Identifying the topology of protein complexes from affinity purification assays

Caroline C. Friedel[*] and Ralf Zimmer
Institut für Informatik, Ludwig-Maximilians-Universität München,
Amalienstraße 17, 80333 München, Germany
Caroline.Friedel@bio.ifi.lmu.de

Abstract: Recent advances in high-throughput technologies have made it possible to investigate not only individual protein interactions but the association of these proteins in complexes. So far the focus has been on the prediction of complexes as sets of proteins from the experimental results while the modular substructure and the physical interactions within protein complexes have been mostly ignored. In this article, we present an approach for identifying the direct physical interactions and the subcomponent structure of protein complexes predicted from affinity purification assays. Our algorithm calculates the union of all maximum spanning trees from scoring networks for each protein complex to extract relevant interactions. In a subsequent step this network is extended to interactions which are not accounted for by alternative indirect paths. We show that the interactions identified with this approach are more accurate in predicting experimentally derived physical interactions than baseline approaches and resolve more satisfactorily the subcomponent structure of the complexes. The usefulness of our approach is illustrated on the RNA polymerases for which the modular substructure can be successfully reconstructed with our method.

1 Introduction

Cellular processes of all sorts are shaped by proteins associated in complexes. Thus, the identification of such complexes and the interactions within the complexes have become a major experimental focus. While direct, physical interactions can be identified by the yeast two-hybrid (Y2H) approach [FS89], affinity purification methods followed by mass spectrometry, such as tandem affinity purification (TAP) [RSR$^+$99], can also identify indirect interactions via other proteins in complexes. Recently, the TAP systems was applied by Gavin *et al.* [G$^+$06] and Krogan *et al.* [K$^+$06] to identify protein complexes in the yeast *Saccharomyces cerevisiae* on a genome-scale .

In the TAP system, epitope tagged proteins (baits) are expressed and purified in consecutive affinity columns [RSR$^+$99]. Proteins interacting directly or indirectly with the bait, so-called preys, are then co-purified with the bait and identified by mass spectrometry. Ideally, the purification of one bait would yield the complete protein complex the bait is involved in. However, due to large false positive and negative rates in the experiments,

[*] Corresponding author.

sophisticated methods are necessary to predict the actual complexes from the purification results.

The first predictions methods were developed by the groups of Gavin *et al.* [G^+06] and Krogan *et al.* [K^+06] themselves. Since the resulting complexes showed only relatively little agreement, advanced methods have been developed recently [PVE^+07, C^+07, HLM07, FKZ08] which improved predictive performance significantly. Here, most approaches use a two-step approach by first calculating interaction scores and then predicting the complexes from those scores. However, the focus so far has been on predicting sets of proteins associated in complexes and not the substructure of the complexes or the physical interactions within these.

A few methods have been developed which analyse the substructure of protein complexes. Aloy *et al.* [A^+04] used interactions from 3D structures and electron microscopy to at least partially resolve interactions between subunits of 54 experimentally derived complexes. The method of Hollunder *et al.* [HBW05, HBW07, HFB^+07] identifies substructures in protein complexes which occur more frequently in different complexes than expected at random. As a consequence, this approach can only identify substructures in protein complexes which occur in more than one complex. Gavin *et al.* [G^+06] distinguished between core elements and modules or attachments in their protein complex predictions but did not predict direct interactions.

Scholtens *et al.* [SVG05] and Bernard *et al.* [BVH07] modeled the physical topology of protein complexes using both affinity purification and Y2H results. However, Scholtens *et al.* used this only as an intermediate step in predicting protein complexes and did not evaluate the actual interactions they predicted. Bernard *et al.* showed that accurate predictions can be obtained with their approach but did not evaluate to what degree their results depend on the Y2H interactions used additionally.

In this article, we investigate if the topology of protein complexes can be predicted from the affinity purification results alone. Here, the topology of a protein complex describes both the direct physical interactions within a complex (the complex scaffold) but also its modular substructure, i.e. the subdivision of the complex into smaller components. Since most methods for predicting protein complexes from affinity purification results calculate interaction scores as an intermediate step, we developed a method to extract interactions relevant for the complex scaffold from these densely connected scoring networks.

Our algorithm calculates the union of all maximum spanning trees from the interaction scores for each complex. The maximum spanning trees are then extended heuristically by interactions which are not accounted for by alternative indirect interactions. We applied our method to confidence scores and protein complexes calculated with the Bootstrap method [FKZ08] from the yeast affinity purification experiments of Gavin *et al.* [G^+06] and Krogan *et al.* [K^+06]. We show that the interactions predicted by our approach are enriched for direct physical interactions determined by Y2H experiments. Furthermore, the distance in the resulting network reflects the functional and localization similarity of the corresponding proteins and the substructure of protein complexes can be resolved in a straightforward way.

2 Methods

In the following, let $C = \{C_1, \ldots, C_n\}$ be a set of protein complexes with C_i a set of proteins and $G = (V, E)$ a weighted network of interaction scores. Here, V is the set of all proteins and E the set of all interactions between them. In the following, we assume that all scores are confidence values in the range of 0 to 1. The function $w : E \rightarrow [0, 1]$ defines the weight, i.e. the confidence score, of each edge. Interactions not contained in the network are given a weight of 0. If the scoring method calculates general scores from $-\infty$ (or 0) to ∞, edge weights are scaled to $[0, 1]$.

Furthermore, we assume that each complex is connected in the network of actual physical interactions. This means that each protein can be reached from every other protein in the same complex by an indirect path of physical (direct) interactions. This network of direct interactions is denoted as the scaffold of the complex in the following. We perform predictions separately for each complex and, consequently, two interactions with the same weight may be predicted as direct in one complex and indirect via other proteins in another one depending both on the association strength within a complex and the existence of alternative paths in this complex.

2.1 Maximum spanning trees

For each complex, we start with a fully or almost fully connected scoring network for interactions between proteins in this complex. In this network, we want to identify a hierarchical subcomponent structure and, thus, not only the largest subcomplexes but also subcomponents of these subcomplexes. This hierarchical structure can be identified with hierarchical clustering algorithms.

The most commonly used variants of hierarchical clustering are average linkage and single linkage clustering. Average linkage uses the average score between two clusters to define their similarity which makes it difficult to assign the actual physical interactions. Single linkage uses the maximum score and, accordingly, the physical interactions can be defined in a straightforward way as the interactions providing the link between two clusters.

Single linkage effectively computes the maximum spanning tree of the network if the resulting dendrogram is unrooted. A spanning tree is a tree which connects all vertices in the network. The maximum spanning tree (MST) is the spanning tree which maximizes the sum of its edge weights and can be calculated efficiently in $O(|E| + |V| \log |V|)$ [CLRS00].

Because of this relationship between hierarchical clustering and MSTs, our algorithm for predicting the scaffold of a complex is based on calculating the MST of the corresponding network. If all interaction weights within the complex are distinct, the MST is unique. As this is generally not the case in scoring networks, many MSTs can exist. As a consequence, we calculate the set of direct interactions in the complex scaffold as the union of all possible MSTs.

To calculate all interactions contained in at least one MST, we do not have to compute all

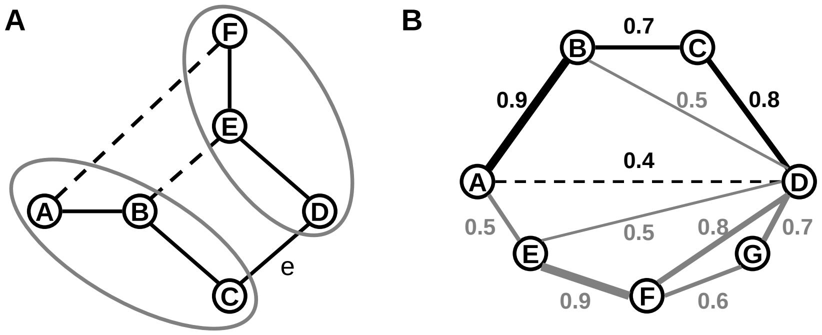

Figure 1: Figure **A** outlines how interactions contained in at least one MST are identified. The solid lines show the MST T calculated first. By removing edge e, a cut into the two sets $\{A, B, C\}$ and $\{D, E, F\}$ is created (grey ellipses). Dashed lines indicate edges crossing that cut with the same weight as e. By replacing e with any of these edges another MST T' is created. Thus, all of these edges are contained in at least one MST and added to the predicted scaffold. Figure **B** illustrates how MSTs are extended. For each possible edge (A, D), we find the optimal shortest path between the two nodes. In this case, this is $A \rightarrow B \rightarrow C \rightarrow D$ (black) which has a weight of $0.9 \cdot 0.7 \cdot 0.8 = 0.504$. Since the weight of edge (A, D) is smaller than this, the edge is discarded. If the weight of (A, D) were larger than 0.504, it would be added to the scaffold network.

possible MSTs, but can find the relevant interactions from one arbitrary MST (see Figure 1). Deleting each edge e in turn from this MST yields a cut *Cut*(T, e) – a partitioning into two sets – of the proteins in the complex. All edges crossing that cut, i.e. connecting proteins not in the same set, with the same weight as e are contained in at least one MST. With this algorithm, all edges in the union of all MSTs can be identified.

Thus, the algorithm for predicting the scaffold consists of two steps. First, an MST is calculated either with the Kruskal or Prim algorithm [Kru56, Pri57]. Second, all other edges contained in an MST are identified with the approach described above.

2.2 Extended MSTs

Although the combination of all MSTs is no longer a tree, the resulting networks are extremely sparse and many protein interactions may still be missing. As a consequence, we add a post-processing step in which we identify interactions which are not yet accounted for by an indirect interaction via other proteins in the MST scaffold. For this purpose, we compare an interaction (u, v) in the original network to the best indirect interaction between u and v in the current scaffold network. If the edge weight is at least as high as a factor α times the weight of the best indirect interaction, the interaction is added to the MST network. The resulting network is denoted as $eMST_\alpha$ and generally, α is set to 1.

For calculating the best indirect interaction we use the fact that all edge weights are confidence values in $[0, 1]$ and, thus, can be interpreted as probabilities. The weight of an indirect interaction is the probability of the optimal path between the corresponding proteins in the current scaffold (without the edge (u, v)). The probability is calculated as the product of the edge probabilities on this path and the optimal path is defined as the path with the highest probability. If we transform edge weights by taking the absolute values of the logarithms, the path with maximum probability is the path with the smallest sum of transformed edge weights. This optimal path between a pair of nodes can then be calculated using Dijkstra's algorithm for shortest paths [CLRS00].

To identify interactions which cannot be explained by a sequence of sufficiently strong indirect interactions, we process candidate interactions in the order of non-increasing edge weights. For each interaction e, we calculate the optimal alternative path P between the corresponding proteins in the current scaffold. The interaction e is added to the scaffold if $w(e) \geq \alpha w(P)$ and the scaffold is updated whenever a new interaction is identified. In the following, we show that this algorithm is correct for $\alpha \leq 1$. This means that there is no edge e in the final scaffold such that an alternative path P exists in the network with $w(e) < \alpha w(P)$.

Proof by contradiction: Assume, there exists such a path P for an edge e. Since the weight of each edge is ≤ 1, we have for each edge $f \in P$ that $w(P) \leq w(f)$. Thus, $w(e) < \alpha w(f) \, \forall f \in P$ and $w(e) < w(f) \, \forall f \in P$ if $\alpha \leq 1$. Thus, all edges on this path have been processed before e and this path was already contained in the network at the time e is added. This is a contradiction to the construction of the scaffold network.

2.3 Baseline prediction algorithms

We compare our algorithm against two baseline predictors. The complete approach predicts all interactions within the complex as direct, physical interactions. The connected approach calculates the network G_τ for each complex where $\forall e \in E_\tau : w(e) \geq \tau$ and τ the largest value such that G_τ is connected.

3 Results

The MST and extended MST approaches were applied to interaction scores and complex predictions calculated from the combined results of the genome-scale TAP experiments of Gavin *et al.* [G$^+$06] and Krogan *et al.* [K$^+$06] in yeast. Here, we used confidence scores and protein complexes predicted with the unsupervised Bootstrap approach we presented recently [FKZ08]. These confidence scores are more accurate than any other scoring method. Furthermore, the medium (BT-409) and high confidence (BT-217) Bootstrap complexes are of the same quality as the best supervised predictions and manually curated protein complexes, respectively.

All bootstrap confidence scores are between 0 and 1 and the original network contains 62,876 interactions. By restricting this to interactions within BT-409 complexes (the complete approach), we obtained 9,918 interactions (15.8% of the original set). The connected approach yielded 5,404 interactions (8.6%), the MST approach 1,658 interactions (2.6%) and the extended MST approach (with $\alpha = 1$) 3,085 interactions (4.9%).

3.1 Reference interactions

To compile a reference set of direct interactions we extracted all yeast protein-protein interactions from the DIP database [SMS$^+$04] determined with the Y2H method. We chose Y2H interactions to make sure that only direct physical interactions are contained in the reference set. Furthermore, we used the genome-scale Y2H results for yeast from the studies of Uetz et al. [U$^+$00] and Ito et al. [I$^+$01]. For the Ito dataset, both the high confidence and complete set were evaluated.

Since large Y2H interaction sets are also available for *Drosophila* [G$^+$03], *C. elegans* [L$^+$04] and human [S$^+$05, R$^+$05], we used orthology assignments from the Inparanoid database [BSOS08] to map these interactions onto yeast. Interactions were mapped if both interaction partners had orthologs in yeast. This resulted in 575 predicted interactions from *Drosophila* to yeast, 170 from *C. elegans* to yeast (70 from the core set defined by Li et al.) and 220 predicted interactions from human to yeast.

Table 1 shows a comparison of the Y2H interaction networks against the BT-409 and BT-217 complexes and manually curated complexes from the MIPS database [M$^+$04]. The first row for each combination indicates the enrichment of Y2H interactions within complexes. Enrichment is calculated as $p_C/p_{\overline{C}}$ where

$$p_C = \frac{|E_C| \cap |E_{Y2H}|}{|E_C|} \quad \text{and} \quad p_{\overline{C}} = \frac{|E_{\overline{C}}| \cap |E_{Y2H}|}{|E_{\overline{C}}|}. \tag{1}$$

Here, E_C is the set of interactions within complexes, $E_{\overline{C}}$ the set of interactions between proteins in different complexes and E_{Y2H} the set of Y2H interactions. The second row of Table 1 specifies the fraction of Y2H interactions contained within complexes.

Complex set	**Y2H interaction network**					
	DIP	**Uetz**	**Ito**	**Ito core**	**All yeast**	**Yeast + Pred.**
MIPS	53.0 [0.06]	106.9 [0.07]	33.9 [0.03]	195.6 [0.09]	52.1 [0.05]	41.4 [0.05]
BT-409	64.5 [0.07]	152.4 [0.14]	53.1 [0.05]	192.7 [0.18]	64.2 [0.07]	59.8 [0.07]
BT-217	75.1 [0.05]	150.9 [0.1]	65.1 [0.03]	205.2 [0.12]	75.2 [0.05]	66.8 [0.05]

Table 1: This table shows for each Y2H network the enrichment (see equation 1) of Y2H interactions within the MIPS, BT-409 and BT-217 complexes. The second row for each combination of network and complex set specifies the fraction of Y2H interactions within protein complexes.

As can be seen, Y2H interactions are significantly enriched within protein complexes and the enrichment values appear to reflect at least partly the confidence of the corresponding Y2H set. The Ito core interactions have much higher enrichment values than the less confident complete Ito set. When adding the less confident predicted interactions to the complete yeast network a less distinctive but still considerable decrease in enrichment can be observed. Interestingly, the enrichment is significantly higher in the Bootstrap complexes than in the MIPS complexes. The average bootstrap score of Y2H interactions in complexes (0.76) is significantly higher than for interactions in complexes not confirmed by Y2H (0.44). Unfortunately, the fraction of interactions in the Y2H network which actually connect proteins in the same complex is very small.

3.2 Assessing the predictive accuracy of complex scaffolds

We evaluated the predictive accuracy of the presented methods using *receiver operating characteristic* (ROC) curves [Faw06]. For this purpose, true positive rates are plotted against false positive rates with decreasing thresholds for predicting a direct interaction. In this case, true positive rate is the fraction of Y2H interactions within the BT-409 complexes recovered by the prediction methods. False positive rate is the fraction of interactions within the BT-409 complexes predicted to be in the scaffold but not contained in the Y2H network.

Figure 2 **A** shows the ROC curve for the complete, connected, MST and extended MST predictions compared against the complete set of yeast Y2H interactions. Similar results can be observed for all Y2H sets. As can be clearly seen, significant improvements in predictive accuracy can be obtained with the MST approach. At a maximum true positive

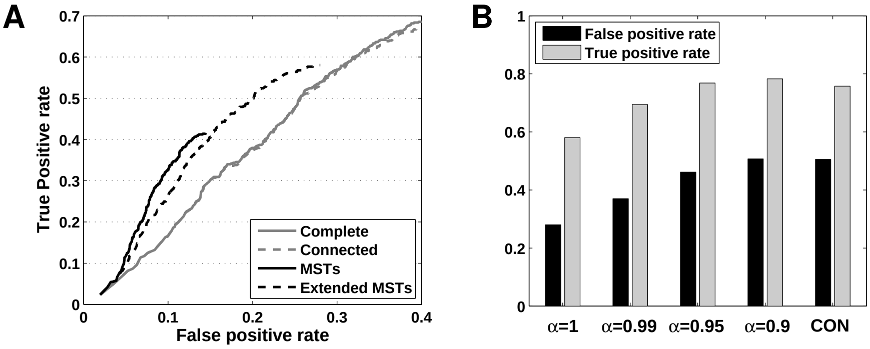

Figure 2: ROC curve (**A**) for the direct interactions predicted by the complete, connected, MST and extended MST (for $\alpha = 1$) approach compared to all yeast Y2H interactions within the BT-409 complexes. Here, the curves for the complete and connected approach are almost identical in this range as 90% of the top scoring interactions of the complete network are also contained in the connected network. The two networks differ mostly in the low scoring and low quality interactions contained additionally in the complete network. Figure **B** illustrates true positive and false positive rates for decreasing values of α and the connected networks (CON).

rate of 41.6%, only 14.5% false positives are predicted. At the same true positive rate, about 22% false positives are predicted by the complete and connected predictions.

However, the higher specificity of the MST approach results in a significantly lower sensitivity. Thus, less than half of the Y2H interactions recovered by the baseline predictions are recovered by the MST approach. By extending the MSTs, the fraction of true positives identified can be increased significantly. Although the false positive rate consequently increases as well, the overall performance of the extended MSTs is nevertheless significantly better than observed for the baseline predictions.

Figure 2 **B** illustrates the true and false positive rates for decreasing values of α used for extending the MSTs. The more conditions are relaxed for extending the networks, the more interactions are added. As a consequence, more true interactions are recovered but also more wrong predictions are made. Nevertheless, at the same false positive rate the extended MSTs can recover more true positives than the connected networks.

3.3 Separation of substructures within complexes

By predicting the topology of protein complexes, we aim to identify substructures within complexes. Proteins which are closely involved with each other should end up very close to each other in the network, i.e. separated by only few interactions. Proteins more distantly related, on the other hand, should be separated by many interactions within the network. To measure the distance of two proteins in the network we calculate the number of interactions on the shortest (unweighted) path between them.

Figure 3 **A** illustrates the correlation between the distance of two proteins and the confi-

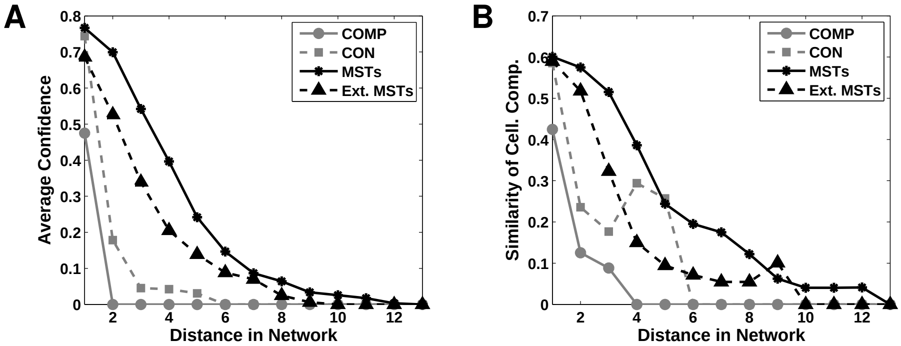

Figure 3: Figure **A** compares the distance between a protein pair in the complete (COMP), connected (CON), MST and extended MST networks against the confidence of this pairwise interaction in the complete network. Averages are taken over all protein pairs with the same distance. Figure **B** compares the distance in the network against the fraction of GO cellular compartment annotations the two protein have in common.

dence of the corresponding interaction in the original bootstrap network. As expected, the more interactions the predicted network contains the shorter are the distances in the network. Thus, largest distances are observed in the MST network and shortest distances in the complete network where most proteins are directly connected. Furthermore, the higher the confidence of an interaction between two proteins is, the smaller is the distance in the resulting scaffold network. Due to the larger distances in the MST and extended MST networks, the rate of decrease is significantly smaller for these networks which allows for a better resolution of the network structure.

We calculated for each pair of proteins the fraction of Gene Ontology (GO) [A+00] annotations they have in common and correlated this with their distance. We used this simple measure instead of more complicated methods as semantic similarity [SDRL06] because semantic similarity within complexes is generally very high [FKZ08]. In this case, the simple overlap measure allows for a more fine-grained analysis of protein function and localization within complexes.

Figure 3**B** shows the results for the cellular component ontology of the GO. Similar results were observed for the biological process and molecular function ontologies. As with interaction confidences, we observe that the similarity of the cellular component assignments tends to decrease with the distance between the corresponding proteins. Furthermore, the rate of decrease is lowest for the sparse MST network. In the extended MSTs, this rate is significantly higher but still by far not as high as in the connected and complete networks.

This indicates that proteins involved in different subcomponents of a complex are separated from each other by many interactions in the predicted scaffolds, whereas proteins involved in the same subcomponents are close to each other. Surprisingly, co-localization scores increase again at a distance of 4 for the connected network and at a distance of 9 for the extended MST network. This is due to the small number of protein pairs with this distance in the corresponding networks. Thus, outliers affect the average co-localization more strongly.

3.4 Analysis of the DNA-directed RNA polymerase complex

To illustrate the value of the predicted interaction scaffolds for the identification of substructures in complexes, we analyzed the DNA-directed RNA polymerase complex. This complex contains 46 proteins in the BT-409 set and effectively consists of three separate RNA polymerase complexes (RNA polymerase I, II and III) which have been clustered into one complex since they have many proteins in common. The crystal structure of polymerase II is known, whereas only little structural information is available for polymerases I and III [CAB+08].

Figure 4 shows the complex connections for the RNA polymerase in the connected and extended MST network. The complex was visualized using the organic layout function of Cytoscape [SMO+03] which clusters closely connected proteins together. In the complete bootstrap network, no substructure can be observed but all proteins form a tight cluster. In the connected network (see Figure 4 **A**), we observe at least a separation between the

RNA polymerase III complex and the remaining proteins but polymerase I and II are too tightly connected to identify the substructure. It is only when proteins are colored by their cellular components that we detect that proteins from the same subcomponents are clustered together.

In the extended MST network (see Figure 4 **B**) the subdivision of the complex into polymerase complexes I, II and III can be clearly observed. The polymerase III complex (light grey) is connected by two proteins (RPC19, RPC5) to the polymerase I complex (dark grey). The latter one is then connected to the polymerase II complex (black) by a group of five proteins (RPB5, RPB6, RPB8, RPB10 and RPB12) contained in all three RNA polymerase complexes.

Interestingly, these five proteins are not directly connected to the other polymerase III proteins although they are contained in this complex. If we relax the criterion for extending an MST ($\alpha = 0.99$), the interaction between RPB10 and RPC5, which has also been identified in Y2H screens [LCST93, FBG$^+$99], is added to the scaffold. This might suggest that the interaction of the common proteins to polymerase III is mediated via this interaction. However, if we look at the crystal structure of polymerase II and the model for polymerase III [CAB$^+$08], we find that none of the common proteins are actually in physical contact in the complexes (possibly apart from RPB10 and RPB12).

Going back to the original purification experiments, we find that of the 7 interactions predicted between the common proteins, 6 interactions are bait-prey interactions which have been found to be very reliable [BCRC04] and 3 of those are identified in both directions. Since the proteins do not appear to physically interact, this is probably a consequence of the common occurrence of these proteins in several different complexes.

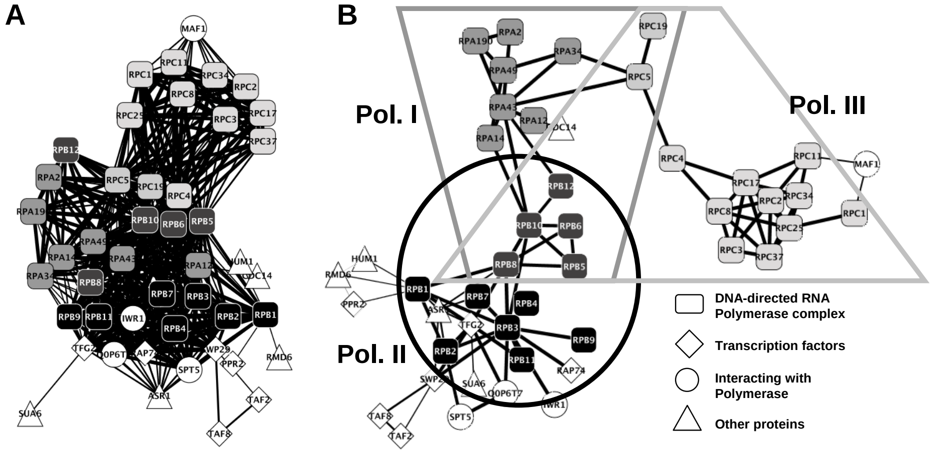

Figure 4: This figure shows the predicted subnetworks for the DNA-directed RNA complex. Figure **A** shows the results for the connected subnetwork and **B** the results for extended MST network. Colors indicate the subcomponents: Polymerase complexes I (dark grey), II (black) and III (light grey). Rectangles denote polymerase proteins and diamonds transcription factors.

Another example which illustrates the problems of affinity purification in distinguishing the actual physical interactions, is the interaction between the RPB3 and RPB9 protein in the polymerase II subcomplex. Although the two proteins are located at different ends of the Polymerase II in the 3D structure, this interaction has a very high confidence score as RPB3 and RPB9 co-purified each other whenever one of these proteins was used as bait (6 times for RPB3 and 5 times for RPB9).

4 Discussion

In this article, we presented an approach for predicting the topology of protein complexes, i.e. the scaffold of direct interactions which spans the complex. First, our method calculates the union of all maximum spanning trees (MSTs) in the interaction score network for a protein complex. In a subsequent step, this network is iteratively extended by interactions which cannot be explained by a path of alternative indirect interactions. The MST approach is applicable to all weighted interaction networks and in particular to interaction scores calculated from affinity purification assays with any of the recently published scoring methods. Confidence scores which are required for extending the MSTs in our algorithm, can be obtained by scaling any type of scores to $[0, 1]$ or using the Bootstrap approach we developed to calculate scores from affinity purification experiments.

Predictive performance of subnetworks calculated from Bootstrap confidence scores was evaluated on experimentally determined direct, physical interactions from Y2H experiments. We showed that predictive accuracy can be increased significantly with our approach compared to baseline predictions. When comparing the individual protein complexes to the Y2H network, we observed that less than half of the complexes both in the predicted complex set and in manually curated complexes contain at least one Y2H interaction, and only 5% to 11% of the complexes are actually non-trivially connected (i.e. they are connected and contain more than two proteins) in the Y2H network. This suggests that many of the direct interactions within complexes have not been identified yet. Here, the interactions predicted by our approach but not found in the Y2H network are promising starting points for experimental verification.

Protein complexes are not simply clumps of proteins but they have an internal substructure in which not all proteins bind closely together. Thus, proteins in the same subcomplex are closely connected by short paths of direct interactions whereas proteins in different subcomponents are separated by many physical interactions in this network. Our results show that both for the network predicted with our approach and for the baseline predictors, the distance between protein pairs is correlated strongly with the corresponding interaction confidence and the similarity of the cellular components these proteins are contained in. However, in the scaffold network predicted by our method, separation of proteins in different subcompartments of a complex is more distinctive and thus, the substructure of the complex can be better resolved.

We illustrated this observation on the complex of DNA-directed RNA polymerases. While the substructure of the complex with three different RNA polymerases can only be partly observed in the baseline predictions, it is clearly evident in the network predicted with our approach. By relaxing the conditions of our algorithm slightly, the substructure of the complex can be further emphasized and important interactions can be identified. Thus, the algorithm presented in this article is valuable for identifying the scaffold of physical protein interactions within complexes as well as their subcomponent structure.

References

[A$^+$00] M. Ashburner et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.

[A$^+$04] Patrick Aloy et al. Structure-based assembly of protein complexes in yeast. *Science*, 303(5666):2026–2029, Mar 2004.

[BCRC04] Joel S Bader, Amitabha Chaudhuri, Jonathan M Rothberg, and John Chant. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, 22:78–85, Jan 2004.

[BSOS08] Ann-Charlotte Berglund, Erik Sjölund, Gabriel Ostlund, and Erik L L Sonnhammer. InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res*, 36(Database issue):D263–D266, Jan 2008.

[BVH07] Allister Bernard, David S. Vaughn, and Alexander J. Hartemink. Reconstructing the Topology of Protein Complexes. In *Proceedings of the 11th Annual International Conference on Research in Computational Molecular Biology, RECOMB 2007, Oakland, CA, USA, April 21-25*, pages 32–46, 2007.

[C$^+$07] Sean R Collins et al. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics*, 6(3):439–450, Mar 2007.

[CAB$^+$08] P. Cramer, K.-J. Armache, S. Baumli, S. Benkert, F. Brueckner, C. Buchen, G.E. Damsma, S. Dengl, S.R. Geiger, A.J. Jasiak, A. Jawhari, S. Jennebach, T. Kamenski, H. Kettenberger, C.-D. Kuhn, E. Lehmann, K. Leike, J.F. Sydow, and A. Vannini. Structure of Eukaryotic RNA Polymerases. *Annual Review of Biophysics*, 37(1):337–352, 2008.

[CLRS00] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, 2nd edition*. MIT Press, McGraw-Hill Book Company, 2000.

[Faw06] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[FBG$^+$99] A. Flores, J. F. Briand, O. Gadal, J. C. Andrau, L. Rubbi, V. Van Mullem, C. Boschiero, M. Goussot, C. Marck, C. Carles, P. Thuriaux, A. Sentenac, and M. Werner. A protein-protein interaction map of yeast RNA polymerase III. *Proc Natl Acad Sci U S A*, 96(14):7815–7820, Jul 1999.

[FKZ08] Caroline C. Friedel, Jan Krumsiek, and Ralf Zimmer. Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast. In *Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology, RECOMB 2008, Singapore, March 30 - April 2*, pages 3–16, 2008.

[FS89] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, Jul 1989.

[G+03] L. Giot et al. A protein interaction map of Drosophila melanogaster. *Science*, 302:1727–36, Dec 2003.

[G+06] Anne-Claude Gavin et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–6, Mar 2006.

[HBW05] Jens Hollunder, Andreas Beyer, and Thomas Wilhelm. Identification and characterization of protein subcomplexes in yeast. *Proteomics*, 5(8):2082–2089, May 2005.

[HBW07] Jens Hollunder, Andreas Beyer, and Thomas Wilhelm. Protein subcomplexes–molecular machines with highly specialized functions. *IEEE Trans Nanobioscience*, 6(1):86–93, Mar 2007.

[HFB+07] Jens Hollunder, Maik Friedel, Andreas Beyer, Christopher T Workman, and Thomas Wilhelm. DASS: efficient discovery and p-value calculation of substructures in unordered data. *Bioinformatics*, 23(1):77–83, Jan 2007.

[HLM07] G. Traver Hart, Insuk Lee, and Edward Marcotte. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, 8(1):236, Jul 2007.

[I+01] T. Ito et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, 98:4569–74, Apr 2001.

[K+06] Nevan J Krogan et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440:637–43, Mar 2006.

[Kru56] J. B Kruskal. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proc. Amer. Math. Soc.*, 7:48–50, 1956.

[L+04] Siming Li et al. A map of the interactome network of the metazoan C. elegans. *Science*, 303:540–3, Jan 2004.

[LCST93] D. Lalo, C. Carles, A. Sentenac, and P. Thuriaux. Interactions between three common subunits of yeast RNA polymerases I and III. *Proc Natl Acad Sci U S A*, 90(12):5524–5528, Jun 1993.

[M+04] H. W. Mewes et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, 32(Database issue):D41–D44, Jan 2004.

[Pri57] R. C. Prim. Shortest connection networks and some generalisations. *Bell System Technical Journal*, 36:1389–1401, 1957.

[PVE+07] Shuye Pu, Jim Vlasblom, Andrew Emili, Jack Greenblatt, and Shoshana J Wodak. Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics*, 7(6):944–960, Mar 2007.

[R+05] Jean-François Rual et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–8, Oct 2005.

[RSR+99] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Sraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–1032, Oct 1999.

[S+05] Ulrich Stelzl et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122:957–68, Sep 2005.

[SDRL06] Andreas Schlicker, Francisco S Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7:302, 2006.

[SMO+03] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003.

[SMS+04] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449–D451, Jan 2004.

[SVG05] Denise Scholtens, Marc Vidal, and Robert Gentleman. Local modeling of global interactome networks. *Bioinformatics*, 21(17):3548–3557, Sep 2005.

[U+00] P. Uetz et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403:623–7, Feb 2000.