BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Cluster-based assessment of protein-protein interaction confidence

Atanas Kamburov[1*], Arndt Grossmann[2], Ralf Herwig[1] and Ulrich Stelzl[2*]

Abstract

Background: Protein-protein interaction networks are key to a systems-level understanding of cellular biology. However, interaction data can contain a considerable fraction of false positives. Several methods have been proposed to assess the confidence of individual interactions. Most of them require the integration of additional data like protein expression and interaction homology information. While being certainly useful, such additional data are not always available and may introduce additional bias and ambiguity.

Results: We propose a novel, network topology based interaction confidence assessment method called CAPPIC (cluster-based assessment of protein-protein interaction confidence). It exploits the network's inherent modular architecture for assessing the confidence of individual interactions. Our method determines algorithmic parameters intrinsically and does not require any parameter input or reference sets for confidence scoring.

Conclusions: On the basis of five yeast and two human physical interactome maps inferred using different techniques, we show that CAPPIC reliably assesses interaction confidence and its performance compares well to other approaches that are also based on network topology. The confidence score correlates with the agreement in localization and biological process annotations of interacting proteins. Moreover, it corroborates experimental evidence of physical interactions. Our method is not limited to physical interactome maps as we exemplify with a large yeast genetic interaction network. An implementation of CAPPIC is available at http://intscore.molgen.mpg.de.

Background

Accurate interaction networks (interactomes) are fundamental to answering questions about how the biochemical machinery of cells organizes matter, processes information, and carries out transformations to perform specific functions leading to various phenotypes. Toward this goal, a number of experimental [1] and computational [2-4] techniques have been devised and applied to map the interactions of human proteins [5-8] and those of model organisms such as yeast [9-12]. Despite their incompleteness [13], current interactome maps already serve as a basis for numerous methods aiming to elucidate biological processes in health and disease [14,15]. Current interactome maps are contaminated with false positive interactions that can make up a considerable portion of the data [13,16-20]. These false positive interactions dim the explanatory light of interaction networks and also decrease the predictive value of methods using such data. It is thus of primary importance to derive confidence values for individual interactions, which can serve to refine current interactome maps or can be used as interaction weights. For example, it has been shown recently that the performance of complex detection approaches is better in confidence-weighted protein-protein interaction networks than in non-weighted networks [19,21].

Several approaches have been proposed for interaction confidence assessment, many of which are reviewed in [19,22,23]. Most of these methods integrate additional data like interaction homology [17], co-expression of genes encoding interacting proteins [17,24,25], or a combination of these and other evidence features [26,27]. The outcome from such methods depends on the additional data sets. Others combine multiple topological features with additional knowledge to achieve better predictions [20,28]. Methods which are able to use network topology

*Correspondence: kamburov@molgen.mpg.de; stelzl@molgen.mpg.de
[1] Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany
[2] Otto-Warburg Laboratory, Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany

alone to predict interaction veracity [29-32] are the tools of choice for interaction confidence assessment if other types of data are limited or biased.

At various levels (globally as well as locally), the topology of interaction networks encodes biological properties which are largely independent of the biochemical function of the individual members of the network [33,34]. This has been demonstrated through analysis of global properties exploiting topological features such as node degree [35] or distance [36,37]. The biological importance of network topology may be even more clear for local structures, as in the case of specific wiring patterns of interaction partners [34]. Likewise, modularity of interaction networks is currently the most successful concept for addressing the dynamics of cellular processes [8,38,39].

Goldberg and Roth [29] proposed a connectivity based approach for interaction confidence assessment where the number of common neighbors of a pair of predicted interaction partners counts in support of the interaction. They defined interaction confidence as the level of enrichment of common network neighbors of interacting proteins. It is quantified by the hypergeometric distribution P-value given the number of common neighbors and total network neighbors of both interacting proteins. The underlying principle of the approach has been established in seminal studies demonstrating that biological networks are marked with short interaction paths separating random pairs of proteins in the network (small-world property), and densely connected local neighborhoods (neighborhood cohesiveness property) [40]. Real protein-protein interactions are expected to meet the network cohesiveness property more frequently than false positives. More recently, Kuchaiev and co-authors [32] proposed another method that embeds interaction networks into a low-dimensional Euclidean space based on network metrics (shortest path length) and then calculates confidence of interactions depending on the Euclidean distance between proteins within that space. The basis of the approach is the geometric graph model that was proposed to better reflect biological networks than e.g. the small-world model [41]. Although the biological basis of the geometric graph model remains elusive, the authors show that it measures network distance more reliably. Both of these topology based methods assign confidence as numerical values to protein-protein interactions in a network and are additionally able to predict new interaction candidates by assigning confidence scores to non-interactions. However, both methods have certain shortcomings. The method by Goldberg and Roth is able to assess the confidence of those interactions whose participants have common neighbors only. Often, however, interacting proteins do not share neighbors. The method of Kuchaiev *et al.* appears limited in that it requires fixing six free parameters. These include algorithm-specific parameters as well as the prior probability for interactions which depends on knowledge about the interactome size.

Here, we propose CAPPIC (cluster-based assessment of protein-protein interaction confidence) – a novel approach that exploits the inherent modular structure of interactomes for confidence assessment of protein-protein interactions. Our method combines the basic principles of the topology based methods described above: high neighborhood interconnectedness of a couple of proteins and short distance between them (the features exploited by Goldberg and Roth and Kuchaiev *et al.*, respectively) are indicators that both proteins participate in the same module. We apply Markov clustering [42] to the line graph [43] of an interaction network to dissect it into modules of interactions. As demonstrated in [44], this strategy can generate interaction clusters that significantly overlap with known biological pathways. Notably, the interaction clusters overlap in their protein constitution. This is biologically more meaningful than clustering the proteins into disjoint modules because pathways and protein machineries are known to overlap [10,21]. The rationale behind our approach is that proteins that are specific to certain modules are expected to have more interactions with proteins that are specific to the same modules than with other proteins [39]. Intuitively, we assign low confidence to interactions that disagree with the modular structure of biological networks and high confidence to those that comply with it. This rationale has also been used as a basis of approaches for the detection of binary interactions [10] or protein complexes [45] from complex purification data or to reveal dynamic interaction patterns during the human spliceosome cycle [8]. While the aim of CAPPIC is to detect false positive interactions, a different approach, which is however also based on the principle of high link density within network modules, has been proposed for identifying false negatives [46].

We applied our method to six large-scale interaction networks from yeast to assess its performance and compare it to previous topology-based methods (Table 1). The six networks were fundamentally different with respect to their biological and topological properties as they have been generated using different techniques. These included: 1) a network that was generated using the protein-fragment complementation assay (PCA) technology [12] (*Tarassov-all*); 2) a sub-network of Tarassov-all obtained by the authors after applying several filtering steps [12] (*Tarassov-hq*); 3) a combined network of interactions found by yeast-two-hybrid (Y2H) screens (*Yu-Ito-Uetz*) comprising the networks published by Yu *et al.* [9], Ito *et al.* [47] and Uetz *et al.* [48] (the integrated data set was retrieved from [9]); 4) a network of interactions predicted by Collins *et al.* [49] from protein complex data resulting from affinity purification assays coupled to mass spectrometry (AP-MS) [10,11] (*Collins*),

Table 1 Yeast interactome maps used in this study for method evaluation

network property	Tarassov-all	Tarassov-hq	Yu-Ito-Uetz	Collins	CPDB-yeast	Costanzo
references	[12]	[12]	[9]	[49]	[51]	[52]
method	PCA	PCA	Y2H	AP-MS	multiple	genetic
node count	2238 (2293)	889 (1124)	1647 (2018)	1002 (1620)	6073 (6075)	4278 (4278)
link count	9360 (9646)	2407 (2770)	2518 (2930)	8313 (9064)	74332 (74333)	63927 (63927)
clustering coefficient	0.14	0.24	0.08	0.72	0.19	0.06
links in triangles	5861 (62%)	1761 (73%)	440 (17%)	8129 (97%)	63385 (85%)	47822 (74%)
mean shortest path length	3.7	5.6	5.6	5.5	2.7	2.9
links with ≥ 3 publications	546 (5%)	419 (17%)	598 (23%)	1635 (19%)	6324 (8%)	2546 (3%)

downloaded from BioGRID [50]; 5) a comprehensive physical interaction network from the interaction meta-database ConsensusPathDB, release 6(yeast) [51] obtained by the integration of multiple publicly accessible interaction repositories (CPDB-yeast); and 6) a genetic interaction map published by Costanzo *et al.* [52] obtained at a stringent experimental cutoff (*Costanzo*). The physical interaction networks constitute a representative benchmark since they result from different, major interaction detection techniques: yeast-two-hybrid, protein-fragment complementation, affinity purification, and integration of interaction data obtained with different methods. We applied our method additionally to the genetic interaction map by Costanzo *et al.* to provide evidence that it is not limited to physical interactome maps. To show that CAPPIC's performance was consistent across taxonomic species, we also applied it to two human networks. The first was obtained by merging the 15 largest, high-quality human yeast-two-hybrid data sets including refs. [5-8] (Additional file 1: Table S1) (*Y2H-human*). The second network corresponded to the top 5% interactions from a probabilistic binary data set generated by Mazloom *et al.* [53] from mass spectrometry-based analysis of 3,290 immuno-precipitation experiments [54] (*Mazloom*). The properties of the two human networks are summarized in Additional file 2: Table S2.

An implementation of CAPPIC is available as a web-based tool called IntScore at http://intscore.molgen.mpg.de [55].

Results

Approach

Assessing protein interaction confidence by random walk interaction clustering

Interaction data are usually modeled as graphs where nodes represent proteins or genes and edges represent interactions between them. For assessing the confidence of every interaction in a network, we apply the following strategy (illustrated in Figure 1). First, the interaction graph is transformed into its line graph [43] where interactions are represented by nodes, and proteins are represented by links that connect their interactions (step 1 in Figure 1). Second, we deploy Markov clustering – an algorithm for network clustering through random walk simulation [42] – on the line graph to dissect it into disjoint clusters of interactions (step 2 in Figure 1). In the third and last step of the approach (step 3 in Figure 1), we evaluate the distribution of interactions among the resulting clusters. It is a key point that interactions of a given protein can be clustered together, or distributed among multiple clusters. A protein is specific to a cluster if the cluster is enriched in interactions of that protein. To quantify this enrichment, we define the fidelity $F_{p,c}$ of a protein p to cluster c as the value of the cumulative hypergeometric distribution function (Equation 1) given $L_{p,c}$, the number of interactions of protein p in cluster c; $L_{p,\cdot}$, the total number of interactions of p (called the degree of p); $L_{\cdot,c}$, the total number of interactions in c; and $L_{\cdot,\cdot}$, the total number of interactions in the network:

$$F_{p,c} = P(X \leq L_{p,c}) = \sum_{k=0}^{L_{p,c}} \frac{\binom{L_{p,\cdot}}{k}\binom{L_{\cdot,\cdot}-L_{p,\cdot}}{L_{\cdot,c}-k}}{\binom{L_{\cdot,\cdot}}{L_{\cdot,c}}} \qquad (1)$$

The value of the fidelity $F_{p,c}$ lies between 0 and 1, with values near or equal to 1 if a protein p is specific to cluster c, i.e. if it has relatively many links in that cluster. For a fixed $L_{p,c}$ it holds that the smaller the cluster (smaller $L_{\cdot,c}$), the greater the fidelity value. Finally, if all the links of two proteins lie within a cluster, the fidelity is greater for the protein with the higher degree.

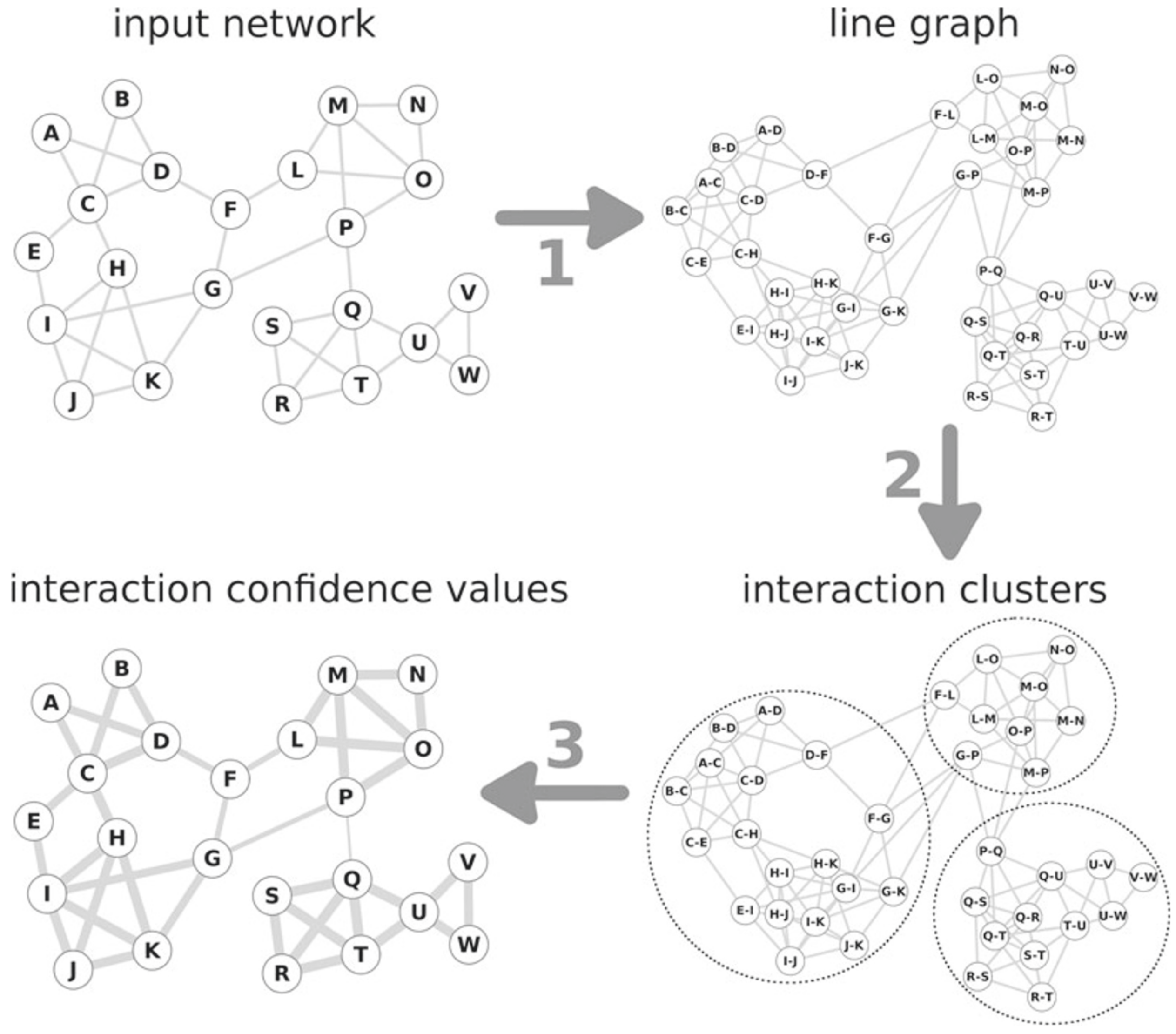

Figure 1 Outline of our interaction confidence assessment method. In the input interaction network (upper left picture), proteins are labeled with letters (A, B, etc.) and interactions between them are represented by edges. In the first step of the approach, we create the line graph of the given network where nodes represent interactions (labeled A–C, A–D, etc.) and edges represent shared interaction participants. In the second step, we use Markov clustering on this line graph to dissect it into interaction clusters. The clustering granularity is optimized in a previous step of the algorithm. Importantly, proteins can be part of more than one cluster. The relative number of interactions of a protein in a cluster determines how specific a protein is to that cluster. In the third step, we calculate confidence values for every interaction based on how specific both proteins are to the respective clusters. The thickness of interaction links in the lower left picture corresponds to the calculated interaction confidence values for this example network.

We define interaction confidence as the product of the fidelity values of both interacting proteins to the cluster c which the interaction has been assigned to:

$$\text{confidence}(l_{p_1,p_2}) = F_{p_1,c} \cdot F_{p_2,c} \quad (2)$$

Interactions get high confidence values if both proteins are specific to the cluster containing the interaction, and low confidence values when one or both of the proteins are not specific to the cluster.

Optimal clustering granularity is reliably determined through partial network rewiring

The interaction confidence scores calculated by CAPPIC are dependent on the granularity of the interaction clustering. It has been previously shown that modules in many complex networks, including protein interaction maps, are organized in a hierarchical manner [56]. Accordingly, interaction clustering can yield protein complexes, cellular machineries, pathways, or higher-order biological processes depending on the clustering granularity. To estimate the clustering granularity for a network that will result in the best discrimination between true and false interactions, we first randomly rewire a small part of the links in that network to generate a false interaction set. In the rewiring procedure, pairs of interactions are selected at random and two of the proteins are swapped so that no real interaction is reconstituted and the network stays connected. This way, two false interactions are generated for two real ones while the degree of each protein is preserved. Then, we calculate interaction confidence values of the resulting partially rewired network as described above using different inflation values. The inflation parameter of the Markov clustering algorithm essentially controls clustering granularity [42]. For every inflation value, we quantify the significance of the difference between confidence score distributions of the rewired and the remaining non-rewired links. This is done with the Wilcoxon rank-sum test under the alternative hypothesis that the confidence scores of the non-rewired links are greater than the confidence scores of the rewired links. The inflation value minimizing the Wilcoxon test P-value is considered optimal.

Experiments have shown that randomly rewiring 3% of the links in the granularity estimation procedure described above is a good choice because this yields a false interaction set of reasonable size while keeping most of the network intact. If the set of false interactions obtained through random rewiring is too small, the granularity estimation will lack statistical power, while if too many interactions are rewired, the network's original modular structure will be altered which will affect the granularity estimate. For all networks CAPPIC was applied on, random rewiring of 1%, 3%, 5%, or 10% of the interactions yielded very similar optimal granularity estimates.

Our granularity estimation strategy builds upon the assumption that the optimal granularity value inferred from a partially rewired network instance (where both false positive and false negative rates are increased compared to the real network) is transferable to the real network. We aimed to scrutinize this reasoning and verified for all reference networks that 1) the estimated optimal granularity was rather independent of the random choice of links for rewiring; and 2) that interaction clusters were similar for the intact and the partially rewired networks clustered with the same inflation value (see Additional file 3: Supplementary Text).

True positive interactions are assigned higher confidence than false positives

We measured the performance of CAPPIC and compared it to previously proposed network topology based interaction confidence assessment methods using five yeast physical interaction networks and one genetic interactome map, covering major interaction inference methods (Table 1). We first constructed positive (literature interactions) and negative (random links) link sets and then evaluated the methods using receiver operating characteristic (ROC) analysis. The positive set for each network consisted of interactions that are reported multiple times in the literature (ranging from 3% to 23% for the six reference networks, Table 1), since such interactions have been shown to be on average more reliable [13,16]. The negative interaction set consisted of links that resulted from a random rewiring of a small sub-set (3%) of the interactions in the respective network. Interactions from the partially rewired instance, ranked with decreasing confidence value were compared successively against the positive and negative benchmark sets to determine the true positive and false positive rates at each step. In general, CAPPIC assigned higher confidence to true interactions than false interactions (Figure 2). The area under the ROC curve (AUC), which quantifies the confidence ranking performance, was as high as 94% for the Collins network. For this data set, at a fixed specificity of 80% our method reached 95% sensitivity. On the other extreme, none of the methods in the analysis showed convincing performance on the combined Y2H network Yu-Ito-Uetz. In this example, Goldberg and Roth's method successfully classified interactions whose proteins shared network neighbors; however, such interactions comprised only 17% of Yu-Ito-Uetz (see 'X'-mark on the green line in Figure 2 and row "links in triangles" in Table 1) while the rest of the interacting protein pairs did not share network neighbors. Goldberg and Roth's method outperformed CAPPIC on the CPDB-yeast and Costanzo networks, whereas the method by Kuchaiev *et al.* did not discriminate (for unclear reasons) between true and false interactions better than random in these two cases. Generally, it performed worse than CAPPIC and Goldberg and Roth's method on all networks. Based on the results for all six networks, we conclude that the method of Goldberg and Roth is able to correctly identify a subset of high-confidence interactions, but will not provide predictions for interactions not involved in triangles. On the other hand, the method by Kuchaiev *et al.* and our approach generate confidence scores for the complete data set, which is often desired when the aim is to assess the confidence of all interactions (e.g. for weighting a non-weighted network) or to filter out a relatively small sub-set of low-confidence interactions. It should be noted that in order to define a reliable negative link set, we destroyed some real interactions (increasing the false negative rate) and simultaneously introduced the same number of false positive interactions into the network. Thus, the AUC values reported here probably slightly underestimate the real performance.

In the case of well-studied organisms such as yeast, data on protein complexes can be used to define the positive interaction sets alternatively to literature evidence as used

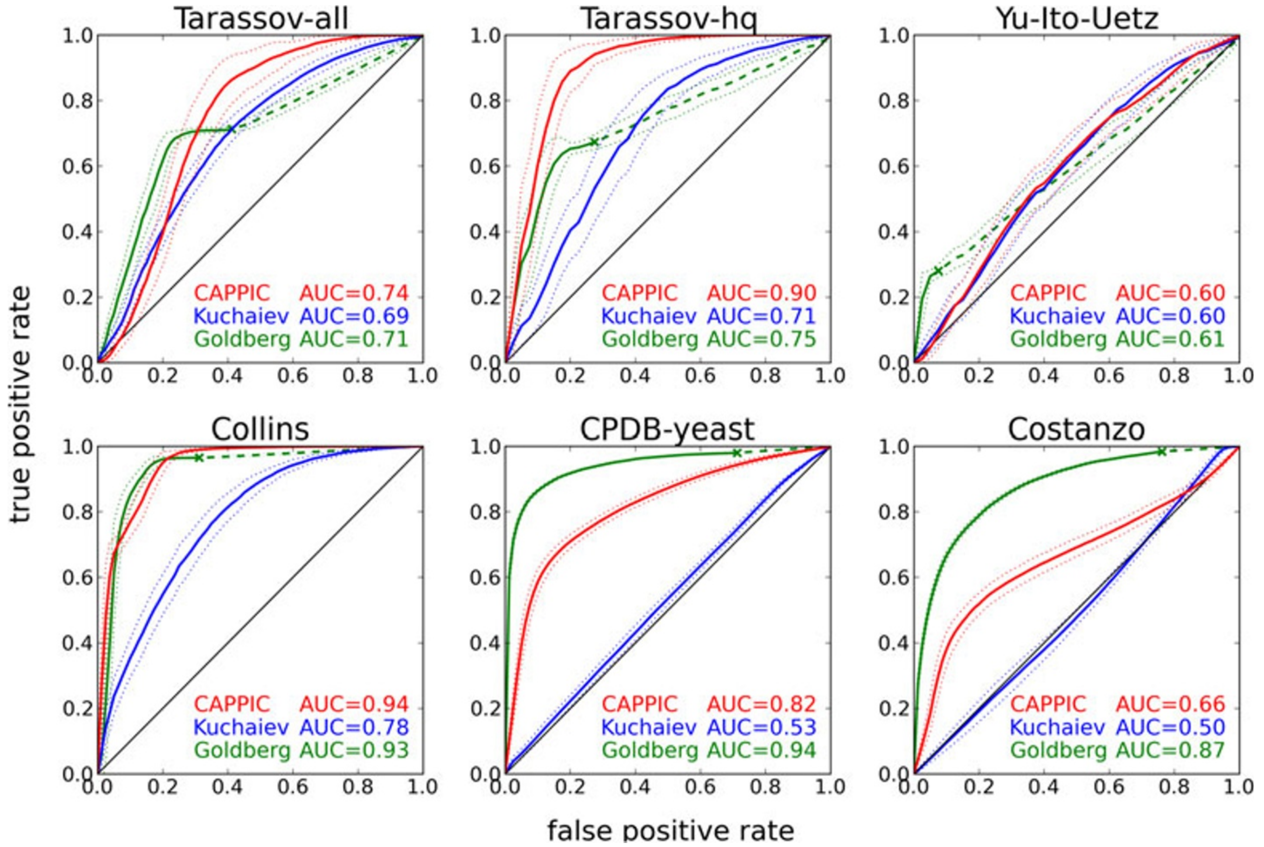

Figure 2 ROC analysis measuring the performance of CAPPIC in comparison to the methods by Goldberg and Roth and Kuchaiev *et al*. False positive rate (1-specificity) is plotted against true positive rate (sensitivity) for each of the six reference networks. Since the definition of a negative interaction set in the performance assessment involves a random process, the ROC plots summarize the outcome of 100 runs. Plots show the average ROC curves (thick lines), their standard error bands (dotted lines), as well as the mean area under the ROC curve (AUC) of all runs. The 'X'-marks on the green ROC curves correspond to the fraction of true/false interactions whose proteins share network neighbors and are thus scored by Goldberg and Roth's method.

above. We used two complex-based positive sets from yeast complexes obtained from CYC2008 [57] and from ref. [58]. The performance of CAPPIC (and often of the reference methods) was better with the complex-based compared to the literature-based positive set for almost all networks (Additional file 4: Figure S1). For example, the AUC for CAPPIC increased from 82% to 87-89% for CPDB-yeast and from 66% to 70-72% for Costanzo when the literature-based positive reference set was replaced by a complex-based one; improvements by 1-2% AUC were also observed for the Tarassov-all, Tarassov-hq and Collins networks (Additional file 4: Figure S1 versus Figure 2 in the Main text). However, despite the better performance with complex-based positive reference sets, such sets are not well-suited for measuring the performance on networks obtained by techniques such as yeast-two-hybrid [9]. This could be the reason for the slight decrease in performance (by 1-2% AUC) on the Yu-Ito-Uetz yeast-two-hybrid network compared to a literature-based positive set (Additional file 4: Figure S1 versus Figure 2 in the Main text). Moreover, the complex-based performance estimate may be positively biased since protein complexes in the reference data may have been defined at least partially on the basis of the analysed interaction networks.

Cluster based confidence scores corroborate experimental interaction evidence

To compare confidence values calculated by CAPPIC with experiment-based interaction scores, we exploited the fact that some of the interactions in Tarassov-all have been designated high-quality by the authors based on experimental interaction intensity [12]. We tested whether our method assigned significantly higher confidence scores to high-quality interactions than to the rest of the interactions in Tarassov-all. As shown in Figure 3, the confidence score distributions of both interaction sub-sets were different. Using the Wilcoxon rank-sum test we confirmed

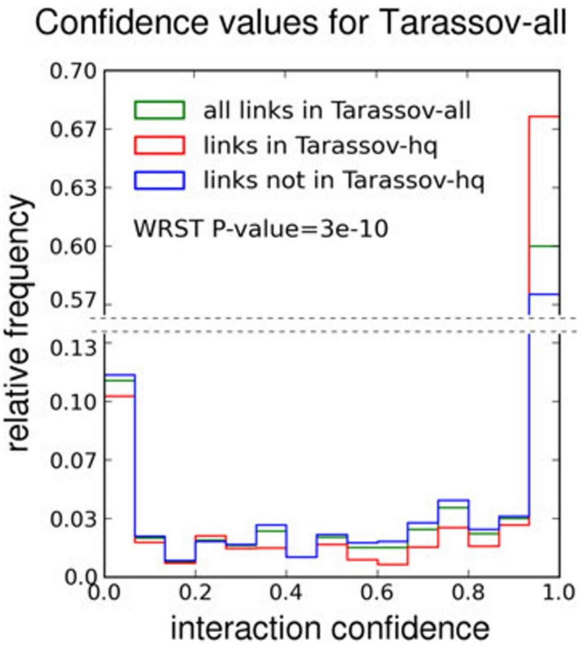

Figure 3 Histogram of confidence scores for interactions in Tarassov-all calculated by our method. The normalized histograms of interaction confidence scores are shown for the complete Tarassov-all network, as well as for its high-quality (Tarassov-hq) and non-high-quality parts. WRST: Wilcoxon rank sum test of the difference between confidence score distributions of both network parts. Note that the Y-axis is interrupted to better show the differences between the three data sets.

that confidence values were greater for high-quality interactions than for the rest of the links in Tarassov-all (P-value $< 3 * 10^{-10}$). The high agreement between cluster based interaction confidence scores and experimental interaction weight for the Tarassov-all network was corroborated by a significant Spearman rank correlation between both ($\rho = 0.3$, p-value $< 10^{-5}$).

High-confidence interactions are more consistent in biological process and cellular compartment annotation

Interacting proteins are expected to participate in related biological processes and to be co-localized in compartments of the cell [59]. Therefore, Gene Ontology (GO) [60] annotations of interacting proteins agree more often than expected by chance. We utilized the semantic similarity of GO biological process and cellular compartment annotations of proteins predicted to interact as a performance measure of our approach. If confidence values reflect the veracity of discovered interactions, we expect interactions with higher confidence score to have a higher average semantic similarity of the proteins' GO annotations. To test this, we ranked interactions from each reference network by confidence score and arranged them into five equal sized bins. The average GO semantic similarity (GOSemSim) values for interacting proteins in each bin are plotted in Figure 4. The GOSemSim generally correlated with interaction confidence. In several extreme cases (e.g. Costanzo), the average GOSemSim of low-confidence interactions was barely distinguishable from the average GOSemSim of random protein pairs (dashed horizontal lines), while the higher-confidence interactions reached average GOSemSim far above the average value of all interactions in the respective network (continuous horizontal lines). These results suggest that there are more false links among the lower-confidence interactions than among the higher-confidence ones.

Furthermore, if low-confidence interactions are removed from interaction clusters, the latter become more consistent regarding the pathway annotations of the contained proteins (see Additional file 3: Supplementary Text). Our approach can thus be used to obtain more refined functional modules in interaction data sets.

The performance of CAPPIC is consistent between yeast and human networks

To exemplify that the performance of CAPPIC is consistent for different taxonomic species, we also applied it to two human networks: Y2H-human (Additional file 1: Table S1) and Mazloom [53]. Figure 5 shows the corresponding ROC plots summarizing the performance of CAPPIC and of the reference methods (analogous to Figure 2), as well as the GO semantic similarity as a function of the CAPPIC score (analogous to Figure 4) for these networks. Notably, the performance of CAPPIC on the Y2H-human and Mazloom human networks was very similar to the performance on the yeast counterparts obtained by analogous techniques (Yu-Ito-Uetz and Collins yeast networks, respectively). For example, CAPPIC achieved 90% AUC on the Mazloom network and 62% AUC on the much sparser yeast-two-hybrid network, outperforming the reference methods in both cases (Figure 5). In the case of the Mazloom network, we also measured the agreement between CAPPIC scores and interaction ranks that were based on evidence from 3,290 co-immunoprecipitation experiments [53]. The CAPPIC scores were calculated independently of the ranks or the confidence values assigned in the original study. The Spearman correlation coefficient between interaction ranks and CAPPIC scores was $\rho = -0.34$ (p-value $< 10^{-5}$). The correlation is negative since interactions with smaller ranks tend to get higher CAPPIC scores. As in the case of the yeast Tarassov-all network described above (that has been obtained by protein-fragment complementation assay), CAPPIC corroborates independent interaction evidence also for this human immuno-precipitation based network.

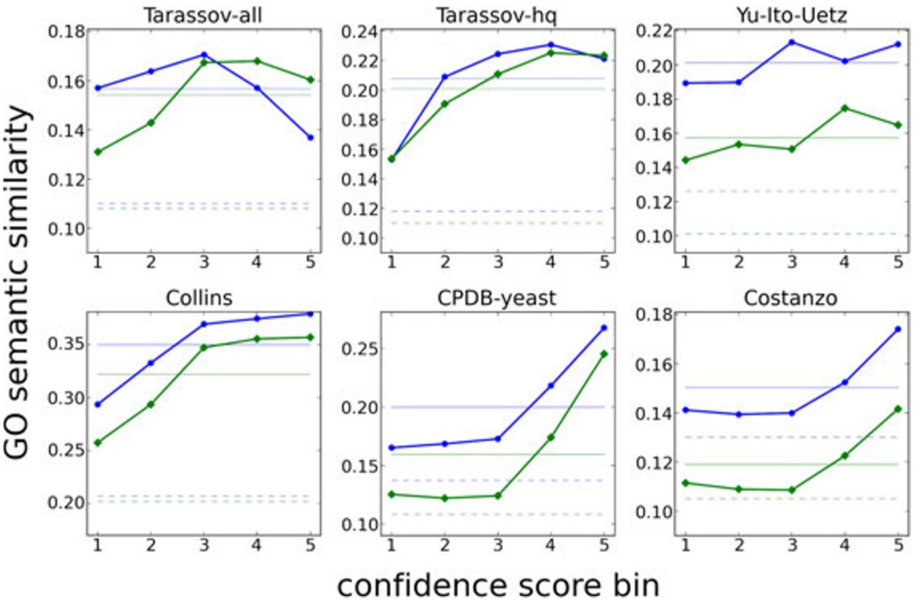

Figure 4 Correlation of CAPPIC interaction confidence with semantic similarity of Gene Ontology co-annotations. Interactions from every network are ranked by confidence and divided into five equal sized bins (X-axis); for each bin, the average semantic similarity of GO biological process (blue) and cellular component (green) annotations of interacting proteins is shown (Y-axis). Additionally, the pale continuous lines correspond to the mean GO semantic similarity over the complete network rather than the separate bins. The dashed lines reflect the average GO semantic similarity of random pairs of proteins from the network.

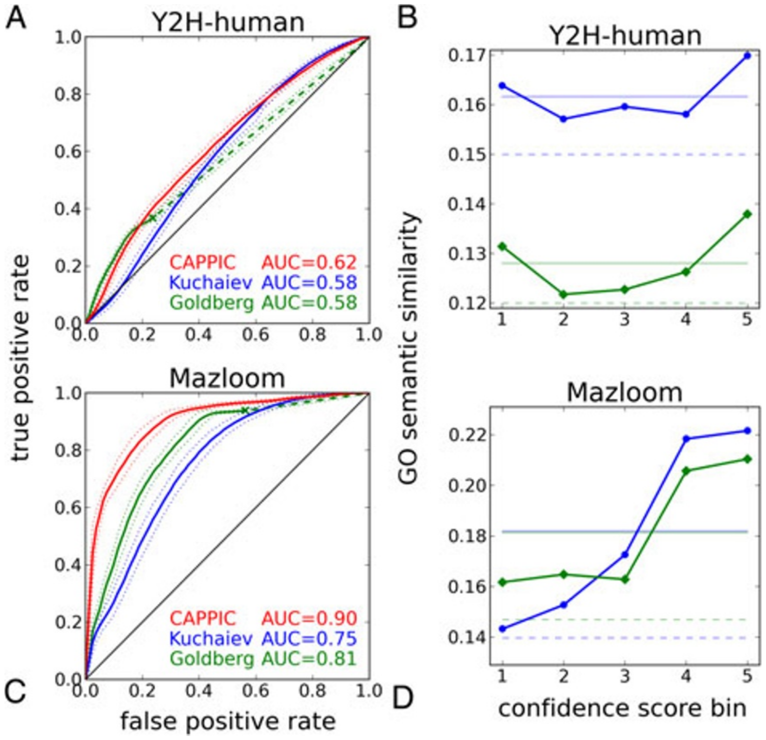

Figure 5 Performance of CAPPIC on human networks. A) and **C)**: ROC plots for Y2H-human and Mazloom, correspondingly (for details, see Figure 2 legend); **B)** and **D)**: correlation of CAPPIC scores with GO semantic similarity for Y2H-human and Mazloom, correspondingly (for details, see Figure 4 legend).

Discussion

Network topology-based approaches are motivated by the fact that the structure of interaction networks is not random but reflects biological functionality [33,34]. Modularity is a topological property that is inherent to protein-protein interaction networks [10,39,56]. We propose a novel method (CAPPIC) to assess the confidence of individual protein interactions in an interaction network. Our method exploits network modularity alone for estimating the confidence of interactions and does not require any additional knowledge about the interacting proteins or the techniques used to generate the data. We demonstrate the power of CAPPIC in discriminating between true and false interactions on the basis of five physical protein interaction networks and one genetic interaction map from yeast, as well as two distinct interaction data sets from human.

CAPPIC compares well to previous topology-based approaches by Goldberg and Roth and Kuchaiev *et al.* in assigning continuous confidence scores to all interactions in a given physical interaction network. The method of Goldberg and Roth is dependent on shared network neighbors of interacting proteins; however, many interacting proteins do not share neighbors. As a result, many interactions are scored with a confidence value of zero. However, integrative approaches operating on networks usually take probabilistic rather than binary data as input. Thus, the goal of confidence assessment is often to assign a continuous score to all interactions rather than to filter for a small subset. In particular, all proteins with a single interaction partner are disregarded by Goldberg and Roth's method, albeit these single protein associations could give important clues about the function of these proteins. Both methods, Kuchaiev *et al.* and CAPPIC, are able to assign continuous scores also to such interactions. In contrast to the method of Kuchaiev *et al.*, CAPPIC does not require any parameter input. The only parameter that influences the resulting confidence scores – clustering granularity – is optimized internally for each individual input network. Our results have shown that the number of clusters obtained at the optimal granularity tends to be small for all reference networks, ranging from 10 to 50 clusters (see Additional file 5: Figure S2 and Figure ST1 in Additional file 3: Supplementary Text). This alleviated our initial concerns that interactions executing essential crosstalks between related pathways could be assigned low confidence. Because the optimal granularity tends to be very coarse, closely related pathways will probably not be separated but clustered together.

CAPPIC should be applicable for weighting any binary network with an inherent modular structure (for examples, see [61]). Notably, it does not consider the technique used to generate the network (unlike other approaches that integrate a fixed, subjective judgment on the reliability of different techniques, e.g. ref. [62]). CAPPIC fails to generate reliable confidence scores in cases where modularity is not pronounced, i.e. if many of the real links within biological modules (complexes, pathways, etc.) are missing. This is probably the case with the Yu-Ito-Uetz and Y2H-human reference networks: here, the topological signal that our method exploits seems to be weaker and it achieves only 60-62% AUC. Absence of modularity in this example is evidenced by the relatively low clustering coefficient [40] of 0.08 which is nine times lower than that of the Collins network where CAPPIC achieves 94% AUC and six times lower than that of the Mazloom network (90% AUC). Moreover, the Yu-Ito-Uetz data set is the sparsest of all yeast reference networks (Table 1). To conclude, results on all example networks suggest that CAPPIC is well suited to score datasets with moderate to high interaction density.

Unlike the reference methods, CAPPIC is able to accommodate experimental evidence weights of interactions. Interaction detection techniques often associate such weights with predicted interactions, reflecting for example the number of times an interaction is observed in repetitions of a yeast-two-hybrid experiment [7,9,13] or the reporter intensity value in the case of a protein-fragment complementation assay [12]. If available, such weights can be exploited by our method in its random walk based interaction clustering step. This can improve the interaction clustering result and consequently increase the performance of confidence assessment. However, since we set out to estimate the performance of CAPPIC in comparison to other methods that cannot accommodate interaction weights, we did not make use of this advantage in this work and considered all interactions equal. Moreover, the ability to incorporate experimental interaction weights helps to avoid interaction data pre-filtering, commonly executed to derive binary interaction networks (where pairs of proteins either interact or not). Such filtering of probabilistic interaction data is inherently associated with data loss. Similarly, it is a common practice to remove interaction hubs in a dataset to improve its quality (*e.g.*, ref. [6]). As exemplified in Additional file 6: Figure S3 for the yeast hubs PHO85 (a Cyclin-dependent kinase; 467 interactions) and UBC7 (an E2 ubiquitin ligase; 622 interactions) in the CPDB-yeast network, CAPPIC assigns on average lower scores to interactions of hubs. However, a considerable fraction of their interactions scores highly: 29% of the interactions of PHO85 and 25% of the interactions of UBC7 are assigned higher CAPPIC scores than the median score of the complete network. This suggests that a complete removal of hubs from the network could unnecessarily remove high-quality protein-protein interactions and emphasizes the utility of confidence scoring.

Our approach can be combined with other lines of interaction evidence like other topological features, protein co-expression, or interaction homology to achieve even better scoring performance [22]. While the aggregation of different features holds the promise of even more reliable interaction confidence assessment, it depends on reference interaction sets. At present, even for yeast the construction of an appropriate reference set is still a daunting task [9].

Conclusions

Since biological interaction networks contain false positives, assessing the confidence of individual interactions in order to weight or filter interaction data is a crucial step that should precede network-based inferences. Here we propose a network topology based method called CAPPIC that estimates interaction confidence by exploiting the network's inherent modularity. CAPPIC requires no reference interaction sets or parameter settings. Based on five large-scale physical interaction networks from yeast, we show that our method compares well to other topology-based approaches. Confidence scores calculated with CAPPIC also correlate well with the Gene Ontology co-annotation of interacting proteins, and corroborate experimental evidence of physical interactions. CAPPIC is limited neither to physical interactome maps nor to yeast networks as it also performs well on a large yeast genetic interaction network and on two human protein-protein interaction data sets.

Methods

Application of Markov clustering algorithm

To cluster a network of interactions, we use the original implementation of the Markov clustering algorithm (version 10-201 downloaded from http://www.micans.org/mcl/sec_software.html). The inflation scan which aims to optimize clustering granularity is carried out in two steps: a coarse scan with step size of 0.1 within a fixed range $I \in [1.1, 2.0]$ (where I is the Markov clustering inflation value) is followed by a fine scan with step size of 0.025 around the optimal inflation value resulting from the coarse scan ± 0.1. In general, the inflation parameter takes values from the interval $I \in (1.0, 30.0]$ with higher values resulting in finer granularity. In all our experiments, the optimal inflation estimate was far below 2.0 (see Figure ST1 in Additional file 3: Supplementary Text), motivating the choice of this value as an upper boundary of the inflation scan.

Receiver operating characteristic analysis

To conduct ROC analysis, we constructed true and false interaction sets. The positive set comprised interactions published in at least three papers in total. An exception was made for the Costanzo network because of the scarcity of genetic interaction data: the positive set in this case consisted of interactions that are also reported in [63]. Literature evidences were retrieved with the interaction evidence mining ConsensusPathDB plugin [64]. The negative interaction set was constructed by randomly rewiring 3% of the interactions in the respective network. For each partially rewired network, we ranked interactions according to confidence as calculated with CAPPIC and reference methods and created receiver operating characteristic (ROC) curves. The performance of a given confidence assessment method in ranking positive interactions higher than negative ones was quantified with the area under the ROC curve (AUC). The AUC is around 50% if a method does not perform better than random interaction ranking, and is closer to 100% the better it ranks positive interactions higher than negative ones. Since the constitution of the negative and positive sets involves a random process (that is, the random selection of interactions for rewiring), we repeated the procedure 100 times and averaged ROC results.

Application of reference methods

We set the number of yeast genes to 6,000 in the method by Goldberg and Roth. The parameters of the method by Kuchaiev *et al.* (implemented as Matlab scripts downloaded from http://www.kuchaev.com/Denoising) were set as follows: priorEdge=0.002945 (which results when the estimated yeast interactome size of 53,000 interactions [65] is divided by the number of all possible protein pairs, 6,000 choose 2); priorNonEdge=1-priorEdge; dim=5 (default); d=3 (default); learnSetSize=min(5,000 or half the number of interactions); delta=1.0; and stopEps=0.01 (default). In the case of Costanzo, dim=3 because the program (run on a standard AMD X2 5600+ machine with 8GB of RAM running Matlab version 7.10.0.499 under Linux) did not return results within five days for a higher number of dimensions.

Assessing semantic similarity of Gene Ontology annotations

For each network, we obtained the GO semantic similarity of biological process and cellular component annotations of interacting proteins using the method proposed by Resnik [66] implemented in the software package GOSemSim version 1.8.0 [67]. GO annotations inferred from physical interaction (GO evidence code 'IPI') were excluded from the semantic similarity calculation to avoid circularity. For each network, interactions were ranked by increasing confidence score and divided into five equal sized bins. The mean semantic similarity values for interacting proteins within each bin were calculated. Additionally, the mean GO semantic similarity for random pairs of proteins from the respective network was assessed by completely rewiring the networks while preserving each protein's degree and then calculating the mean GO semantic similarity of links in those randomized networks.

Additional files

Additional file 1: Table S1. Interaction data sets merged to construct the Y2H-human network.The table lists the studies that contribute yeast-two-hybrid interactions for the merged Y2H-human network. The file is in XLS format and is viewable e.g. with LibreOffice or Microsoft Excel.

Additional file 2: Table S2. Properties of the Y2H-human and Mazloom networks. The table shows the properties of the human networks used in the analysis (analogous to Table 1 in the main text). The file is in XLS format and is viewable e.g. with LibreOffice or Microsoft Excel.

Additional file 3: Supplementary Text. This file contains additional text and figures demonstrating the validity of the partial random rewiring approach for clustering parameter optimization, as well as text and figures showing that CAPPIC scores can be used for interaction cluster de-noising. The file is in PDF format and is viewable e.g. with Adobe Reader.

Additional file 4: Figure S1. ROC plots with complex-based positive reference sets. Receiver operating characteristic analysis results for the yeast reference networks where complex-based positive reference sets have been used. Complexes were obtained from ref. [57] (A) and from ref. [58] (B). The figure is otherwise analogous to Figure 2.

Additional file 5: Figure S2. Cluster number and sizes for the yeast reference networks clustered with the optimal granularity. Yeast reference networks were clustered at the optimal inflation value into 10-50 interaction clusters. Here, the cluster sizes in terms of number of interactions (blue line, left-hand-side Y-axis) and number of genes/proteins (green line, right-hand-side Y-axis) per cluster are plotted.

Additional file 6: Figure S3. Distribution of CAPPIC scores for the hubs PHO85 and UBC7 in comparison to the whole data set.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AK and US conceived the method. AG and RH provided feedback on the method and contributed ideas. AK developed the method and carried out the experiments. AK and US wrote the manuscript, AG and RH provided feedback on the manuscript. All authors read and approved the manuscript.

Acknowledgements

This work was funded by the European Commission under its 7FP grant diXa (grant number 283775) and the German Ministry for Education and Research under its grants MedSys - PREDICT (grant number 0315428A) and NGFNplus (NeuroNet TP3, grant number 01GS08171).

Received: 26 March 2012 Accepted: 16 August 2012
Published: 10 October 2012

References

1. Stelzl U, Wanker EE: **The value of high quality protein-protein interaction networks for systems biology.** *Curr Opin Chem Biol* 2006, **10**(6):551–558. [http://www.ncbi.nlm.nih.gov/pubmed/17055769]. [PMID: 17055769]
2. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Sci (New York, N.Y.)* 2003, **302**(5644):449–453. [http://www.ncbi.nlm.nih.gov/pubmed/14564010]. [PMID: 14564010]
3. Kamburov A, Goldovsky L, Freilich S, Kapazoglou A, Kunin V, Enright AJ, Tsaftaris A, Ouzounis CA: **Denoising inferred functional association networks obtained by gene fusion analysis.** *BMC Genomics* 2007, **8:**460. [http://www.ncbi.nlm.nih.gov/pubmed/18081932]. [PMID: 18081932]
4. Pazos F, Juan D, Izarzugaza JMG, Leon E, Valencia A. *Methods in Mol Biol (Clifton, N.J.)* 2008, **484:**523–535. [http://www.ncbi.nlm.nih.gov/pubmed/18592199]. [PMID: 18592199]
5. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122**(6):957–968. [http://www.ncbi.nlm.nih.gov/pubmed/16169070]. [PMID: 16169070]
6. Bandyopadhyay S, Chiang Cy, Srivastava J, Gersten M, White S, Bell R, Kurschner C, Martin CH, Smoot M, Sahasrabudhe S, Barber DL, Chanda SK, Ideker T: **A human MAP kinase interactome.** *Nat Methods* 2010, **7**(10):801–805. [http://www.ncbi.nlm.nih.gov/pubmed/20936779]. [PMID: 20936779]
7. Vinayagam A, Stelzl U, Foulle R, Plassmann S, Zenkner M, Timm J, Assmus HE, Andrade-Navarro MA, Wanker EE: **A directed protein interaction network for investigating intracellular signal transduction.** *Sci Signaling* 2011, **4**(189):rs8. [http://www.ncbi.nlm.nih.gov/pubmed/21900206]. [PMID:21900206]
8. Hegele A, Kamburov A, Grossmann A, Sourlis C, Wowro S, Weimann M, Will CL, Pena V, Lührmann R, Stelzl U: **Dynamic protein-protein interaction wiring of the human spliceosome.** *Mol Cell* 2012, **45**(4):567–580. [http://www.ncbi.nlm.nih.gov/pubmed/22365833]. [PMID: 22365833]
9. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual J, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, Smet Ad, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabasi A, et. al: **High-quality binary protein interaction map of the yeast interactome network.** *Sci (New York, N.Y.)* 2008, **322**(5898):104–110. [http://www.ncbi.nlm.nih.gov/pubmed/18719252]. [PMID: 18719252]
10. Gavin A, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier M, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G: **Proteome survey reveals modularity of the yeast cell machinery.** *Nature* 2006, **440**(7084):631–636. [http://www.ncbi.nlm.nih.gov/pubmed/16429126]. [PMID: 16429126]
11. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, et. al: **Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.** *Nature* 2006, **440**(7084):637–643. [http://www.ncbi.nlm.nih.gov/pubmed/16554755]. [PMID: 16554755]
12. Tarassov K, Messier V, Landry CR, Radinovic S, Molina MMS, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW: **An in vivo map of the yeast protein interactome.** *Sci (NewYork, N.Y.)* 2008, **320**(5882):1465–1470. [[http://www.ncbi.nlm.nih.gov/pubmed/18467557]. [PMID: 18467557]]
13. Venkatesan K, Rual J, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh K, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, Smet Ad, Dann E, Smolyar A, Vinayagam A, Yu H, zeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, et. al: **An empirical framework for binary interactome mapping.** *Nat Methods* 2009, **6:**83–90. [http://www.ncbi.nlm.nih.gov/pubmed/19060904]. [PMID: 19060904]
14. Ideker T, Sharan R: **Protein networks in disease.** *Genome Res* 2008, **18**(4):644–652. [http://www.ncbi.nlm.nih.gov/pubmed/18381899]. [PMID: 18381899]
15. Vidal M, Cusick ME, Barabasi A: **Interactome networks and human disease.** *Cell* 2011, **144**(6):986–998. [http://www.ncbi.nlm.nih.gov/pubmed/21414488]. [PMID: 21414488]
16. Mering Cv, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417**(6887):399–403. [http://www.ncbi.nlm.nih.gov/pubmed/12000970]. [PMID: 12000970]
17. Deane CM, Salwinski L, Xenarios I, Eisenberg D: **Protein interactions: two methods for assessment of the reliability of high throughput**

observations. *Mol & Cell Proteomics: MCP* 2002, **1**(5):349–356. [http://www.ncbi.nlm.nih.gov/pubmed/12118076]. [PMID: 12118076]

18. Jansen R, Gerstein M: **Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction.** *Curr Opin Microbiol* 2004, **7**(5):535–545. [http://www.ncbi.nlm.nih.gov/pubmed/15451510]. [PMID: 15451510]
19. Kritikos GD, Moschopoulos C, Vazirgiannis M, Kossida S: **Noise reduction in protein-protein interaction graphs by the implementation of a novel weighting scheme.** *BMC Bioinf* 2011, **12:**239. [http://www.ncbi.nlm.nih.gov/pubmed/21679454]. [PMID: 21679454]
20. Yu J, Murali T, Finley J, Russell L: **Assigning confidence scores to protein-protein interactions.** *Methods Mol Biol (Clifton, N.J.)* 2012, **812:**161–174. [http://www.ncbi.nlm.nih.gov/pubmed/22218859]. [PMID: 22218859]
21. Nepusz T, Yu H, Paccanaro A: **Detecting overlapping protein complexes in protein-protein interaction networks.** *Nat Methods* 2012, **9:**471–472. [http://www.ncbi.nlm.nih.gov/pubmed/22426491]. [PMID: 22426491]
22. Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker: **A direct comparison of protein interaction confidence assignment schemes.** *BMC Bioinf* 2006, **7:**360. [http://www.ncbi.nlm.nih.gov/pubmed/16872496]. [PMID: 16872496]
23. Chua HN, Wong L: **Increasing the reliability of protein interactomes.** *Drug Discovery Today* 2008, **13**(15-16):652–658. [http://www.ncbi.nlm.nih.gov/pubmed/18595769]. PMID: 18595769
24. Kemmeren P, Berkum NLv, Vilo J, Bijma T, Donders R, Brazma A, Holstege FCP: **Protein interaction verification and functional annotation by integrated analysis of genome-scale data.** *Mol Cell* 2002, **9**(5):1133–1143. [http://www.ncbi.nlm.nih.gov/pubmed/12049748]. [PMID: 12049748]
25. Deng M, Sun F, Chen T: **Assessment of the reliability of protein-protein interactions and protein function prediction.** *Pacific Symp on Biocomputing. Pacific Symp Biocomputing* 2003:140–151. [http://www.ncbi.nlm.nih.gov/pubmed/12603024]. [PMID: 12603024]
26. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species.** *Proc Nat Acad Sci USA* 2005, **102**(6):1974–1979. [http://www.ncbi.nlm.nih.gov/pubmed/15687504]. [PMID: 15687504]
27. Li D, Liu W, Liu Z, Wang J, Liu Q, Zhu Y, He F: **PRINCESS, a protein interaction confidence evaluation system with multiple data sources.** *Mol & Cell Proteomics* 2008, **7**(6):1043–1052. [http://www.mcponline.org/content/7/6/1043.abstract]
28. Bader JS, Chaudhuri A, Rothberg JM, Chant J: **Gaining confidence in high-throughput protein interaction networks.** *Nat Biotechnol* 2004, **22:**78–85. [http://www.ncbi.nlm.nih.gov/pubmed/14704708]. [PMID: 14704708]
29. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world.** *Proc Nat Acad Sci USA* 2003, **100**(8):4372–4376. [http://www.ncbi.nlm.nih.gov/pubmed/12676999]. [PMID: 12676999]
30. Saito R, Suzuki H, Hayashizaki Y: **Construction of reliable protein-protein interaction networks with a new interaction generality measure.** *Bioinformatics (Oxford, England)* 2003, **19**(6):756–763. [http://www.ncbi.nlm.nih.gov/pubmed/12691988]. [PMID: 12691988]
31. Chen J, Hsu W, Lee ML, Ng S: **Discovering reliable protein interactions from high-throughput experimental data using network topology.** *Artif Intelligence Med* 2005, **35**(1-2):37–47. [http://www.ncbi.nlm.nih.gov/pubmed/16055319]. [PMID: 16055319]
32. Kuchaiev O, Rasajski M, Higham DJ, Przulj N: **Geometric de-noising of protein-protein interaction networks.** *PLoS Comput Biol* 2009, **5**(8):e1000454. [http://www.ncbi.nlm.nih.gov/pubmed/19662157]. [PMID: 19662157]
33. Barabasi A, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**(2):101–113. [http://www.ncbi.nlm.nih.gov/pubmed/14735121]. [PMID: 14735121]
34. Alon U: **Network motifs: theory and experimental approaches.** *Nat Rev Genet* 2007, **8**(6):450–461. [http://www.ncbi.nlm.nih.gov/pubmed/17510665]. [PMID: 17510665]
35. Goh K, Cusick ME, Valle D, Childs B, Vidal M, Barabasi A: **The human disease network.** *Proc Nat Acad Sci USA* 2007, **104**(21):8685–8690. [http://www.ncbi.nlm.nih.gov/pubmed/17502601]. [PMID: 17502601]
36. Berger SI, Ma'ayan A, Iyengar R: **Systems pharmacology of arrhythmias.** *Sci Signaling* 2010, **3**(118):ra30. [http://www.ncbi.nlm.nih.gov/pubmed/20407125]. [PMID: 20407125]
37. Yosef N, Ungar L, Zalckvar E, Kimchi A, Kupiec M, Ruppin E, Sharan R: **Toward accurate reconstruction of functional protein networks.** *Mol Syst Biol* 2009, **5:**248. [http://www.ncbi.nlm.nih.gov/pubmed/19293828]. [PMID: 19293828]
38. Alexander RP, Kim PM, Emonet T, Gerstein MB: **Understanding modularity in molecular networks requires dynamics.** *Sci Signaling* 2009, **2**(81):pe44. [http://www.ncbi.nlm.nih.gov/pubmed/19638611]. [PMID: 19638611]
39. Fortunato S: **Community detection in graphs.** *Phys R* 2010, **486**(3-5):75–174. [http://www.sciencedirect.com/science/article/B6TVP-4XPYXF1-1/2/99061fac6435db4343b2374d26e64ac1]
40. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393**(6684):440–442. [http://www.ncbi.nlm.nih.gov/pubmed/9623998]. [PMID: 9623998]
41. Higham DJ, Rasajski M, Przulj N: **Fitting a geometric graph to a protein-protein interaction network.** *Bioinformatics (Oxford, England)* 2008, **24**(8):1093–1099. [http://www.ncbi.nlm.nih.gov/pubmed/18344248]. [PMID: 18344248]
42. Dongen Sv: *A Cluster algorithm for graphs. PhD thesis, Centrum voor Wiskunde en Informatica*. Netherlands: Amsterdam; 2000.
43. Whitney H: **Congruent graphs and the connectivity of graphs.** *Am J Mathematics* 1932, **54:**150–168. [http://www.jstor.org/stable/2371086]. [ArticleType: research-article / Full publication date: Jan. 1932 / Copyright 1932 The Johns Hopkins University Press]
44. Pereira-Leal JB, Enright AJ, Ouzounis CA: **Detection of functional modules from protein interaction networks.** *Proteins* 2004, **54:**49–57. [http://www.ncbi.nlm.nih.gov/pubmed/14705023]. [PMID: 14705023]
45. Friedel CC, Krumsiek J, Zimmer R: **Bootstrapping the interactome: unsupervised identification of protein complexes in yeast.** *J of Comput Biol: A J of Comput Mol Cell Biol* 2009, **16**(8):971–987. [http://www.ncbi.nlm.nih.gov/pubmed/19630542]. [PMID: 19630542]
46. Yu H, Paccanaro A, Trifonov V, Gerstein M: **Predicting interactions in protein networks by completing defective cliques.** *Bioinf (Oxford, England)* 2006, **22**(7):823–829. [http://www.ncbi.nlm.nih.gov/pubmed/16455753]. [PMID: 16455753]
47. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Nat Acad Sci USA* 2001, **98**(8):4569–4574. [http://www.ncbi.nlm.nih.gov/pubmed/11283351]. [PMID: 11283351]
48. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403**(6770):623–627. [http://www.ncbi.nlm.nih.gov/pubmed/10688190]. [PMID: 10688190]
49. Collins SR, Kemmeren P, Zhao X, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Krogan NJ. *Mol & Cell Proteomics: MCP* 2007, **6**(3):439–450. [http://www.ncbi.nlm.nih.gov/pubmed/17200106]. [PMID: 17200106]
50. Stark C, Breitkreutz B, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Auken KV, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M: **The BioGRID interaction database: 2011 update.** *Nucleic Acids Res* 2011, **39**(Database issue):D698—704. [http://www.ncbi.nlm.nih.gov/pubmed/21071413]. [PMID: 21071413]
51. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R: **ConsensusPathDB: toward a more complete picture of cell biology.** *Nucleic Acids Res* 2011, **39**(Database issue):D712–717. [http://www.ncbi.nlm.nih.gov/pubmed/21071422]. [PMID: 21071422]
52. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, Prinz J, Onge RPS, VanderSluis B, Makhnevych T, Vizeacoumar FJ, Alizadeh S, Bahr S, Brost RL, Chen Y, Cokol M, Deshpande R, Li Z, Lin Z, Liang W, Marback M, Paw J, Luis BS, Shuteriqi E, Dyk Nv, et. al: **The genetic landscape of a cell.** *Sci (New York, N.Y.)* 2010, **327**(5964):425–431. [http://www.ncbi.nlm.nih.gov/pubmed/20093466]. [PMID: 20093466]
53. Mazloom AR, Dannenfelser R, Clark NR, Grigoryan AV, Linder KM, Cardozo TJ, Bond JC, Boran ADW, Iyengar R, Malovannaya A, Lanz RB, Ma'ayan A: **Recovering protein-protein and domain-domain interactions from**

aggregation of IP-MS proteomics of coregulator complexes. *PLoS Comput Biol* 2011, **7**(12):e1002319. [http://www.ncbi.nlm.nih.gov/pubmed/22219718]. [PMID: 22219718]

54. Malovannaya A, Lanz RB, Jung SY, Bulynko Y, Le NT, Chan DW, Ding C, Shi Y, Yucer N, Krenciute G, Kim B, Li C, Chen R, Li W, Wang Y, O'Malley BW, Qin J: **Analysis of the human endogenous coregulator complexome.** *Cell* 2011, **145**(5):787–799. [http://www.ncbi.nlm.nih.gov/pubmed/21620140]. [PMID: 21620140]
55. Kamburov A, Stelzl U, Herwig R: **IntScore: a web tool for confidence scoring of biological interactions.** *Nucleic Acids Res* 2012, **40**(Web Server issue):W140-146. [http://www.ncbi.nlm.nih.gov/pubmed/22649056]. [PMID: 22649056]
56. Ravasz E. *Methods Mol Biol (Clifton, N.J.)* 2009, **541:**145–160. [http://www.ncbi.nlm.nih.gov/pubmed/19381526]. [PMID: 19381526]
57. Pu S, Wong J, Turner B, Cho E, Wodak SJ: **Up-to-date catalogues of yeast protein complexes.** *Nucleic Acids Res* 2009, **37**(3):825–831. [http://www.ncbi.nlm.nih.gov/pubmed/19095691]. [PMID: 19095691]
58. Benschop JJ, Brabers N, van Leenen, D, Bakker LV, van Deutekom, H W M, van Berkum, N L, Apweiler E, Lijnzaad P, Holstege FCP, Kemmeren P: **A consensus of core protein complex compositions for Saccharomyces cerevisiae.** *Mol Cell* 2010, **38**(6):916–928. [http://www.ncbi.nlm.nih.gov/pubmed/20620961]. [PMID: 20620961]
59. Oliver S: **Guilt-by-association goes global.** *Nature* 2000, **403**(6770):601–603. [http://www.ncbi.nlm.nih.gov/pubmed/10688178]. [PMID: 10688178]
60. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25–29. [http://www.ncbi.nlm.nih.gov/pubmed/10802651]. [PMID: 10802651]
61. Ahn Y, Bagrow JP, Lehmann S: **Link communities reveal multiscale complexity in networks.** *Nature* 2010, **466**(7307):761–764. [http://www.ncbi.nlm.nih.gov/pubmed/20562860]. [PMID: 20562860]
62. Schaefer MH, Fontaine J, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA: **HIPPIE: Integrating protein interaction networks with experiment based quality scores.** *PloS One* 2012, **7**(2):e31826. [http://www.ncbi.nlm.nih.gov/pubmed/22348130]. [PMID: 22348130]
63. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z, Brown GW, Ingles CJ, Emili A, Allis CD, Toczyski DP, Weissman JS, Greenblatt JF, Krogan NJ: **Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map.** *Nature* 2007, **446**(7137):806–810. [http://www.ncbi.nlm.nih.gov/pubmed/17314980]. [PMID: 17314980]
64. Pentchev K, Ono K, Herwig R, Ideker T, Kamburov A: **Evidence mining and novelty assessment of protein-protein interactions with the ConsensusPathDB plugin for Cytoscape.** *Bioinformatics (Oxford, England)* 2010, **26**(21):2796–2797. [http://www.ncbi.nlm.nih.gov/pubmed/20847220]. [PMID: 20847220]
65. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7**(11):120. [http://www.ncbi.nlm.nih.gov/pubmed/17147767]. [PMID: 17147767]
66. Resnik P: **Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language.** *J of Artif Intelligence Res* 1999, **11:**95–130. [http://citeseerx.ist.psu.edu/viewdoc/summary?doi:10.1.1.50.3785]
67. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S: **GOSemSim: an R package for measuring semantic similarity among GO terms and gene products.** *Bioinf (Oxford, England)* 2010, **26**(7):976–978. [http://www.ncbi.nlm.nih.gov/pubmed/20179076]. [PMID: 20179076]

doi:10.1186/1471-2105-13-262
Cite this article as: Kamburov *et al.*: Cluster-based assessment of protein-protein interaction confidence. *BMC Bioinformatics* 2012 **13**:262.