JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 8, 2009

Pp. 971–987
DOI: 10.1089/cmb.2009.0023

Research Article

Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast

CAROLINE C. FRIEDEL, JAN KRUMSIEK, and RALF ZIMMER

ABSTRACT

Protein interactions and complexes are important components of biological systems. Recently, two genome-wide applications of tandem affinity purification (TAP) in yeast have increased significantly the available information on interactions in complexes. Several approaches have been developed to predict protein complexes from these measurements, which generally depend heavily on additional training data in the form of known complexes. In this article, we present an unsupervised algorithm for the identification of protein complexes which is independent of the availability of such additional complex information. Based on a Bootstrap approach, we calculate intuitive confidence scores for interactions more accurate than all other published scoring methods and predict complexes with the same quality as the best supervised predictions. Although there are considerable differences between the Bootstrap and the best published predictions, the set of consistently identified complexes is more than four times as large as for complexes derived from one data set only. 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 species even if few complexes are already known.

Key words: algorithms, computational molecular biology, protein complexes, interactomics.

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) (Uetz et al., 2000; Ito et al., 2001; Giot et al., 2003; Li et al., 2004; Rual et al., 2005; Stelzl et al., 2005; LaCount et al., 2005) and affinity purification followed by mass spectrometry, e.g., co-immuno-precipitation (Co-IP)(Ho et al., 2002) or tandem affinity purification (TAP) (Gavin et al., 2002, 2006; Krogan et al., 2006). Since new methods are often applied first to the yeast *Saccharomyces cerevisiae*, its interactome is the most thoroughly studied.

While Y2H identifies only direct physical interactions, affinity purification can also identify indirect interactions via other proteins in protein complexes. In genome-scale studies on complexes, the TAP system was applied to yeast separately by Gavin et al. (2006) and Krogan et al. (2006). In the first experiment, 1,993 distinct TAP-tagged proteins (baits) were purified successfully and 2,760 distinct proteins (preys) were

Institut für Informatik, Ludwig-Maximilians-Universität München, München, Germany.

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. After the publication of the original results, improved prediction methods were developed by Pu et al. (2007) based on the scoring method of Collins et al. (2007) and by Hart et al. (2007). These methods used the data from both purification experiments. Recently, two new methods were proposed by Rungsarityotin et al. (2007) and Zhang et al. (2008) but only applied to the Gavin data. Furthermore, a complex prediction method was previously developed by Scholtens et al. (2005) but only applied to two smaller studies in yeast (Gavin et al., 2002; Ho et al., 2002).

A comparison of the different approaches outlines the important steps in complex determination (Fig. 1A). First, purification experiments have to be combined for higher prediction quality. Second, scoring weights for individual protein interactions have to be determined from the purifications. Third, these scores should be converted to confidence values between 0 and 1 which quantify the confidence of the prediction algorithm. Such confidence values are more easily interpretable than arbitrary scores between 0 (or $-\infty$) and ∞. Although the method by Hart et al. (2007) calculates p-values as interactions scores which have a clear defined minimum and maximum, these scores cannot be used as confidence values, since p-values depend strongly on the data set size and, thus, are difficult to compare between experiments.

In a fourth step, a cutoff has to be applied on the network to select the most confident interactions. The restricted network then has to be clustered to obtain individual complexes. 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 proteins can be involved in more than one complex.

Apart from the combined method by Collins et al. (2007) and Pu et al. (2007), all described approaches leave out at least one of these steps. Although both Rungsarityotin et al. (2007) and Scholtens et al. (2005) calculated interaction scores in their algorithms, these interactions scores are only used as an intermediate step and cannot be obtained from the output of the algorithms. Figure 1A gives an overview on which

A

	GA	KR	CP	HA	RU	ZH	SVG	BT
Experiment combination	-	-	S	U	-	-	-	U
Interaction scores	U	S	S	U	U	U	U	U
Confidence values	-	S	S	-	-	-	-	U
Network cutoff	U	S	S	S	-	S	-	U
Clustering	S	S	S	S	U	U	U	U
Shared Proteins	S	-	S	-	-	U	U	U

B

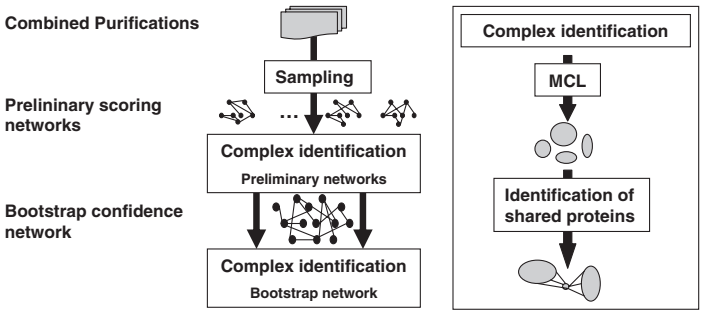


FIG. 1. Outline of complex prediction methods. (**A**) Major steps involved in complex identification and whether they are realized by the approaches of Gavin et al. (2006) (GA), Krogan et al. (2006) (KR), Collins et al. (2007) and Pu et al. (2007) (CP), Hart et al. (2007) (HA), Rungsarityotin et al. (2007) (RU), Zhang et al. (2008) (ZH), Scholtens et al. (2005) (SVG), and the Bootstrap approach described here (BT). Furthermore, we indicate whether a supervised approach based on additional training data is used (S, light gray) or an unsupervised approach (U, white). Dark gray indicates that the step is omitted. (**B**) Major steps of our Bootstrap approach.

methods implement which steps in a supervised or unsupervised way. Here, supervised means that additional training data is used apart from the purification experiments alone.

With the exception of the method by Rungsarityotin et al. (2007), all approaches described above rely more or less heavily on the availability of such additional training data in the form of known complexes for at least one step. For yeast, manually curated complexes can be taken from the MIPS database (Mewes et al., 2004) and the study of Aloy et al. (2004). Furthermore, complexes can be automatically extracted from Gene Ontology (GO) annotations (Ashburner et al., 2000). Unfortunately, the resulting complexes are of lower quality than the manually curated ones. Recently, manually curated complexes have become available for human and other mammals which cover about 12% of the protein-coding genes in human (Ruepp et al., 2008). For other organisms, such information on complexes 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 several of the above mentioned approaches distinguish between test and training set by choosing one of the yeast reference sets for training and a different one for testing, these sets overlap to a large degree and, thus, are not sufficiently disjoint to guarantee a reliable performance estimate.

To deal with these problems, we developed an unsupervised algorithm for the identification of protein complexes from the purification data alone which implements all steps described above (Friedel et al., 2008). Since no additional information on protein complexes is required, our approach is not limited to yeast and other organisms for which many protein complexes have already been identified but can be applied easily to large-scale purification experiments for any species.

We show that our approach is equivalent to the best supervised methods both with regard to functional and localization similarity 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 on the combined data sets are much more consistent than the original Gavin and Krogan complexes. However, although the latest predictions are of similar quality, significant differences 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 six important steps (Fig. 1A). Purification experiments are combined by pooling them. Interaction confidences are determined by first identifying preliminary complexes for Bootstrap samples from the set of purifications (Efron, 1979; Efron and Tibshirani, 1994). The resulting confidence network is then clustered with the MCL algorithm (van Dongen, 2000), and proteins shared between complexes are identified in a post-processing step. Here, 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 (Fig. 1B). Details of the algorithm are described in the following.

2.1. Combination of experiments

While Gavin et al. (2006) used only one mass spectrometry method to identify proteins co-purified with a bait, Krogan et al. (2006) used two separate methods (MALDI-TOF and LC-MS/MS). Since mass spectrometry is one potential source of false positives and false negatives in the experiments, this approach increases the coverage and accuracy of the method. Proteins identified with both methods are highly confident, however the coverage is very low if only these proteins are considered (Krogan et al., 2006). On the other hand, if all proteins identified by at least one method are used, the quality of the results is decreased.

As a trade-off between coverage and accuracy, we use the following approach. Prey proteins identified with both mass spectrometry methods will be counted twice in a purification, while proteins only identified with one method will be counted only once. Thus, interactions between proteins identified with two methods will get a higher weight in the calculation of interaction scores described in the next section. Since Gavin et al. used only one mass spectrometry method, preys are only counted once for purifications from this experiment in the pooled set of both experiments.

2.2. Bootstrap sampling

To determine reliable confidence scores, the Bootstrap technique (Efron, 1979; Efron and Tibshirani, 1994) 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 (Felsenstein, 1985). In the following, let $\Phi = (\phi_1, \ldots, \phi_n)$ be the list of purifications and V the set of proteins. Each purification ϕ_i consists of one bait protein $b_i \in V$ and the preys $p_{i,1}, \ldots, p_{i,m} \in V$ identified for this bait in the purification

$$\phi_i = \langle b_i, [p_{i,1}, \ldots, p_{i,m}] \rangle. \quad (1)$$

From the list of purifications Φ, l Bootstrap samples are created (in our case $l = 1,000$). Each Bootstrap sample $S_j(\Phi)$ is created by drawing n purifications with replacement from Φ. This means that the Bootstrap sample $S_j(\Phi)$ contains the same number of purifications as Φ and each purification ϕ_i can be contained in $S_j(\Phi)$ once, multiple times or not at all. Multiple copies of the same purification are treated as separate purifications.

For each Bootstrap sample $S_j(\Phi)$, we calculate preliminary interaction scores for each pair of proteins co-purified at least once. For this purpose, we use the socio-affinity scores described by Gavin et al. (2006) which compare the number of co-occurrences of two proteins against the random expectation using a combination of spoke and matrix model (Bader and Hogue, 2002). No additional training data is required to compute them from a set of purifications.

The preliminary scoring network for each sample $S_j(\Phi)$ is then defined as $G_j = (V, E_j)$ with $E_j = \{(u, v) | u, v \in V \wedge w_{S_j(\Phi)}(u, v) \geq \tau_P\}$. Here, $w_{S_j(\Phi)}(u, v)$ is the socio-affinity score for the interaction between u and v in the Bootstrap sample $S_j(\Phi)$ and τ_p a pre-defined cutoff which is applied to this network to filter for the most confident interactions and to allow for fast computation. From these preliminary networks, we then determine preliminary complex predictions for each individual Bootstrap sample $S_j(\Phi)$ with the algorithm described in the following section.

2.3. 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.

2.3.1. Clustering. Networks are clustered using the Markov clustering algorithm (MCL) developed by van Dongen (2000). In a recent study (Brohee and van Helden, 2006), this algorithm was found to be superior to several other graph clustering algorithms for the identification of protein complexes. As a consequence, many 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 very 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 based on MCL used additional training data in the form of known complexes to choose the optimal inflation parameter. In this article, we suggest using an intrinsic measure which compares the resulting clusters against the original network from which the clusters were obtained. 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 van Dongen (2000). 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 cluster the preliminary networks G_j for each sample $S_j(\Phi)$ with gradually increasing inflation parameters. For each inflation parameter, we calculate the average efficiency over all samples. We found that, for our networks, the efficiency always reaches a maximum for a certain inflation parameter and decreases on either side of this value. 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.

2.3.2. 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 manner similar to that of Pu et al. (2007), 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 scoring 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 interaction strength of the protein to the complex which is required to add the protein depends on the strength of interactions within the 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

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

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.

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 the protein p_i has to have sufficiently strong interactions. The threshold 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.

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. This value of γ is identified with a binary search. As we only want to add proteins to complexes to which they are clearly associated, we set λ very high at 0.95. 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.

2.4. Calculation of confidence scores and final complexes

The final confidence scores are then determined by calculating the so-called Bootstrap network G_{BT} 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). Thus, the confidence score c_{BT} for an interaction (u,v) is defined as

$$c_{BT}(u, v) = \frac{1}{l} \sum_{j=1}^{l} \delta_{S_j(\Phi)}(u, v) \tag{3}$$

where $\delta_{S_j(\Phi)}(u, v)$ is 1 if u and v are in the same complex for Bootstrap sample $S_j(\Phi)$ and 0 otherwise.

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 by choosing a stringent cut-off τ_P 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 threshold τ_C 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 approach used by Hart et al. (2007) of first restricting the network and then repeating complex identification but yields almost identical complexes.

2.5. Criteria for the evaluation of complex quality

2.5.1. Functional co-annotation within complexes. Since protein complexes are formed to carry out a specific function, the function 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) (Ashburner et al., 2000). 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 (Resnik, 1999; Lin, 1998; Lord et al., 2003; Schlicker et al., 2006). Here, we use the relevance similarity proposed recently by Schlicker et al. (2006). 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'' ontologies. The final co-annotation score is then calculated as the geometric mean of the ''biological process'' and ''molecular function'' GO scores.

2.5.2. 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 Huh et al. (2003).

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

$$L = \frac{\sum_j \max_i l_{i,j}}{\sum_j |C_j|} \tag{4}$$

Here, $l_{i,j}$ 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.

We do not calculate the co-localization score of a complex C_j as $\max_i l_{i,j} / \sum_i l_{i,j}$ as suggested by Pu et al. (2007). This method would give a score of 0.5 to a complex of two proteins where each protein is assigned to the same two co-localization categories. However, such a complex is perfectly co-localized and, accordingly, is given a value of 1 by our co-localization score.

2.5.3. 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) manually curated complexes from the MIPS database (Mewes et al., 2004) (214 complexes after removing redundant complexes) and the study of Aloy et al. (2004) (101 complexes) as well as (b) complexes extracted from the SGD database (Dwight et al., 2002) 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 (ftp://genome-ftp.stanford.edu/).

We used the definition of sensitivity and PPV for protein complexes by Brohee and van Helden (2006). 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:

$$Sn = \frac{\sum_i \max_j T_{i,j}}{\sum_i |R_i|} \quad \text{and} \quad PPV = \frac{\sum_j \max_i T_{i,j}}{\sum_j \sum_i T_{i,j}}. \tag{5}$$

Thus, sensitivity evaluates which fraction of proteins in known complexes are recovered and positive predictive value determines how good the predicted complexes match to the known reference complexes. For the calculation of sensitivity only reference complexes are considered for which $\max_j T_{i,j} \geq 1$.

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 28% of the retrieved preys between the two experiments. This is comparable to the agreement between purifications of the same bait within the Krogan data set (23%), but significantly smaller than within the Gavin set (49%).

From these purifications the final Bootstrap network was calculated. We used a cut-off $\tau_P = 8$ on the socio-affinity scores to derive preliminary networks for each Bootstrap sample. We chose a more stringent threshold than the one recommended by Gavin et al. (2006) 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 times more). Thus, the more stringent threshold τ_P 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. Because of this relatively small network size, no additional cut-off is necessary before predicting protein complexes with our approach. Here, we obtained 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 with a threshold $\tau_C = 0.32$. 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, 940 distinct proteins, 54 shared) at $\tau_C = 0.69$ which has a similar size as the MIPS complexes (214 complexes, 1190 distinct proteins, 119 shared). We compared our selection approach against the less efficient method described by Hart et al. (2007), of first restricting the network and clustering afterwards, and found that the differences observed were negligible with sensitivities in both directions of about 0.97.

3.1. Evaluation of interaction networks

The quality of the Bootstrap network in comparison to other suggested interaction scores was evaluated using a *receiver operating characteristic* (ROC) curve (Fawcett, 2006). 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. True negative interactions were defined as interactions between proteins assigned to different MIPS complexes and cellular localizations by Huh et al. (2003).

Figure 2A shows the resulting ROC curves for the Gavin, Krogan, Collins, Hart, and Bootstrap scores as well as socio-affinity scores and Dice coefficients (Zhang et al., 2008) calculated from the combined experiments with the ProCope software package (Krumsiek et al., 2008). We see that for all networks calculated from the combined data, the curve is generally steeper and reaches a higher level than for the scores 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 and Dice coefficients performed worst. Nevertheless, they still performed slightly better for the given range than the socio-affinity scores computed from the combined experiments.

Figure 2B illustrates the maximum possible true positive rate and corresponding false positive rate if all interactions in the scoring networks are predicted. Here, we see that only the socio-affinity scores and Dice coefficients can obtain a slightly higher true positive rate than the Bootstrap confidence scores. However, this results in false positive rates 25 times as high as for the Bootstrap confidence scores and can only be obtained by including interactions with very low scores. For all other scoring networks, the true positive rate of the Bootstrap network cannot be reached, not even for high false positive rates.

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 (Fig. 3A). Furthermore, the Gavin core set was evaluated which contains only the core components defined by Gavin et al. (2006). In this analysis, we focused only on the predictions made on the combined set of purifications as well as the predictions from the original publications. The prediction method by Scholtens et al. (2005) could not be evaluated since the available R

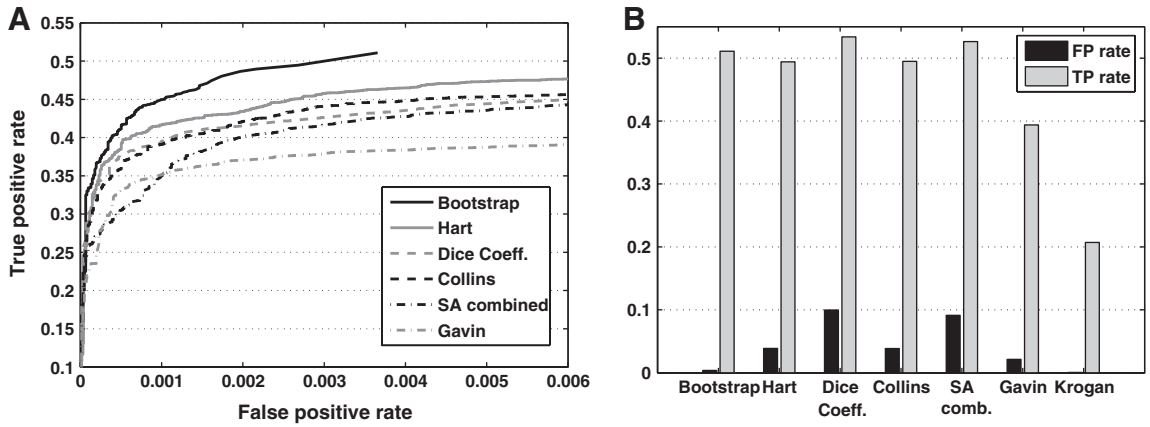


FIG. 2. Accuracy of the interactions predicted with the Bootstrap approach. (**A**) ROC curves are given for the Bootstrap, Hart, Dice coefficient and Collins scores, the socio-affinity scores for the combined data (SA combined) and the Gavin purifications alone (from top to bottom). Krogan scores are omitted since they performed significantly worse than any of the other scoring methods. True positive rates on the *y*-axis are plotted against false positive rates on the *x*-axis for gradually decreasing thresholds to predict an interaction from the scoring networks. (**B**) Maximum false positive and true positive rate for each scoring network.

implementation did not scale up to the large data sets by Gavin et al. (2006) and Krogan et al. (2006). A supplementary table containing the detailed results of the comparisons can be found at www.bio.ifi.lmu.de/Complexes/Bootstrap. Co-annotation scores are also given separately for the biological process and molecular function ontologies. However, rankings of prediction methods for either ontology are almost identical.

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 previously published prediction 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.

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

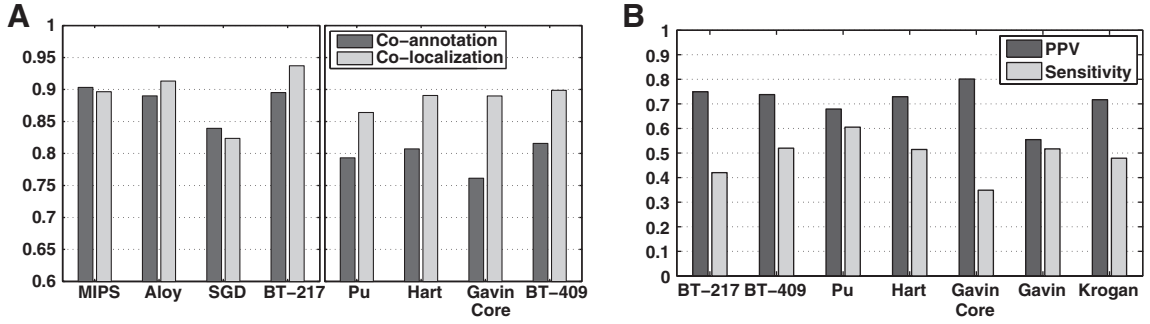


FIG. 3. Functional and localization similarity and predictive accuracy of complexes. (**A**) Co-annotation (dark gray) and co-localization (light gray) scores for the MIPS, Aloy and SGD complexes as well as the highly confident BT-217 complexes on the left hand side and the Pu, Hart, Gavin core, and BT-409 predictions on the right hand side. (**B**) PPV (dark gray) and sensitivity (light gray) are shown for the BT-217, BT-409, Pu, Hart, Gavin core, Gavin, and Krogan complexes.

show similar co-annotation and higher co-localization scores than the manually 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 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 3B. 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 lower in return. 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. Assessing predictions from the Gavin data alone

Since the recently published methods by Rungsarityotin et al. (2007) and Zhang et al. (2008) have only been applied to the Gavin data, we also applied our Bootstrapping algorithm to this data set alone to compare it against these two prediction approaches. Due to the smaller size of the Gavin data set, we used a slightly lower τ_P of 7 on the preliminary networks to obtain 381 complexes with 1970 distinct proteins.

We compared our results against the original Gavin predictions, the Markov random field (MRF) predictions by Rungsarityotin et al. (2007) both from the matrix and spoke model and the predictions by Zhang et al. (2008). From the MRF matrix prediction we also selected the 381 complex predictions with highest quality score and for the Zhang et al. predictions we distinguished between complete and core predictions.

We found that functional and localization similarity (Fig. 4) is always higher in the Bootstrap complexes than both in the complete and high-confidence MRF and Zhang predictions. PPV is slightly higher in the MRF and Zhang core predictions, but on the other hand their sensitivity is considerably lower than in the Bootstrap predictions. This shows that the complexes identified with our approach are of higher overall quality and accuracy than complexes predicted with these more recently published methods.

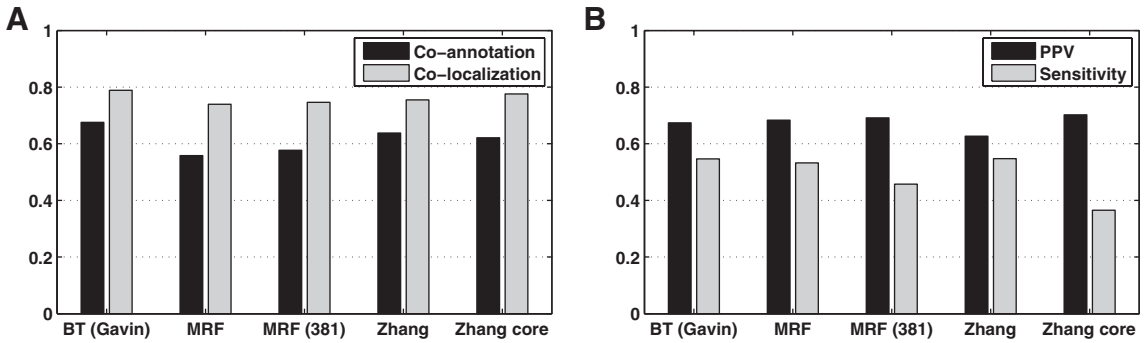


FIG. 4. Functional and localization similarity (**A**) and positive predictive value and sensitivity (**B**) for the predictions on the Gavin data alone. Results are shown for the Bootstrap (BT) predictions on the Gavin set, the Markov random field (MRF) predictions of Rungsarityotin et al. (2007) (for the matrix model, results for the spoke model are similar), the 381 best MRF matrix predictions and the predictions by Zhang et al. (2008) both for the complete and core set.

The Bootstrap predictions on the Gavin data have significantly higher functional and localization similarity and predictive performance than the complete Gavin complexes. To compare our predictions against the high confidence Gavin core predictions, we extracted a more confident set of 210 complexes with 1127 distinct proteins using a τ_C of 0.32 from our complete predictions. Although this complex set contains fewer complexes than the Gavin core predictions (423 complexes with 1128 proteins), the number of proteins is almost identical. We found that localization similarity is very similar between both predictions ($\sim$0.88), whereas functional similarity is significantly higher in the Bootstrap complexes (Bootstrap, 0.81; Gavin core, 0.73). Furthermore, the Gavin core predictions obtain a higher PPV (Bootstrap, 0.7; Gavin core, 0.8) than the Bootstrap complexes but only by having a significantly lower sensitivity (Bootstrap, 0.46; Gavin core, 0.35).

3.5. 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.

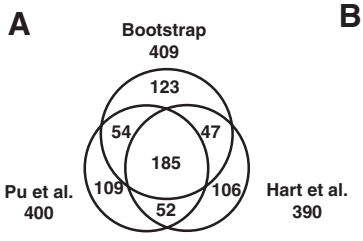
First, we calculated PPV and sensitivity values in both directions for each possible pairwise combination of prediction methods. Here, we observed an average PPV of 0.85 and sensitivity of 0.72. This suggests that the agreement between each pair of these new predictions is much higher than between the Gavin and Krogan complexes for which we observe an average PPV of 0.26 and sensitivity of 0.29.

In a second step, we calculated for each pair of prediction methods the number of complexes with (a) no significant overlap (at least two proteins) to the other set, (b) a significant overlap to exactly one complex in the other set which again has no other overlaps (consistent complexes), and (c) significant overlaps but without a one-to-one correspondence as in (b) (inconsistent complexes). 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.

This analysis showed that 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 supported by only one method. This is much lower than observed between the Gavin and Krogan complexes, where 42% and 64% of the complexes, respectively, have no significant overlap to the other set

For the consistent complexes, results are shown in Figure 5A. For more than half of complexes in this group the predictions agree exactly. In the remaining cases, 28–34% additional proteins are added by each method to the common core identified by all two or three predictions. Here, 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 with 185 complexes (46%) than between the Gavin and Krogan complexes where only 45 complexes ($<$10%) have a clear one-to-one correspondence between the predictions.

For the inconsistent complexes, we observe a one-to-many relationship in pairwise comparisons in $\sim$71% of cases. This means that a complex predicted by one method is subdivided into several complexes by the respective other method. Furthermore, inconsistent complexes agree in about 83% of the proteins between complex predictions. This indicates that although predictions identify approximately the same sets of proteins, they disagree with regard to the partitioning.



B

	Bootstrap	Pu et al.	Hart et al.
Bootstrap		71.0%	77.8%
Pu et al.	85.9%		65.4%
Hart et al.	82.5%	81.1%	

FIG. 5. Consensus between best supervised predictions and the Bootstrap approach. (**A**) Venn diagram of the number of complexes which can be assigned consistently between the BT-409, Pu, and Hart complexes. (**B**) Fraction of one-to-many mappings among inconsistent predictions for each pairwise comparison (above the diagonal) and the average protein overlap between inconsistent protein complexes (below the diagonal).

When we analyse the distribution of complex sizes for each prediction method (Fig. 6A), we observe that it follows approximately a power-law distribution with ~50% of the complexes consisting of only two proteins and only 6–7% of complexes having a size larger than 10 proteins. Thus, the majority of protein complexes are heterodimers and only in few cases do we observe very large multi-subunit structures.

We also evaluated average complex size separately for complexes supported by only one or two methods or identified consistently or inconsistently by all methods (Fig. 6B). Our results showed that complexes supported by only one method have very small sizes with 75–87% of complexes consisting only of an interaction of two proteins. Complexes supported by two methods show only slightly larger sizes. Medium complex sizes are observed for consistent complexes, where more than 50% contain more than two proteins. The inconsistent complexes generally are largest and 58–70% of those involve at least five proteins. We also found that average complex scores (Fig. 6C) were smallest for complexes supported by only 1 or 2 methods and largest in complexes identified consistently by all complexes. For the inconsistent complexes only medium complex scores are observed.

For each group of complexes (support 1 or 2, consistent or inconsistent in a comparison of all predictions), we calculated functional similarity and co-localization separately (Fig. 7). These results show that complexes identified consistently by all methods have the highest functional similarity and co-localization, while complexes supported by only one method generally have the worst. Apart from these overall

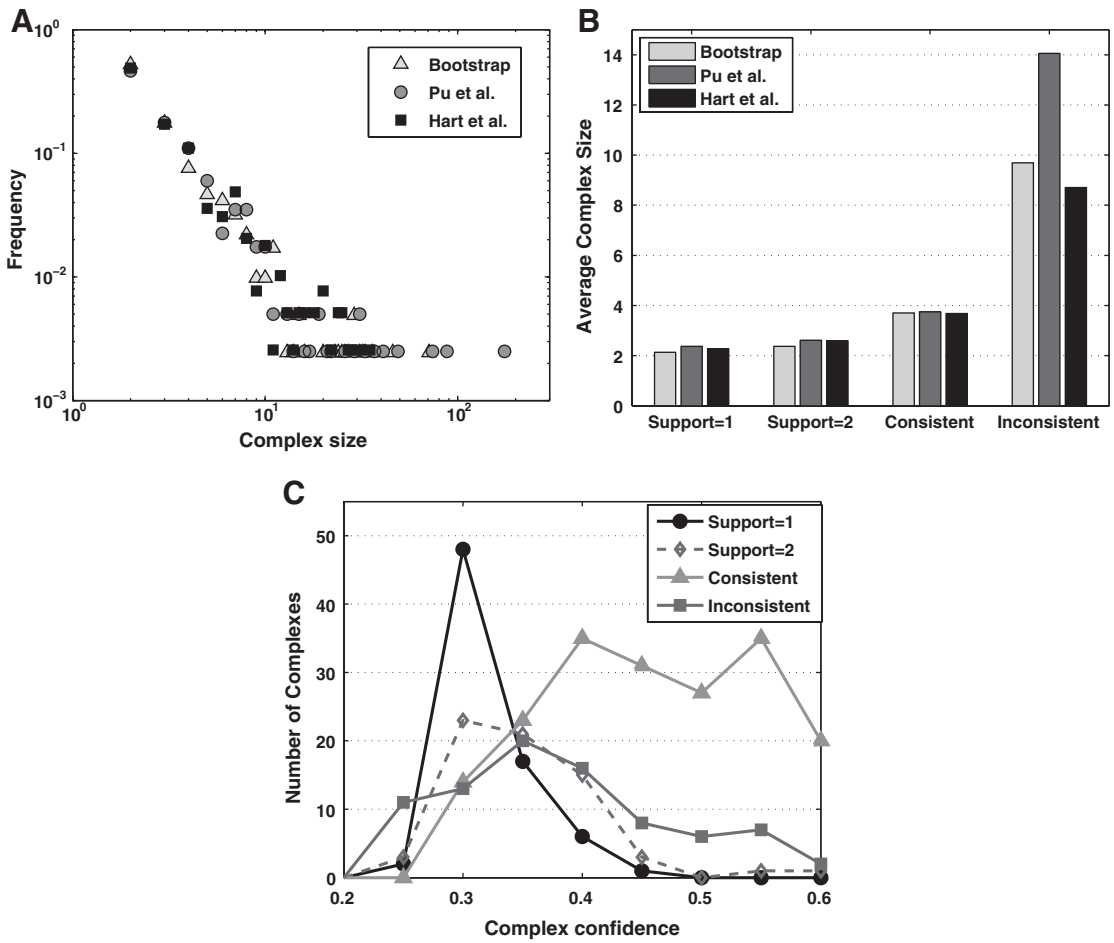


FIG. 6. Size and confidence of predicted protein complexes. (**A**) Distribution of complex size for the Bootstrap, Pu and Hart predictions. (**B**) Average complex size for all methods. (**C**) Distribution of complex confidence scores for the Bootstrap complexes (results for the Pu and Hart predictions show similar tendencies). Here, we distinguished between protein complexes which are supported by no other method (support = 1), by only one other method (support = 2) or can be mapped consistently or inconsistently between the Bootstrap, Pu and Hart predictions.

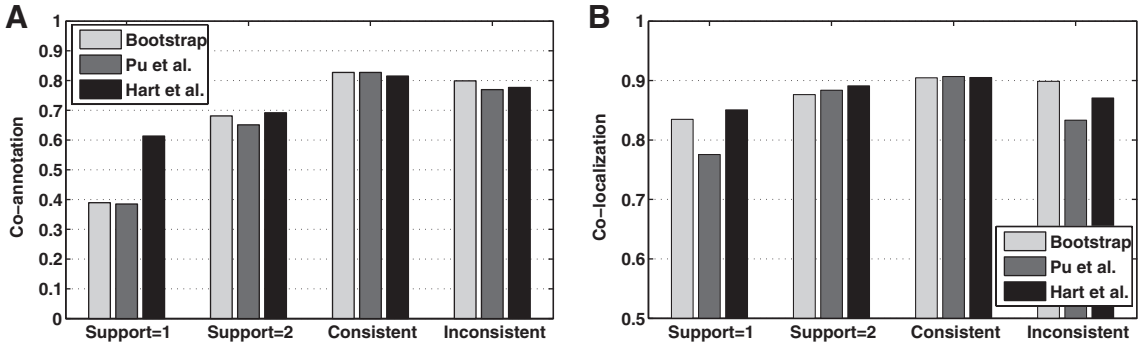


FIG. 7. Comparison of complexes identified consistently and inconsistently by different methods. Average co-annotation (**A**) and co-localization scores (**B**) are given for the four groups described in Figure 6.

tendencies, striking differences can be observed between the methods. Complexes predicted only by the Hart method perform significantly better than complexes predicted only by either the Bootstrap or the Pu approach. On the other hand, the Hart predictions for the inconsistent complexes are of lower overall quality than the Bootstrap predictions and the consistent complexes show lower functional similarities than both the Bootstrap and Pu predictions. On average, the Pu complexes perform worst across all groups.

3.6. Evaluation of shared proteins

For the identification of proteins shared between different complexes, the parameter λ was set to 0.95. To evaluate the influence of the choice of λ, we repeated the identification step with different values of λ. We then extracted more confident predictions at the same threshold used for the BT-409 predictions ($\tau_c = 0.32$) and compared positive predictive value, sensitivity, and co-annotation and co-localization scores. The corresponding results as well as the number of shared proteins identified are shown in Figure 8. As can be seen, the number of shared proteins identified increases approximately linearly with λ. At the same time, positive predictive values and functional similarity scores decrease with λ while sensitivity increases. Co-localization scores, on the other hand, stay approximately the same. These results show that by decreasing λ, the number of recovered proteins for each complex can be increased at the cost of also including more wrongly predicted subunits.

Furthermore, we compared shared proteins between the BT-409 complexes and the Pu predictions which, contrary to the Hart complexes, contain partially overlapping complexes. A detailed analysis of the shared proteins showed that in both cases, around 31% of the shared proteins identified are contained in only two very large overlaps between complexes (>16 proteins). In the Pu predictions, this is due to three protein complexes that cover slightly different but overlapping parts of the preribosome. In the Bootstrap predictions, one large overlap is also found between different preribosome subcomplexes and another between the spliceosome and a splicing factor complex.

Around 50% of shared proteins are shared between complexes whose intersection contains not more than three proteins. We evaluated manually the accuracy of those predictions which were not also contained in larger intersections (Fig. 8C). Predictions were classified as correct if we found evidence in the literature that they are contained in at least two of the complexes they were predicted for and wrong if we found clear evidence against this. We furthermore included an additional category of potential candidates for which there was neither conclusive evidence for or against. Proteins with unknown function were also treated separately. Our results show that the true positive rate is approximately the same for the Bootstrap and Pu predictions while the false positive rate is significantly higher in the Pu predictions. Since the Pu predictions were trained on known protein complexes from the MIPS database, this may indicate overfitting effects to large protein overlaps such as observed for the preribosome which results in lower prediction quality for smaller complex overlaps. Even if we include the potential candidates into the false positive predictions, our results show that the unsupervised Bootstrap approach can identify shared proteins with at least the same prediction accuracy as the supervised approach by Pu et al. (2007).

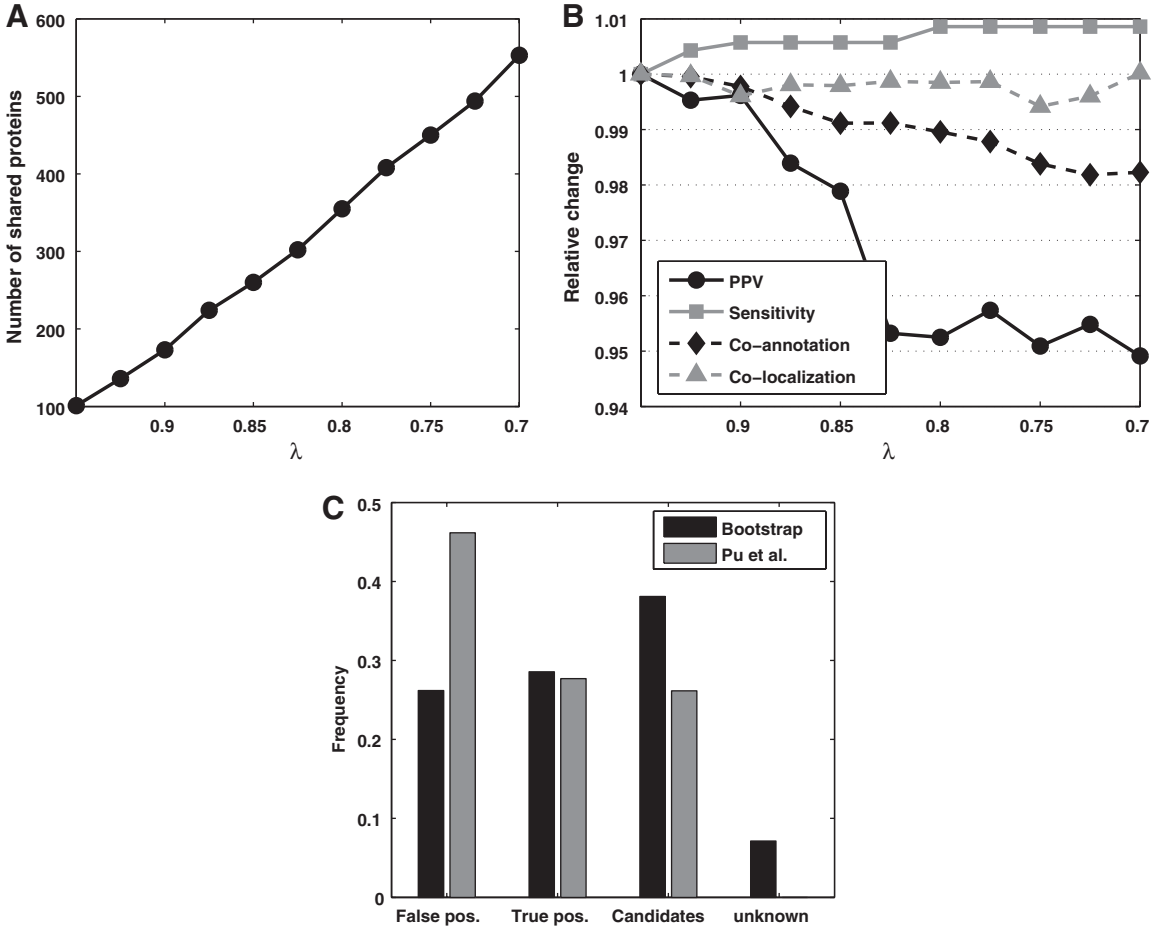


FIG. 8. Evaluation of shared proteins. (**A**) Number of shared proteins for increasing λ. (**B**) Positive predictive values (PPV) and sensitivity as well as co-annotation and co-localization scores divided by the corresponding values for the BT-409 set. Thus, values >1 indicate an increase and <1 a decrease in quality compared to the BT-409 set. (**C**) Shared proteins identified by the Bootstrap and Pu approach were evaluated manually if the intersection of complexes they are part of contains at most two other proteins; the fraction of false positives, true positives, candidates, and proteins with unknown function are shown for each approach. Proteins were classified as candidates if the functional annotation supports the association with at least two different complexes, but no clear evidence is found.

3.7. Comparison of example complexes

We compared the BT-409, Pu and Hart predictions on example complexes which are supported by only two methods or identified consistently or inconsistently by all three methods. Figure 9A shows a complex which is only found in the BT-409 and Hart predictions. Both predictions cluster the alpha and beta subunits of phosphofructokinase together. Furthermore, each approach predicts one additional protein of unrelated function. These additional proteins are probably false positives and, accordingly, the interactions to the additional proteins are scored significantly lower by both methods than the interaction between the two phosphofructokinase subunits.

Results for the SET3 histone deacetylase complex identified consistently by all approaches are shown in Figure 9B. Here, the Hart approach predicts five additional proteins of unrelated or unknown function, but fails to identify other components of the histone deacetylase complex: CPR1 and HST1 which are both identified by the Bootstrap and Pu approach. However, HST1 is only found in the BT-893 but not the BT-409 set. It has been shown that HST1 is only a non-essential subunit of this complex but an essential subunit of another complex (Pijnappel et al., 2001). This may explain why the Hart approach does not identify this protein and why interaction scores of HST1 to this complex are quite weak both in the Bootstrap and Pu predictions.

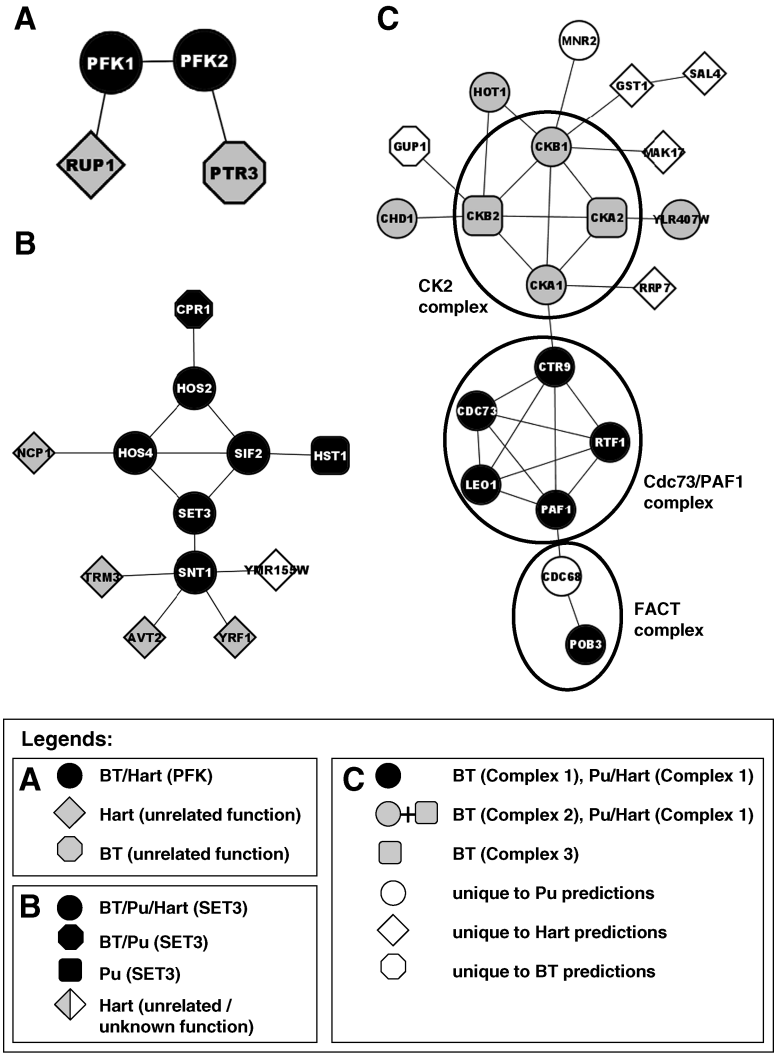


FIG. 9. Comparison of BT-409, Pu, and Hart predictions for complexes. (**A**) Phosphofructokinase (PFK) complex (black) predicted only by the Bootstrap and Hart approach. (**B**) SET3 histone deacetylase complex (black) predicted by all three approaches. (**C**) Comparison between the Bootstrap, Pu, and Hart predictions for the protein kinase CK2 complex and the PAF1 complex. Here, three complexes are predicted by the Bootstrap approach, while both Pu et al. and Hart et al. predict only one complex. Important interactions were extracted with the MST approach by Friedel and Zimmer (2008). Actual biological complexes found in the MIPS data set or literature are circled.

Figure 9C shows an example for an inconsistently identified complex. Here, the Bootstrap approach predicts three complexes: the PAF1 complex plus one protein of the FACT complex, the casein kinase CK2 complex with associated proteins as well as a smaller subcomponent of the CK2 complex consisting of one alpha and one beta subunit. The Pu and Hart methods group these complexes together into one complex. It has been shown that the CK2 complex associates with the PAF1 and FACT complexes during transcription (Krogan et al., 2002). However, since the CK2 complex is involved in many more biological processes (Ahmed et al., 2002), this association is not permanent which is reflected in the Bootstrap predictions. The additional proteins added to the CK2 complex might be some of the many targets of the CK2 protein kinase (Ahmed et al., 2002).

This last example illustrates the problems of complex prediction approaches in identifying the correct subdivision between interacting complexes or even more for largely overlapping complexes. Here, known difficult cases are the spliceosome and the RNA polymerases. Both the Bootstrap and the Pu approach cluster the several subunits of the spliceosome together with additional transcription factors. Contrary to that the Hart approach divides this complex into several smaller complexes which at least partially cor-

respond to subcomplexes of the spliceosome. In the case of the RNA polymerases which have several proteins in common, all approaches predict only one large complex. This shows that additional investigation of the predicted complexes is necessary for instance with the maximum spanning tree approach we presented recently (Friedel and Zimmer, 2008).

4. DISCUSSION

In this article, we presented an algorithm for the prediction of protein complexes from purification experiments alone. It implements all necessary steps from the combination of different experiments up to the identification of shared proteins in an unsupervised manner. 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. 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 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 other published scoring methods, in particular also better than the socio-affinity scores 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. (2007) and Hart et al. (2007). 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 complexes not supported by all methods or identified inconsistently. This suggests that evidence in the experimental data for these consistently identified complexes is relatively clear while evidence for other complexes is more ambiguous. Such ambiguous information may be observed due to measurement errors but also if complexes are not permanently associated or connected by weak interactions.

As low confidence scores indicate more unreliable complexes, 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 discarded 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.

4.1. Availability

Bootstrap scores for the combined purification experiments and the BT-893, BT-409 and BT-217 complexes can be downloaded from www.bio.ifi.lmu.de/Complexes/Bootstrap. Implementations of the Bootstrap algorithm and most other methods described here (e.g., socio-affinity scores) are included in the ProCope package (Krumsiek et al., 2008) which can be freely obtained from www.bio.ifi.lmu.de/Complexes/ProCope.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Ahmed, K., Gerber, D.A., and Cochet, C. 2002. Joining the cell survival squad: an emerging role for protein kinase CK2. *Trends Cell Biol*. 12, 226–230.

Aloy, P., et al. 2004. Structure-based assembly of protein complexes in yeast. *Science* 303, 2026–2029.

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

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

Brohee, S., and van Helden, J. 2006. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinform.* 7, 488.

Collins, S.R., et al. 2007. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae. Mol. Cell Proteomics* 6, 439–450.

Dwight, S.S., et al. 2002. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the gene ontology (GO). *Nucleic Acids Res*. 30, 69–72.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7, 1–26.

Efron, B., and Tibshirani, R. 1994. *An Introduction to the Bootstrap.* Chapman & Hall, London.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recogn. Lett.* 27, 861–874.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.

Friedel, C.C., and Zimmer, R. 2008. Identifying the topology of protein complexes from affinity purification assays. *Proc. German Conf. Bioinform.* 30–43.

Friedel, C.C., Krumsiek, J., and Zimmer, R. 2008. Bootstrapping the interactome: Unsupervised identification of protein complexes in yeast. *RECOMB 2008* 3–16.

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

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

Giot, L., et al. 2003. A protein interaction map of *Drosophila melanogaster. Science* 302, 1727–1736.

Hart, G.T., Lee, I., and Marcotte, E. 2007. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinform.* 8, 236.

Ho, Y., et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183.

Huh, W.-K., et al. 2003. Global analysis of protein localization in budding yeast. *Nature* 425, 686–691.

Ito, T., et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* 98, 4569–4574.

Krogan, N.J., et al. 2002. RNA polymerase II elongation factors of *Saccharomyces cerevisiae*: a targeted proteomics approach. *Mol. Cell Biol*. 22, 6979–6992.

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

Krumsiek, J., Friedel, C.C., and Zimmer, R. 2008. ProCope—protein complex prediction and evaluation. *Bioinformatics* 24, 2115–2116.

LaCount, D.J., et al. 2005. A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* 438, 103–107.

Li, S., et al. 2004. A map of the interactome network of the metazoan *C. elegans. Science* 303, 540–543.

Lin, D. 1998. An information-theoretic definition of similarity. *Proc. 15th Int. Conf. Mach. Learn.* 296–304.

Lord, P.W., Stevens, R.D., Brass, A., et al. 2003. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19, 1275–1283.

Mewes, H.W., et al. 2004. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*. 32, D41–D44.

Pijnappel, W.W., et al. 2001. The S. cerevisiae SET3 complex includes two histone deacetylases, Hos2 and Hst1, and is a meiotic-specific repressor of the sporulation gene program. *Genes Dev*. 15, 2991–3004.

Pu, S., Vlasblom, J., Emili, A., et al. 2007. Identifying functional modules in the physical interactome of *Saccharomyces cerevisiae. Proteomics* 7, 944–960.

Resnik, P. 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intellig. Res.* 11, 95–130.

Rual, J.-F., et al. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178.

Ruepp, A., et al. 2008. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*. 36, D646–D650.

Rungsarityotin, W., Krause, R., Schodl, A., et al. 2007. Identifying protein complexes directly from high-throughput TAP data with markov random fields. *BMC Bioinform.* 8, 482.

Schlicker, A., Domingues, F.S., Rahnenführer, J., et al. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinform.* 7, 302.

Scholtens, D., Vidal, M., and Gentleman, R. 2005. Local modeling of global interactome networks. *Bioinformatics* 21, 3548–3557.

Stelzl, U., et al. 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968.

Uetz, P., et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* 403, 623–627.

van Dongen, S. 2000. Graph clustering by flow simulation [Ph.D. dissertation]. University of Utrecht.

Zhang, B., Park, B.-H., Karpinets, T., et al. 2008. From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics* 24, 979–986.

Address correspondence to:
Dr. Caroline C. Friedel
Institut für Informatik
Ludwig-Maximilians-Universität München
Amalienstraße 17
80333 München, Germany

E-mail: Caroline.Friedel@bio.ifi.lmu.de