratio of the number of observed edges between the neighbors of a vertex to the number of all possible edges between the neighbors. If the neighbors of a vertex form a clique (a subset of fully interconnected nodes), then this ratio becomes 1. We defined the clustering coefficient feature of an edge as the average of the local clustering coefficients of the vertices connected by that edge. Qi *et al.* (2008) used clustering coefficients of vertices to identify protein complexes in an interaction network.

2.2.4 Neighborhood ratio The neighborhood ratio (also known as the Jaccard index) of an edge that connects vertices v_i and v_j is the ratio of the number of common neighbors of v_i and v_j to the total number of all neighbors of v_i and v_j. A similar scoring has been used to improve functional networks in budding yeast (Lee *et al.*, 2004).

2.2.5 Global betweenness Global betweenness of an edge is defined as the number of shortest paths passing through the edge normalized by the number of all possible shortest paths. This is a measure of centrality of the edge in the graph. Global betweenness of edges has been used to detect community structures in networks (Newman, 2004).

2.2.6 Local betweenness Local betweenness is same as global betweenness, but defined for the local neighborhood of the edge. We took the subgraph that consists of only the immediate neighbors of the vertices connected by the query edge.

A high global betweenness score indicates more centrality in the overall network, while local betweenness is less dependent on the position of the edge in the global network. When calculating both types of betweenness scores, the graph was assumed to be un-weighted. When we calculated the betweenness based on weighted edges, betweenness scores became very sparse. For human, only 26% of the edges had non-zero local betweenness and the situation was even worse for the global betweenness with only 7% non-zero scores. This is due to a small set of very strong functional edges traversing most shortest paths even though they often introduced a considerable number of 'hops'. A correlation analysis showed that the

Table 1. Dataset sizes and database resources used in the study

Species	Functional network	Binary interactions	Co-complex interactions	Functional-only interactions	External validation set
Human	(STRING modified)	(BioGRID, KEGG)	(BioGRID, CORUM)	(KEGG)	(CRGhigh)
	14 063 genes	2624 genes	4884 genes	4470 genes	1392 genes
	924 616 interactions	6484 interactions	16 317 interactions	69 611 interactions	1239 interactions
Budding Yeast	(STRING modified)	(BioGRID, KEGG)	(BioGRID)	(KEGG)	(IntAct)
	5176 genes	1488 genes	2622 genes	1320 genes	1320 genes
	146 685 interactions	2061 interactions	10 050 interactions	11 355 interactions	1626 interactions
			Physical interactions		
Fission yeast	(STRING modified)	(BioGRID, KEGG, DIP, Mint, IntAct)		(KEGG)	(Orthologs from budding yeast)
	3286 genes	906 genes		853 Genes	1208 genes
	60 874 interactions	1161 interactions		3760 interactions	3117 interactions

Functional network is the remaining STRING network (STRING modified) after discarding the 'experimental' and 'database' evidence channels from the original STRING network. Rest of the set of interactions is the intersection of available interactions in the database resources with the modified STRING network. Interactions that exist in the binary or co-complex sets (physical interactions) are removed from the external validation sets.

weighted betweenness scores were highly correlated with the 'extended MST', 'shortest path ratio' and the 'combined STRING score', all of which are reflecting connection strength of the functional associations (Supplementary Fig. S2). Thus, there was also considerable redundancy between the weighted betweenness and other features, leading us to only consider un-weighted betweenness scores.

2.2.7 Shortest path to edge weight ratio For a given edge, this is the ratio of the shortest path between the vertices connected by the edge to the weight of the edge itself. Note that this feature assumes the maximum value of 1 if the shortest path is the edge itself. This feature penalizes very weak links connecting two vertices that are also indirectly connected through other, much stronger edges. In such case, the ratio will be significantly below 1.

2.3 STRING features

Five features were taken from the STRING database (version 8.3): genomic neighborhood, gene fusion, phylogenetic profile, co-expression and text mining in addition to the re-calculated combined score. 'Genomic neighborhood' indicates whether the two genes are closely located in genomes. 'Gene fusion' evidence shows fusion events in orthologous genes. 'Phylogenetic profile' shows evidence of co-occurrence of the gene pair in different organisms. 'Co-expression' indicates predicted association between genes based on observed patterns of simultaneous expression of genes. Finally, STRING parses a large body of publications for statistically significant co-occurrence of gene names and incorporates it as text mining evidence.

2.4 Training and testing

We trained Random Forest (Breiman, 2001) for three-class classification (binary, co-complex and functional-only) using the 'randomForest' package from R (Liaw and Wiener, 2002). To avoid a bias toward larger classes, we randomly sampled from the larger sets to obtain training sets of approximately even size. For a given edge to be predicted, Random Forest returns probabilities for each class that add up to 1. Probability of being physical is calculated as the sum of predicted probabilities of binary and co-complex classes. Binary versus co-complex classification is only performed for the interactions where the probability of being physical is above 0.5.

For all three organisms studied, we performed two types of validations. First, we performed 5-fold cross-validation. At each step, 1-fold of the data is left aside for testing and the remaining folds are used for training. A Random Forest with 1000 trees was trained for each fold. Second, we assessed the performance of the model with an external validation set. In this case, external data were used as positive (physical) test and an equal sized random sample from the functional-only set that has not been used in training was used as the negative test set. We filtered the external validation test set so that it does not contain any sample from the training set.

We used receiver operator characteristics (ROCs) and precision–recall (PR) curves to assess the classification performance. The area under ROC curve (AUC) is a popular threshold-independent measure for model comparison. It directly reflects the difference in comparison to a random ranking (Davis and Goadrich, 2006). However, in real-life applications, the precision of the highest ranking predictions is of greater relevance, which is better visualized by PR curves. We therefore decided to use both means for assessing the performance of the models.

3 RESULTS AND DISCUSSION

3.1 Network features significantly improve prediction of physical interactions

We used the human, budding yeast and fission yeast networks from STRING version 8.3 (Jensen *et al.*, 2009) as the initial functional networks for predicting physical interactions. After discarding 'experimental and database' evidences from STRING (to avoid re-using evidence on which our reference set is based), we constructed a functional interaction graph for each species. We assigned weights to edges in the graph such that stronger functional interactions get lower weights. We derived seven network topological features from this graph (Section 2) and assessed their utility for distinguishing physical from non-physical interactions.

Initially, we restricted the analysis to human and budding yeast, for which more interactions have been measured than for fission yeast. For these two species, we obtained experimentally validated interactions for all three classes from BioGRID, KEGG and CORUM (Kanehisa *et al.*, 2010; Ruepp *et al.*, 2010; Stark *et al.*, 2011), which we used subsequently for supervised learning and testing.

We trained multi-class Random Forest classifiers (Breiman, 2001) using network features only, original STRING evidences only and the combination of both. We decided to use Random Forest because it is virtually assumption free, very adaptable and has been shown to outperform other machine learning methods for similar tasks (Elefsinioti *et al.*, 2011; Qi *et al.*, 2006, 2009). Herein, we compare Random Forest to Support Vector Machines (Burges,

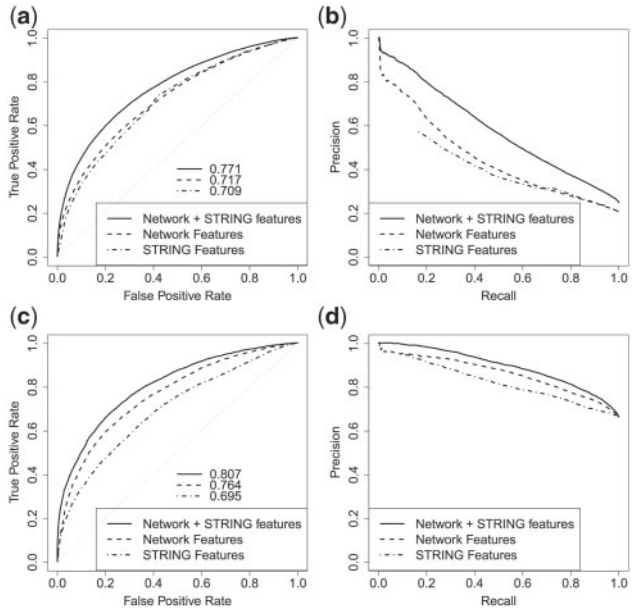

Fig. 2. Contribution of network features to the prediction of physical interactions. ROC curves (left) and PR curves (right) for physical versus functional-only classification after 5-fold cross-validation in **(a** and **b)** human and **(c** and **d)** budding yeast. AUC is shown in the panels

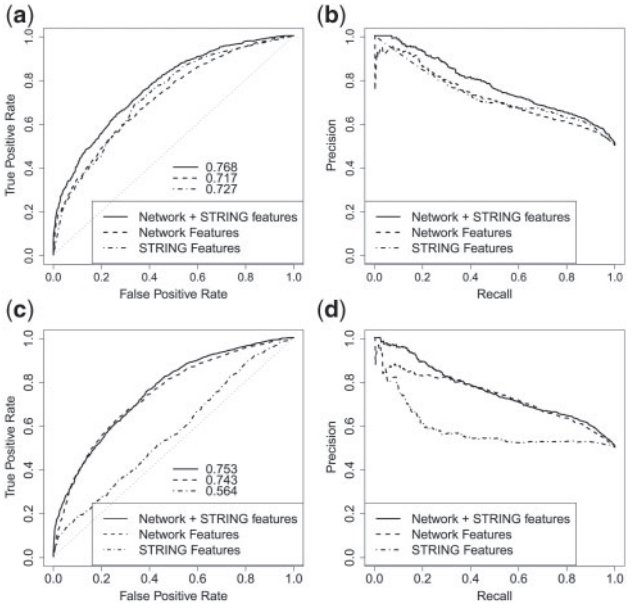

Fig. 3. ROC curves (left) and PR curves (right) for validation with the external data. (**a** and **b**) Validation of the human model using CRGhigh. (**c** and **d**) Validation of the budding yeast model using IntAct. AUC is shown in the panels, respectively

1998) using default parameters and the radial basis function as kernel. The comparison of the different machine learning methods on the human network confirmed the high performance of Random Forest (Supplementary Fig. S3). Subsequently, we distinguished network features (i.e. features derived from the topology of the functional network) and 'STRING features'. This latter class of features encompasses those that have been used for constructing the functional network including the integrated evidence score (Combined Score). Figure 2 shows ROC curves and PR curves for physical versus functional-only classification after 5-fold cross-validation. Using the network features alone, we could better distinguish physical from functional-only interactions compared to using the STRING features alone. This analysis shows that network features are useful for predicting physical binding and that the topological structure of the functional network also contains important information about the physical interactome. In both organisms, combining network features with STRING features increased the performance, indicating that network topology provides complementary information that is not directly available in the features used for creating the functional network. The improvement is more pronounced for high confidence interactions, which are actually more important for predicting new physical interactions.

If a data source contains groups of similar samples, it is possible that these similar samples scatter into train and test folds, which may cause an overestimation of the performance based on cross-validation. In order to overcome this problem, we also validated our models by using test data obtained from two other sources that were not used for training (Aranda *et al.*, 2010; Bossi and Lehner, 2009). Interactions that are already in our training data were excluded from this analysis. This type of validation is only performed for physical versus non-physical classifications, because we did not have sufficient information about the types of experiments to sub-classify the measurements into binary and co-complex. A random set of samples from the functional-only set was used as a negative test set. In human, CRGhigh was used as the external validation set (see Section 2.1). Physical interactions extracted from the IntAct database were used as external validation data for budding yeast. This test gave similar results as the cross-validation for both human and budding yeast (Fig. 3). Network features generalize better than STRING features, and the difference is much more striking in budding yeast. In both species, combining network and STRING features improves the performance. These results indicate that using network features in combination with machine learning is not simply over-fitting to the topological structure of the training set, but it captures general information about the physical associations underlying the functional network.

We also checked if the interactions predicted with the help of network features are biased for proteins with specific functions. We performed Gene Ontology (GO) (Ashburner *et al.*, 2000) enrichment analysis in human for proteins involved in interactions that are only predicted when using network features. We selected all interactions that can be predicted with high confidence (probability of being physical >0.8) and used the proteins that are involved in these interactions as background. Then we selected the subset of these interactions that could not be predicted with STRING features alone (Random Forest score <0.3). Prediction of these interactions crucially relies on the inclusion of network features in the model. We used 'GO Term Finder' (Boyle *et al.*, 2004) for the enrichment analysis and observed an enrichment for 195 GO molecular function terms representing very diverse functions in the ontology (Supplementary Table S1). Thus, we do not have evidence that network features introduce a bias for specific types of proteins. Node degree and pleiotropy (multi-functionality) of

genes are highly correlated, which can create substantial biases when predicting the function of genes using network information (Gillis and Pavlidis, 2011). Thus, any highly connected set of proteins is biased toward being enriched for many GO terms. Network features are most useful for classifying interactions in densely connected regions of the functional network, which explains the large number of very heterogeneous GO terms being enriched among those genes.

3.2 Feature importance is network dependent

We also investigated the importance of features for distinguishing each class. Random Forest provides a way of measuring the importance of individual features for the model by permuting the values of the corresponding feature among samples and measuring the mean decrease in classification accuracy. If a single feature is highly important for the model, permutation of that feature results in a large decrease in accuracy. Figure 4 shows the feature importance values calculated by Random Forest for human and budding yeast, separately for each class, in addition to the overall importance. Both, STRING features and network features are important for the successful classification of interactions. An important result of our analysis is the variable importance of features when comparing different species. For example, both the 'global' and 'local betweenness' scores are important for functional-only interactions for budding yeast. In contrast, the global betweenness in human is the most important feature for the binary class. Similarly, 'co-expression' is most important for co-complex interactions in human, yet it is the least important for co-complex interactions in human. We suspect that these differences reflect differences in data availability and experimental biases rather than fundamental biological differences. Budding yeast has been subjected to extensive testing of protein binding using AP-MS technologies (Gavin *et al.*, 2002, 2006; Krogan *et al.*, 2006). Such experiments have not been conducted at a comparable scale (let alone comparable coverage) in human cells. Thus, the budding yeast network is strongly enriched for densely connected subnetworks compared to the human network. Such differences will evidently affect the prediction strategy chosen by the machine learning algorithm. There are also similarities. In both organisms, text mining, combined score and 'shortest path ratio' are more important for the functional-only class than for the binary and co-complex classes (Fig. 4), suggesting that these features are critical for distinguishing functional-only interactions from physical association.

One drawback of the permutation importance in Random Forest is that in the case of correlated predictors, importance of one of these individual predictors may be underestimated due to the compensation from the other (Nicodemus *et al.*, 2010). We also calculated the Gini importance, which measures how well a given feature splits the data into homogenous partitions. Gini importance integrates the decrease in Gini impurity over all trees of the forest whenever the corresponding feature is used for splitting (Breiman, 2001). Although it does not directly indicate importance of the feature in the final model, it is less affected by correlation with other features and gives a more direct evaluation of the distinguishing power of the feature. With respect to Gini importance, network features are in general more important than STRING features (Supplementary Fig. S4).

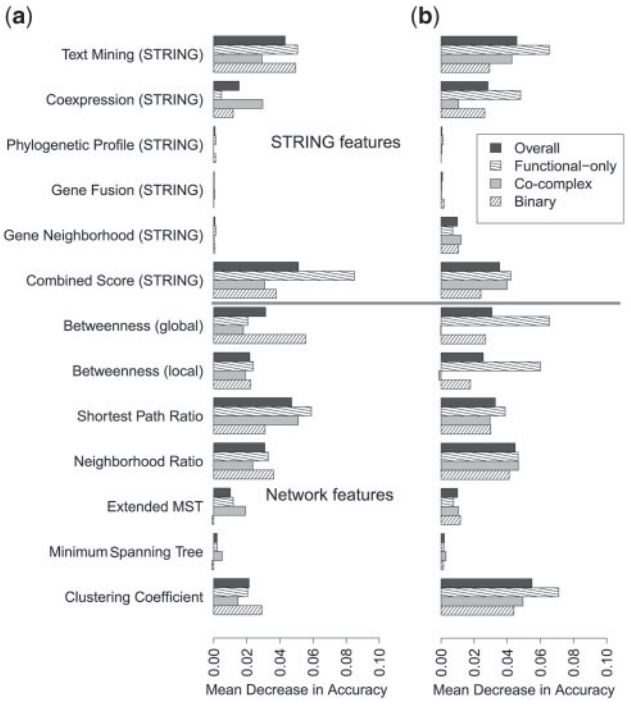

Fig. 4. Random Forest feature importance (Mean Decrease in Accuracy) for **(a)** human and **(b)** budding yeast. Feature importance is shown for the entire classification task (Overall) and for predicting specific types of interactions ('Binary', 'Complex' and 'Functional-only', respectively). STRING features were taken from the STRING database as is. The combined score was recalculated excluding experimental evidences (see main text for details)

Random Forest feature importance values do not show direct correlation between the features and the classes. They only indicate the importance of a feature for the non-linear model (i.e. conditional on all other features used). In order to quantify the relationship between individual features and the response variables (class assignment), we performed *t*-tests for individual features indicating how much a given feature separates the different interaction classes (Fig. 5). According to the *t*-test, most features show similar discriminative power in human and yeast with respect to distinguishing functional from physical interactions. Two notable exceptions are clustering coefficient and co-expression, which perform very differently in the two species. Co-expression separates physical from functional interactions only in human (Fig. 5a) but not in budding yeast (Fig. 5b). This difference might be due to the simpler transcriptional regulatory program in yeast. Hence, budding yeast genes participating in the same pathway might be co-expressed even if they do not bind directly, whereas this is less likely to happen in mammals (Zhang *et al.*, 2004). Interestingly, co-expression is important for correctly classifying functional interactions in the yeast Random Forest model when combined with the other features (Fig. 4b). This observation underlines the complexity of relationships between features being exploited by Random Forest. MST and extended MST are associated with physical interactions (Fig. 5), yet they are not very important for the Random Forest model (Fig. 4). This is mainly because these features are sparse, i.e. they provide information only for a relatively small number of edges. Shortest path ratio is strongly correlated with MST and extended

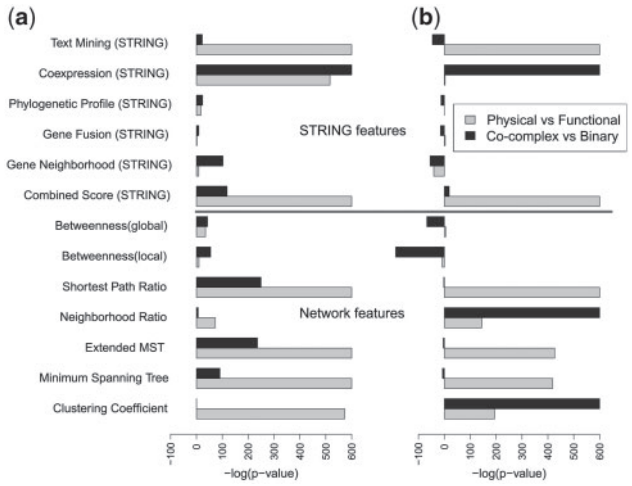

Fig. 5. Correlation between features and classes for **(a)** human and **(b)** budding yeast. Negative log-*P*-values are shown for *t*-tests comparing the mean values of each feature between the two classes (physical versus non-physical or complex versus binary, respectively). Positive bars indicate features that are higher in the first class; negative bars indicate features that are higher in the second class. High absolute values indicate that the respective feature is directly predictive for the classification. However, the *t*-test ignores possible interactions between features that can be exploited by Random Forest

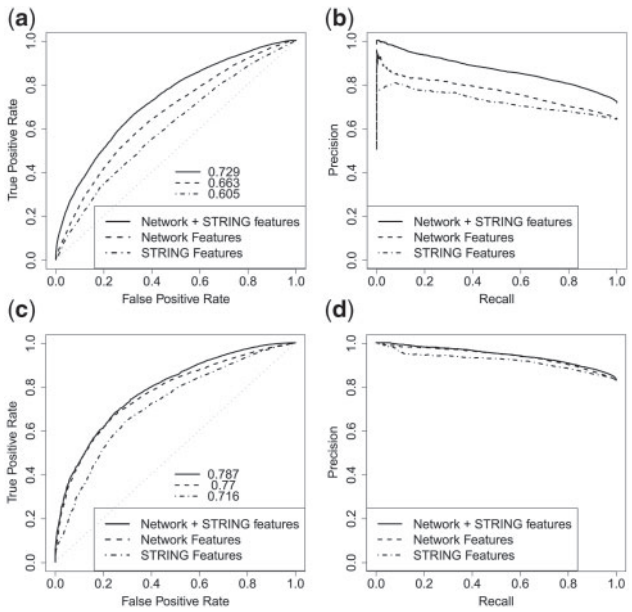

Fig. 6. ROC curves (left) and PR curves (right) for complex versus binary classification after 5-fold cross-validation in (**a** and **b**) human and (**c** and **d**) budding yeast

MST. The MST-based features may seem less relevant because of this redundancy. However, removing shortest path ratio from the model had no significant effect on the importance scores of MST and extended MST (Supplementary Fig. S5).

3.3 Distinguishing different types of physical interactions

Next, we investigated in more detail how well we can separate the two types of physical interactions (co-complex versus binary interactions). Figure 6 shows the ROC curves and PR curves for co-complex versus binary predictions. The AUC is higher in both species when using network features alone than using STRING features alone. Thus, network topology can be used effectively for distinguishing different types of interactions. As shown before, the importance of features is mostly network dependent (Fig. 4). Relative importances of features with respect to binary and co-complex classes vary between human and budding yeast. For example, in human the two betweenness features are important for predicting both co-complex and binary interactions, whereas in budding yeast it is irrelevant for predicting co-complex interactions, but important for the classification of binary interactions.

A complication for the prediction of co-complex interactions is the fact that AP-MS measurements usually do not infer the topology of protein complexes, i.e. the structure of direct binding events remains elusive. Protein complexes are often identified by these experiments such that a single protein (bait) is 'pulled down' (purified) using an antibody together with all the proteins (preys) belonging to the same multi-protein complex. The actual topology of the binding of proteins in the complex is not revealed by such an experiment. Thus, sets of commonly associated proteins have been turned into 'networks' using either the spoke or matrix model (Bader and Hogue, 2002). In the matrix model, all possible pairwise interactions among the members of the complex (bait and preys) are assumed. In the spoke model, interactions are only assigned between the bait and each prey. Both models have shortcomings. However, we can safely assume that protein complexes can be represented as densely connected sub-networks. Random Forest was capable of detecting these sub-networks while exploiting their distinct characteristics with respect to the network features that we inferred. We investigated to which extent our predictions are affected by representing protein complexes using the spoke versus matrix models. BioGRID does not provide details about the representation of protein complexes. Thus, we limited this analysis to human complexes from CORUM, where protein complex co-membership is defined. From each CORUM complex, we randomly chose a protein as a 'bait protein' and then simulated both the spoke and matrix models. It turns out that the importance of network features is not much affected by the choice of the model (see Supplementary Analysis).

3.4 Prediction of novel physical interactions in fission yeast

The original motivation for this project was to reliably predict physical interactomes in species where only a functional network is available. The fission yeast is an important model species for which various high-throughput functional and genetic screens have been performed (Kapitzky *et al.*, 2010; Pancaldi *et al.*, 2010; Roguev *et al.*, 2008). However, a comprehensive physical interactome is still lacking. Hence, we applied our method to predict novel physical interactions in fission yeast after training on high-confidence physical interactions from BioGRID, DIP, MINT, IntAct and KEGG databases (Aranda *et al.*, 2010; Ceol *et al.*, 2010; Kanehisa *et al.*,

2010; Stark *et al.*, 2011; Xenarios *et al.*, 2000). We trained a two-class classifier for physical versus non-physical interactions, because we did not have enough information on the training data for a more fine-grained classification. We estimated the performance of the classifier using 5-fold cross-validation (Supplementary Fig. S6). We selected a threshold that gives a precision of 0.8 on cross-validation, yielding 2853 novel predicted interactions that are not in the positive training set (Supplementary Table 2). The main source of evidence for these interactions in STRING is text mining. The majority (2851) of these interactions have text mining evidence, but only 246 have another type of STRING evidence.

To validate the model further, we used a set of interactions mapped from budding yeast using orthology, i.e. 'interologs' (Matthews *et al*., 2001). We only mapped interologs for budding yeast proteins with unique orthologs in fission yeast, resulting in 2586 mapped interactions that are not present in the previous reference set. The Random Forest model successfully predicted 61% of these interactions at a Random Forest probability threshold of 0.5. Supplementary Figure S7 shows that predictions are enriched for interactions mapped from budding yeast.

4 CONCLUSIONS

We presented a method to predict physical PPIs based on a functional interaction network. Hence, the goal was not the *de novo* prediction of interactions. Rather, links existing in the STRING network were classified as being physical or functional, etc. We showed that features extracted from the topological properties of the functional network can be used to predict physical binding, and we systematically tested the importance of each of the seven features for the prediction task. Furthermore, we have demonstrated that network features help distinguishing different types of interactions. Obviously, this approach can be extended in the future to distinguish even more different types of interactions (such as regulatory versus structural), and it may be applicable to other types of networks (e.g. transcriptional regulatory networks, micro RNA networks or genetic networks).

We are aware of the shortcomings of our classification in 'stable' protein complexes and 'transient' binary interactions. Complexes are obviously dynamic and binary interactions can be stable. Nevertheless, the Random Forest classifier could distinguish these classes, which we operationally defined based on the experimental techniques used. Our analysis suggests that these classes are at least enriched for different types of interactions.

Finally, the application to fission yeast demonstrated that the strategy outlined here can in fact be used to complement physical interactomes in less studied model organisms. The latest version of STRING generated functional networks for about 1100 fully sequenced organisms (Szklarczyk *et al.*, 2011). Thus, there is a great potential for predicting physical interactions based on the structure of these networks.

Funding: The EC FP7 projects PhenOxiGEn (HEALTH-F4-2008-223539), SyBoSS (FP7-HEALTH-2009-242129) and the Klaus Tschira Foundation.

Conflict of Interest: none declared.

REFERENCES

Albert,I. and Albert,R. (2004) Conserved network motifs allow protein–protein interaction prediction. *Bioinformatics (Oxford, England)*, **20**, 3346–3352.

Aragues,R. *et al.* (2008) Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics*, **9**, 172.

Aranda,B. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bader,G.D. and Hogue,C.W.V. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.

Bossi,A. and Lehner,B. (2009) Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.*, **5**, 260.

Boyle,E.I. *et al.* (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)*, **20**, 3710–3715.

Breiman,L. (2001) Random forests. *Machine Learn.*, **45**, 5–32.

Bu,D. (2003) Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Res.*, **31**, 2443–2450.

Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discov.*, **2**, 121–167.

Ceol,A. *et al.* (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.

Chen,J. and Yuan,B. (2006) Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics (Oxford, England)*, **22**, 2283–2290.

Davis,J. and Goadrich,M. (2006) The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning–ICML '06*. ACM Press, New York, NY, pp. 233–240.

Elefsinioti,A. *et al*. (2011) Large-scale de novo prediction of physical protein–protein Association. *Mol. Cell. Proteom. MCP*, **10**, M111.010629.

Friedel,C.C. and Zimmer,R. (2009) Identifying the topology of protein complexes from affinity purification assays. *Bioinformatics (Oxford, England)*, **25**, 2140–2146.

Gavin,A.-C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gavin,A.-C. *et al*. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

Gillis,J. and Pavlidis,P. (2011) The impact of multifunctional genes on 'guilt by association' analysis. *PLoS One*, **6**, e17258.

Jensen,L.J. *et al.* (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Kapitzky,L. *et al.* (2010) Cross-species chemogenomic profiling reveals evolutionarily conserved drug mode of action. *Mol. Syst. Biol.*, **6**, 451.

Krogan,N.J. *et al*. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, **440**, 637–643.

Kuchaiev,O. *et al.* (2009) Geometric de-noising of protein–protein interaction networks. *PLoS Comput. Biol.*, **5**, e1000454.

Lee,I. *et al.* (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555.

Li,H. and Liang,S. (2009) Local network topology in human protein interaction data predicts functional association. *PLoS One*, **4**, e6410.

Liaw,A. and Wiener,M. (2002) Classification and regression by random forest. *R News*, **2**, 18–22.

Matthews,L.R. *et al*. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or 'interologs'. *Genome Res.*, **11**, 2120–2126.

Milenković,T. and Przulj,N. (2008) Uncovering biological network function via graphlet degree signatures. *Cancer Inform.*, **6**, 257–273.

Newman,M.E.J. (2004) Detecting community structure in networks. *Eur. Phys. J. B—Condensed Matter*, **38**, 321–330.

Nicodemus,K.K. *et al.* (2010) The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, **11**, 110.

Pancaldi,V. *et al.* (2010) Meta-analysis of genome regulation and expression variability across hundreds of environmental and genetic perturbations in fission yeast. *Mol. Biosyst.*, **6**, 543–552.

Przulj,N. (2011) Protein–protein interactions: making sense of networks via graph-theoretic modeling. *BioEssays*, **33**, 115–123.

Przulj,N. *et al.* (2004) Functional topology in a network of protein interactions. *Bioinformatics (Oxford, England)*, **20**, 340–348.

Qi,Y. *et al.* (2006) Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, **63**, 490–500.

Qi,Y. *et al.* (2008) Protein complex identification by supervised graph local clustering. *Bioinformatics (Oxford, England)*, **24**, i250–i258.

Qi,Y. *et al.* (2009) Systematic prediction of human membrane receptor interactions. *Proteomics*, **9**, 5243–5255.

Roguev,A. *et al.* (2008) Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science (New York, NY)*, **322**, 405–410.

Ruepp,A. *et al.* (2010) CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.*, **38**, D497–D501.

Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA.*, **100**, 12123–12128.

Stark,C. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.,* **39**, D698–704.

Szklarczyk,D. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.

Vanunu,O. *et al.* (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.

Vazquez,A. *et al.* (2003) Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol.*, **21**, 697–700.

Xenarios,I. *et al.* (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.*, **28**, 289–291.

Zhang,W. *et al.* (2004) The functional landscape of mouse gene expression. *J. Biol.*, **3**, 21.