Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast

Caroline C. Friedel*, Jan Krumsiek, and Ralf Zimmer

Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstraße 17, 80333 München, Germany
Caroline.Friedel@bio.ifi.lmu.de

Abstract. Protein interactions and complexes are major components of biological systems. Recent genome-wide applications of tandem affinity purification (TAP) in yeast have increased significantly the available information on such interactions. From these experiments, protein complexes were predicted with different approaches first from the individual experiments only and later from their combination. The resulting predictions showed surprisingly little agreement and all of the corresponding methods rely on additional training data. In this article, we present an unsupervised algorithm for the identification of protein complexes which is independent of the availability of additional complex information. Based on a bootstrap approach, we calculated intuitive confidence scores for interactions which are more accurate than previous scoring metrics. The complexes determined from this confidence network are of similar quality as the complexes identified by the best supervised approaches. Despite the similar quality of the latest predictions and our predictions, considerable differences are still observed between all of them. Nevertheless, the set of consistently identified complexes is more than four times as large as for the first two studies. Our results illustrate that meaningful and reliable complexes can be determined from the purification experiments alone. As a consequence, the approach presented in this article is easily applicable to large-scale TAP experiments for any organism.

1 Introduction

Cellular processes are shaped by proteins physically associated in complexes. Accordingly, significant efforts are put into the experimental identification of such protein interactions. Commonly used techniques are yeast two-hybrid (Y2H) [1,2] and affinity purification followed by mass spectrometry (e.g. Co-immunoprecipitation (Co-IP)[3] or tandem affinity purification (TAP)[4,5,6]). Since new methods are generally first applied to the yeast *Saccharomyces cerevisiae*, its interactome is the most thoroughly studied one.

While Y2H identifies only direct physical interactions, affinity purification can also identify indirect interactions in protein complexes. In the so far only

* Corresponding author.

M. Vingron and L. Wong (Eds.): RECOMB 2008, LNBI 4955, pp. 3–16, 2008.

genome-scale studies on complexes, the more effective TAP system [7] was applied to yeast separately by Gavin et al. [5] and Krogan et al. [6]. In the first experiment, 1,993 distinct TAP-tagged proteins (baits) were purified successfully and 2,760 distinct proteins (preys) identified in these purifications. In the second experiment, 2,357 baits were purified and 4,087 preys identified.

Because of the large size of these data sets, sophisticated methods were developed in both studies to infer individual protein complexes from the purification data. However, the resulting complexes showed surprisingly little overlap (see results). After the publication of the original results, improved prediction methods were developed by Pu et al. [8] based on the scoring method of Collins et al. [9] and by Hart et al. [10]. These methods used the data from both purification experiments.

A comparison of the different approaches outlines the important steps in complex determination (see Figure 1 **A**). First, purification experiments have to be combined for higher prediction quality. Second, propensities for individual protein interactions have to be determined from the purifications. Third, these propensities should be converted to confidence values assessing the likelihood of each protein interaction. In a fourth step, the network has to be clustered to obtain individual complexes after restricting it to the most confident interactions. If the corresponding clustering method only produces disjoint complexes, a final step has to be included to identify proteins shared between complexes. This is necessary because in biological systems proteins can be involved in more than one complex. Apart from the combined method by Collins et al. [9] and Pu et al. [8], all previous approaches leave out at least one of these steps. Figure 1 **A** gives an overview on which methods implement which steps in a supervised or unsupervised way.

All previous approaches rely more or less heavily on the availability of additional training data for at least one step. For yeast, hand-curated complexes can be taken from MIPS [11] and the study of Aloy et al. [12]. Furthermore, complexes can be automatically extracted from GO annotations [13]. Unfortunately, the resulting complexes are of lower quality than the hand-curated ones (see results). Recently, hand-curated complexes have become available for human and other mammals which cover about 12% of the protein-coding genes in human [14]. For other organisms, such complex information is not available which limits the applicability of the supervised approaches significantly.

Even if reference sets are available as for yeast, a large fraction of them have to be set aside as independent test sets to evaluate the quality of predicted complexes. Although most of the above mentioned approaches distinguished between test and training set by choosing one of the yeast reference sets for training and a different one for testing, these sets are not sufficiently disjoint to guarantee a reliable performance estimate. Indeed, generally more than half of the complexes in each reference set have a significant overlap to at least one complex in each of the other sets.

In this article, we present an unsupervised algorithm for the identification of protein complexes from the purification data alone which implements all steps

A

	GA	KR	CP	HA	BT
Experiment combination	-	-	S	U	U
Propensity calculation	U	S	S	U	U
Confidence values	-	S	S	-	U
Clustering	S	S	S	S	U
Shared Proteins	S	-	S	-	U

B

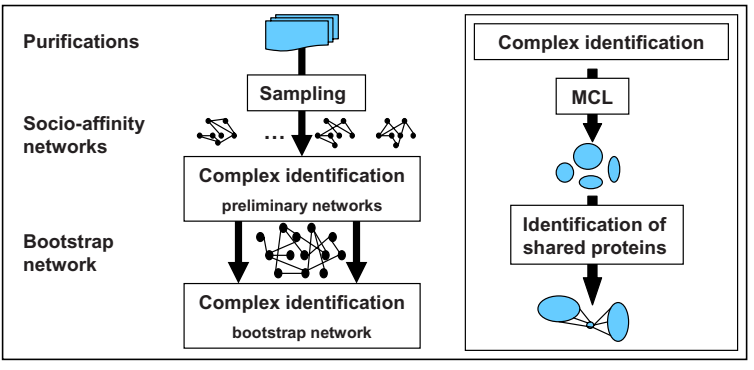

Fig. 1. Table **A** lists the major steps involved in complex identification and if they are realized by the approaches of Gavin et al. [5] (GA), Krogan et al. [6] (KR), Collins et al. and Pu et al. [9,8] (CP), Hart et al. [10] (HA) and the approach described in this article (BT). Furthermore, we indicate if a supervised approach based on additional training data is used (S, red) or an unsupervised approach (U, green). Figure **B** illustrates the major steps of the bootstrap approach described in this article. See methods for more details.

described above. Since no additional information on protein complexes is required, our approach is not limited to yeast but can be applied easily to large-scale purification experiments for other organisms.

Our results show that this approach is equivalent to the best supervised methods both with regard to functional and localization similarity observed in the resulting complexes as well as predictive performance with regard to known yeast complexes. We find that our predictions and the Pu and Hart predictions are much more consistent than the original Gavin and Krogan complexes. However, although the latest predictions are of similar quality, significant divergences between the predicted complexes are still observed. This clearly illustrates the need for further investigations of the individual protein complexes.

2 Methods

The algorithm we propose for the unsupervised identification of protein complexes implements all five necessary steps (see Figure 1 **A**). Purification experiments

are combined by pooling them. Interaction confidences are determined by first identifying preliminary complexes for bootstrap samples from the set of purifications [15,16]. The resulting confidence network is then clustered with the MCL algorithm [17] and proteins shared between complexes are identified in a post-processing step. Here, all necessary parameters are tuned based on intrinsic measures calculated from the networks and complexes alone. The last two steps are also used to determine the preliminary complexes for the bootstrap samples (see Figure 1 **B**). Details of the algorithm are described in the following.

2.1 Bootstrap Sampling

To determine reliable confidence scores, the bootstrap technique [15,16] is used to estimate how stable interactions are under perturbations of the data. A similar approach is utilized successfully for assigning confidence to phylogenetic trees [18]. For this purpose, 1,000 bootstrap samples are created from the set of N purifications. A purification consists of one bait protein and its preys in this purification. To create one bootstrap sample, N purifications are drawn with replacement. Each purification can be contained in a bootstrap sample once, multiple times or not at all. Multiple copies of the same purification are treated as separate purifications.

For each bootstrap sample, we then calculate preliminary interaction propensities in the form of socio-affinity scores as described by Gavin et al. [5]. These scores compare the number of co-occurrences of two proteins against the random expectation using a combination of spoke and matrix model [19]. No additional training data is required to compute them from a set of purifications. From these preliminary networks, we then determine preliminary complex predictions for each individual bootstrap sample with the algorithm described in the following section. To allow for reasonably fast computation, only edges are included in the preliminary networks whose weight exceeds a certain threshold τ.

2.2 Identification of Protein Complexes

The algorithm for the prediction of complexes from a network consists of two steps: clustering of the network and subsequent identification of shared proteins.

Clustering. Networks are clustered using the Markov clustering algorithm (MCL) developed by van Dongen [17]. In a recent study [20], this algorithm was found to be superior to a selection of other graph clustering algorithms for the identification of protein complexes. Apart from the approach of Gavin et al., all previous approaches to complex identification from purification data use this method. The running-time of the MCL procedure is in $O(Nk^2)$ for a network with N nodes and a maximum degree of k. Thus, it is reasonably fast for sparse networks. Its most important parameter is the *inflation parameter* which influences the granularity of the identified clusters, i.e. their size and number. The higher the inflation parameter, the smaller are the resulting clusters and the more clusters are identified.

All previous approaches used additional training data in the form of known complexes to chose the optimal inflation parameter. In this article, we suggest to use an intrinsic measure which compares the resulting clusters against the original network. For this purpose, we use a performance measure for graph clustering proposed by van Dongen, the so-called *efficiency*. Details for the calculation of efficiency can be found in [17]. Basically, a clustering is highly efficient if proteins in the same cluster are connected by edges with high weights and proteins in different clusters have no or only low weight connections.

To determine the optimal inflation parameter, we clustered the socio-affinity networks for each sample using several, gradually increasing inflation parameters. For each inflation parameter we then calculated the average efficiency over all samples. We found that the efficiency always reaches a maximum for a certain inflation parameter and decreases on either side of this value if the socio-affinity networks were reasonably sparse. Accordingly, the optimal inflation parameter is chosen as the one with the highest average efficiency across all samples. This inflation parameter is then used to cluster the preliminary networks for the bootstrap samples.

Identification of shared proteins. The MCL algorithm, as most clustering methods, identifies only disjoint clusters. In real biological systems, however, proteins can be contained in more than one complex. If a protein has such strong associations with two complexes, the MCL procedure will either cluster those two complexes together or, if further associations between the complexes are missing, cluster this protein with only one of these complexes. We address this problem in a similar way as Pu et al. [8] by post-processing the clusters obtained from the MCL algorithm. Contrary to them, we do not optimize this step based on proteins shared between known yeast complexes, but again use an intrinsic measure based on the original network.

The following criteria are used for adding shared proteins. First, a protein is only added to another complex if it has sufficiently strong interactions to this complex. Second, the tighter the associations within this complex, the stronger have to be the interactions of the protein to this complex. Third, for large complexes strong interactions are only required to some of the complex proteins or, alternatively, weaker interactions to most of them. As a consequence, a protein p_i can be added to a complex C if

$$s(p_i, C) \geq \alpha \cdot s(C) \cdot \left(|C|^{-\gamma} / \, 2^{-\gamma} \right) \tag{1}$$

with $s(p_i, C)$ the average interaction score of p_i to proteins of C and $s(C)$ the average interaction score within the complex. Interactions not contained in the network are given a weight of zero.

The threshold for adding a protein to a complex decreases both with complex size and with complex confidence. To control the influence of complex size, a power function was chosen since it decreases steeply at first but then levels off for larger values. The power function is normalized to yield 1 for complexes of size 2. In this case the threshold depends only on the strength of the interaction between the two proteins.

This threshold definition has two parameters, α and γ. Here, α defines how much weaker than $s(C)$ the connections to the complex are allowed to be and γ controls to which fraction of the complex sufficiently strong interactions have to exist. Both parameters are set such that the weighted average score over all complexes after the post-processing is at least as high as a fraction λ of the original average score. For this purpose, α is set to λ and γ to the largest possible value such that this requirement is still met. Here, λ was set to 0.95 to add proteins only to complexes to which they have a strong association. Note that proteins are added to complexes in parallel. Accordingly, the complex memberships and the average complex score are not updated until all proteins have been processed and the result does not depend on the order of the proteins.

2.3 Calculation of Confidence Scores and Final Complexes

The final confidence scores are then determined by calculating the so-called bootstrap network from the complexes identified for each bootstrap sample. In the bootstrap network, two proteins are connected by an edge if they are clustered together in at least one sample. The fraction of samples for which they are contained in the same complex provides the weight for the corresponding edge and the confidence for this association (between 0 and 1).

Final complexes are then obtained by applying the complex identification algorithm on this bootstrap network. For this purpose, the optimal inflation parameter determined in the previous step is chosen. No threshold is applied to the network before complex identification but the size of the network is limited beforehand by choosing stringent cut-offs for the preliminary socio-affinity networks.

More confident predictions can be obtained from the original complexes in the following way. First, all edges are removed from the network with weight lower than a specific threshold and then connected components are calculated for each complex separately in this restricted network. As a consequence, complexes can either shrink, be subdivided or be removed completely. This approach is more efficient than the alternative approach of first restricting the network and then repeating complex identification but yields almost identical complexes (see results).

2.4 Criteria for the Evaluation of Complex Quality

Functional co-annotation within complexes. Since protein complexes are formed to carry out a specific function, the functions of proteins in the same complex should be relatively homogeneous. We evaluate the functional similarity between proteins predicted to be in the same complex by using the protein annotations of the Gene Ontology (GO) [13]. The functional similarity of two proteins is quantified in terms of the *semantic similarity* of GO terms annotated to these proteins. Several variations of semantic similarity have been described [21,22,23,24]. Here, we use the relevance similarity proposed recently by Schlicker

at al. [24]. This measure is based both on the closeness of two GO terms to their common ancestors as well as the level of detail of these ancestors.

The GO score of a complex is the average relevance similarity of all protein pairs in this complex. The GO score for a set of complexes is the weighted mean over all complex scores and determined separately for the "biological process" and "molecular function" taxonomies. The final co-annotation score is then calculated as the geometric mean of the two values.

Co-localization within complexes. Since complexes can only be formed if the corresponding proteins are actually located together in the cell, a second quality measure is based on the similarity of protein localizations within a complex. For this purpose, we used the localization assignments and categories determined experimentally in yeast by Hu et al. [25].

The co-localization score for a complex is defined as the maximum fraction of proteins in this complex which are found at the same localization. The average co-localization score is calculated as the weighted average over all complexes and is defined as

$$L = \frac{\sum_j \max_i l_{i,j}}{\sum_j |C_j|} \tag{2}$$

Here, l_{ij} is the number of proteins of complex C_j assigned to the localization group i and $|C_j|$ is the number of proteins in the complex C_j with localization assignments.

Sensitivity and positive predictive value. To evaluate the accuracy of the predictions, *sensitivity* (Sn) and *positive predictive value* (PPV) were calculated with regard to the following reference sets: (a) hand-curated complexes from MIPS [11] (214 complexes after removing redundant complexes) and Aloy et al. [12] (101 complexes) as well as (b) complexes extracted from the SGD database [26] based on GO annotations (189 complexes). To compile the SGD set, GO-slim complex annotations to all yeast genes were taken from the SGD ftp site.

We used the definition of sensitivity and PPV for protein complexes by Brohee and van Helden [20]. Both measures are calculated from the number $T_{i,j}$ of proteins shared between a reference complex R_i and a predicted complex C_j:

$$Sn = \frac{\sum_i \max_j T_{i,j}}{\sum_i |R_i|} \quad \text{and} \quad PPV = \frac{\sum_j \max_i T_{i,j}}{\sum_j \sum_i T_{i,j}}. \tag{3}$$

3 Results

Bootstrap confidence values were calculated from the combined Krogan and Gavin purification experiments. This combined set contains 6498 purifications with 2995 distinct baits, more than half of which (1617) were purified more than once. On average, separate purifications of the same bait agree in about 27% of the retrieved preys between the two experiments. This is similar to the agreement between purifications of the same bait within the Krogan data set, but significantly smaller than within the Gavin set.

From these purifications the final bootstrap network was calculated. Only socio-affinity scores of at least 8 were included in the preliminary networks for the bootstrap samples. We chose a more stringent threshold than the one recommended by Gavin et al. for two reasons. First, the preliminary networks are much denser for the combined data than for the Gavin data alone and, as a consequence, the runtime of the MCL algorithm is increased considerably. Second, the final bootstrap network contains many more interactions than each individual preliminary network (in our case 20 time as many). Thus, the more stringent threshold both reduces runtime of the bootstrapping step and at the same time limits the size of the resulting bootstrap network.

The final bootstrap network contains 62,876 interactions between 5195 distinct proteins. Protein complexes were then determined from the bootstrap network with our method, resulting in 893 complexes (denoted as BT-893) which contain 5187 distinct proteins (397 of those shared between complexes).

To compare our results against the smaller Pu and Hart predictions, more confident complexes were extracted from the original set at a threshold of 0.32. This set contains 409 complexes with 1692 distinct proteins (101 shared between complexes) and will be referred to as BT-409 in the following. It is comparable in size with the Pu predictions of 400 complexes with 1914 distinct proteins (141 shared) and the Hart predictions of 390 complexes with 1689 distinct proteins (none shared). We also extracted a second set of 217 complexes (BT-217) at a threshold of 0.69 which has a similar size as the MIPS complexes. We compared our selection approach against the less efficient method of first restricting the network and clustering afterwards and found that the differences observed were negligible with mutual sensitivities of 0.97 on average.

3.1 Evaluation of Interaction Networks

The quality of the bootstrap network in comparison to previously suggested interaction propensities was evaluated using a *receiver operating characteristic* (ROC) curve [27]. In a ROC curve, true positive rates are plotted against false positive rates for gradually decreasing thresholds. True positive interactions were defined as interactions between proteins in the same MIPS complex. The large and small ribosomal subunits were excluded since they would otherwise make up 44% of the true positive interactions. A set of true negative interactions was defined by randomly selecting interactions between proteins assigned to different MIPS complexes and cellular localizations by Hu et al. [25].

Figure 2 **A** shows the resulting ROC curves for the Gavin, Krogan, Collins, Hart and bootstrap scores as well as socio-affinity scores calculated from the combined experiments. We see that for all networks calculated from the combined data, the curve is steeper and reaches a higher level than for the propensities calculated from each experiment alone. Furthermore, the bootstrap scores calculated with our method performed best at separating true interactions from noise. Among the scoring methods proposed after the publication of the original purification studies, the Collins scores performed worst. Nevertheless, they

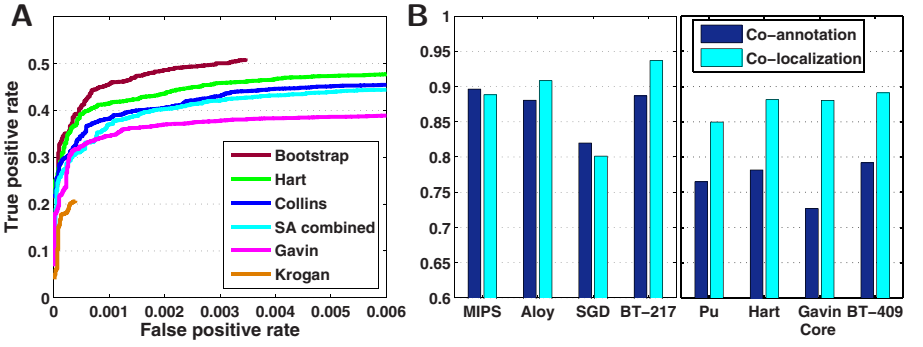

Fig. 2. ROC curves (**A**) are given for the Bootstrap, Hart and Collins scores, the socio-affinity scores for the combined data (SA combined) and the Gavin and Krogan scores (from top to bottom). True positive rates on the y coordinate are plotted against false positive rates on the x coordinate for gradually decreasing thresholds. Figure **B** illustrates co-annotation (dark blue) and co-localization (light blue) scores for the MIPS, Aloy and SGD complexes and the highly confident BT-217 complexes on the left hand side and the Pu, Hart, Gavin core and BT-409 predictions on the right hand side.

still perform slightly better for the given range than the socio-affinity scores computed from the combined experiments.

3.2 Co-annotation and Co-localization within Complexes

To assess the quality of the predicted complexes, we calculated the co-annotation and co-localization scores for the MIPS, Aloy and SGD complexes as well as for the Pu, Hart, Gavin and Krogan predictions and the BT-409 and BT-217 complexes (see Figure 2 **B**). Furthermore, the Gavin core set was evaluated which contains only the core components defined by Gavin et al. [5].

The lowest functional and localization similarity is observed for the Gavin and Krogan complexes (data not shown). By restricting to the more confident core components in the Gavin predictions, both co-annotation and co-localization can be increased significantly by 17% and 25%, respectively. Among all previous approaches, the highest co-annotation scores are obtained by the Pu and Hart predictions and the highest co-localization scores by the Hart predictions and the Gavin core set.

Functional and localization similarity is highest in the MIPS, Aloy and BT-217 complexes. Among the reference complexes, the SGD complexes perform worst. While co-annotation is still higher than in the best predictions, co-localization is significantly lower. This suggests that these automatically derived complexes are considerably less reliable than hand-curated ones.

When evaluating the complexes identified by our approach, we find that the BT-409 complexes perform significantly better than the Pu and Gavin core

complexes with regard to functional and localization similarity and slightly better than the Hart complexes. Moreover, the highly confident BT-217 complexes show similar co-annotation and higher co-localization scores than the hand-curated MIPS and Aloy complexes. It should be noted though that a large fraction of the BT-217 complexes is already well-known as 64% of these complexes share at least two proteins with one of the reference complexes. In the BT-409 set, this applies only to 43% of the complexes.

3.3 Validation on Reference Complexes

By comparing the predicted complexes against the current knowledge on protein complexes in the form of the MIPS, Aloy and SGD reference complexes, the sensitivity and and the PPV of the corresponding methods can be estimated. One should keep in mind, though, that these estimates may be unprecise due to the incompleteness of current knowledge.

Results for the comparison against the MIPS complexes are shown in Figure 3 **A**. Similar trends are observed for the comparison against all reference sets. The worst results are obtained by the original Gavin complexes, followed by the Krogan complexes. Here, the Gavin complexes are generally more sensitive but less accurate in their predictions than the Krogan complexes. By restricting the Gavin complexes to the core components, the PPV can be increased beyond that of any other prediction. However, this improvement comes at the cost of a very low sensitivity.

When comparing the performance of the BT-409, Pu and Hart complexes, we observe that none of the predictions is clearly superior to the other two. Although the sensitivity of the Pu complexes is slightly higher than for the other two approaches, the corresponding PPV is in return lower. Thus, it appears that all predictions cover the reference complexes with similar quality. The PPV of the BT-409 complexes can be increased slightly by restricting to the more confident BT-217 complexes, however the loss of sensitivity again is considerable.

3.4 Towards a Consensus of Complex Predictions

Although functional and localization similarity within complexes is slightly higher for the BT-409 predictions than for the Pu and Hart predictions, the comparison against the reference complexes yielded very similar results for all three sets. In order to appreciate how much these predictions agree or diverge, we compared them at the level of the individual complexes.

First, we calculated pairwise PPV and sensitivity values by taking either set of complexes once as prediction and once as reference for each possible pairwise combination. Here, we observed an average PPV of 0.85 and sensitivity of 0.72. This suggests that the agreement between each pair of these new predictions is much higher than between the Gavin and Krogan complexes for which we observe a mutual PPV of 0.26 and a sensitivity of 0.29.

In a second step, we calculated for each pair of predictions the number of complexes with (a) no significant overlap (at least 2 proteins) to the other set,

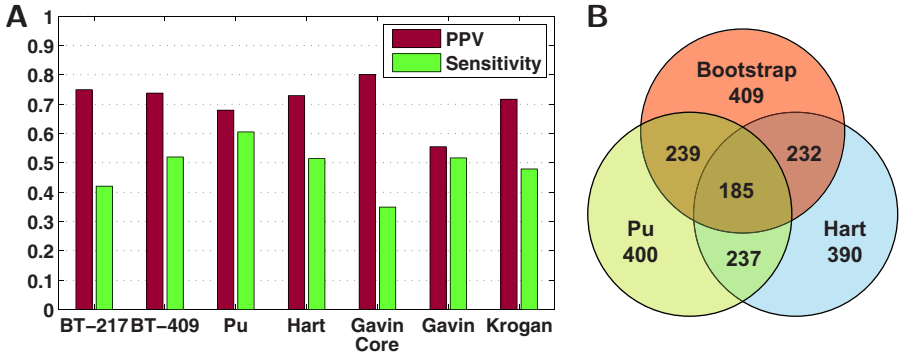

Fig. 3. PPV (red) and sensitivity (green) are shown in **A** for the BT-217, BT-409, Pu, Hart, Gavin core, Gavin and Krogan complexes. **B** illustrates the number of complexes which can be assigned consistently between the BT-409, Pu and Hart complexes. Results are shown for each pairwise comparison as well as the comparison of all three sets.

(b) a significant overlap to exactly one complex in the other set which again has no other overlaps and (c) significant overlaps but without an one-to-one correspondence as in (b). In the second case, we also distinguished between complexes with an exactly matching counterpart and complexes which contain additional proteins in at least one of the predictions. The same analysis was also performed for all three sets together.

For the second group of complexes, results are shown in Figure 3 **B**. For more than half of complexes in this group the predictions agree exactly. For the remaining complexes, each prediction adds on average 28-34% proteins to the proteins common to all two or three predictions. As can be seen, the consensus of each pair of predictions is much higher than for all predictions taken together. Nevertheless, even in the latter case the consensus is still larger at 46% (185 complexes) than between the Gavin and Krogan complexes where less than ten percent (45 complexes) have a clear one-to-one correspondence between the predictions.

Furthermore, 42% of the Gavin complexes and 64% of the Krogan complexes have no significant overlap to the respective other set. Contrary to that, only about 25% of complexes in the pairwise comparisons of the BT-409, Pu and Hart complexes and 16% in the comparison of all three predictions are unique to the corresponding prediction. These complexes are also characterized by considerably lower scores in the respective network and by smaller sizes than complexes in the other groups. Highest confidence values and medium complex sizes are observed for the second group while the last group generally contains the largest complexes.

4 Discussion

In this article, we presented an algorithm for the prediction of protein complexes from purification experiments alone. It implements all necessary steps in

an unsupervised manner from the combination of different experiments up to the identification of shared proteins. Accordingly, it does not depend on the availability of additional information on protein complexes and interactions and is not limited to organisms for which such an extensive knowledge exists as in the case of yeast. Therefore, our method can be applied to large-scale TAP experiments for any organism.

Intuitive and accurate confidence scores for protein interactions were obtained by an application of the bootstrap technique. For this purpose, our complex identification method was applied to preliminary networks calculated from bootstrap samples to estimate the stability of interactions. The resulting confidence scores distinguished better between correct and wrong interactions than all previous scoring methods, in particular also better than the scoring method used for the preliminary networks.

The same complex identification method was then applied to the complete bootstrap network to yield a large set of complex predictions. From this large set, we extracted approximately the same number of high-confidence complexes (BT-409) as the so far best methods by Pu et al. and Hart et al. The comparison of functional and localization similarity within complexes showed slightly better results for the BT-409 complexes compared to the Pu and Hart predictions. Furthermore, the predictive performance with regard to known reference complexes proved to be equivalent. This suggests that meaningful complexes can be derived from the purification experiments without additional training data.

When analyzing the individual BT-409, Pu and Hart complexes, we found that about 60% of the complexes have a one-to-one correspondence in pairwise comparisons. Here, each prediction shows approximately the same agreement with each of the other predictions. When combining all three predictions, the fraction of complexes identified consistently drops to 46%. This shows, that the consensus between each pair of predictions is larger than between all three of them. Nevertheless, the degree of agreement is still significantly higher than observed between the original Gavin and Krogan predictions.

In general, complexes in the consensus set are assigned higher confidence values by each method than more diverging complexes. A possible explanation is that these complexes are most stable and strongly connected and thus are detected more consistently by all methods than more transient or weakly connected complexes. Furthermore, low confidence scores may be an indication for a higher degree of uncertainty. Thus, the confidence of complexes should be taken into account for any analysis based on these protein complexes. However, since the more confident complexes tend to be already covered to a large degree by existing biological knowledge, new information may be found preferentially in the less confident ones. Thus, these should not be discounted per se but validated in additional experiments. Here, the original large set of complexes (BT-893) identified in this study can be a valuable resource for biological hypothesis generation and testing.

5 Availability

Bootstrap scores for the combined purification experiments, as well as the BT-893, BT-409 and BT-217 complexes can be downloaded from `http://www.bio.ifi.lmu.de/Complexes`.

References

1. Uetz, P., et al.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623–627 (2000)
2. Ito, T., et al.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA 98, 4569–4574 (2001)
3. Ho, Y., et al.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415, 180–183 (2002)
4. Gavin, A.-C., et al.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147 (2002)
5. Gavin, A.-C., et al.: Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631–636 (2006)
6. Krogan, N.J., et al.: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440, 637–643 (2006)
7. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417, 399–403 (2002)
8. Pu, S., Vlasblom, J., Emili, A., Greenblatt, J., Wodak, S.J.: Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. Proteomics 7, 944–960 (2007)
9. Collins, S.R., et al.: Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol. Cell. Proteomics 6, 439–450 (2007)
10. Hart, G.T., Lee, I., Marcotte, E.: A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics 8, 236 (2007)
11. Mewes, H.W., et al.: MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res. 32, 41–44 (2004)
12. Aloy, P., et al.: Structure-based assembly of protein complexes in yeast. Science 303, 2026–2029 (2004)
13. Ashburner, M., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000)
14. Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V., Mewes, H.W.: CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. 36, 646–650 (2008)
15. Efron, B.: Bootstrap methods: Another look at the jackknife. The Annals of Statistics 7, 1–26 (1979)
16. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, Boca Raton (1994)
17. van Dongen, S.: Graph Clustering by Flow Simulation. Ph.D. thesis, University of Utrecht (2000)
18. Felsenstein, J.: Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791 (1985)

19. Bader, G.D., Hogue, C.W.V.: Analyzing yeast protein-protein interaction data obtained from different sources. Nat. Biotechnol. 20, 991–997 (2002)
20. Brohee, S., van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7, 488 (2006)
21. Resnik, P.: Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
22. Lin, D.: An information-theoretic definition of similarity. In: Proc. 15th International Conf. on Machine Learning, pp. 296–304. Morgan Kaufmann, San Francisco (1998)
23. Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19, 1275–1283 (2003)
24. Schlicker, A., Domingues, F.S., Rahnenführer, J., Lengauer, T.: A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 7, 302 (2006)
25. Huh, W.-K., et al.: Global analysis of protein localization in budding yeast. Nature 425, 686–691 (2003)
26. Dwight, S.S., et al.: Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res. 30, 69–72 (2002)
27. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition 27, 861–874 (2006)