BMC Bioinformatics

Predicting co-complexed protein pairs using genomic and proteomic data integration

BMC Bioinformatics 2004, 5:38

Lan V Zhang (lan_zhang@student.hms.harvard.edu)
Sharyl L Wong (sharyl_wong@student.hms.harvard.edu)
Oliver D King (oliver_king@hms.harvard.edu)
Frederick P Roth (fritz_roth@hms.harvard.edu)

ISSN 1471-2105

Article type Research article

Submission date 03 Nov 2003

Acceptance date 16 Apr 2004

Publication date 16 Apr 2004

Article URL http://www.biomedcentral.com/1471-2105/5/38

Predicting co-complexed protein pairs using genomic and proteomic data integration

Lan V. Zhang[1], Sharyl L. Wong[1], Oliver D. King[1] and Frederick P. Roth[1, *]

[1] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

[*] Corresponding author

Key words:

Protein-protein interaction, protein complex, decision tree, data integration, machine learning

Contact information:

Lan V. Zhang: lan_zhang@student.hms.harvard.edu
Sharyl L. Wong: sharyl_wong@student.hms.harvard.edu
Oliver D. King: oliver_king@hms.harvard.edu
Frederick P. Roth: fritz_roth@hms.harvard.edu

Abstract

Background:

Identifying all protein-protein interactions in an organism is a major objective of proteomics. A related goal is to know which protein pairs are present in the same protein complex. High-throughput methods such as yeast two-hybrid (Y2H) and affinity purification coupled with mass spectrometry (APMS) have been used to detect interacting proteins on a genomic scale. However, both Y2H and APMS methods have substantial false-positive rates. Aside from high-throughput interaction screens, other gene- or protein-pair characteristics may also be informative of physical interaction. Therefore it is desirable to integrate multiple datasets and utilize their different predictive value for more accurate prediction of co-complexed relationship.

Results:

Using a supervised machine learning approach — probabilistic decision tree, we integrated high-throughput protein interaction datasets and other gene- and protein-pair characteristics to predict co-complexed pairs (CCP) of proteins. Our predictions proved more sensitive and specific than predictions based on Y2H or APMS methods alone or in combination. Among the top predictions not annotated as CCPs in our reference set (obtained from the MIPS complex catalogue), a significant fraction was found to physically interact according to a separate database (YPD, Yeast Proteome Database), and the remaining predictions may potentially represent unknown CCPs.

Conclusions:

We demonstrated that the probabilistic decision tree approach can be successfully used to predict co-complexed protein (CCP) pairs from other characteristics. Our top-scoring CCP predictions provide testable hypotheses for experimental validation.

Background

Proteins are the major executors of the genetic program. Many proteins participate in cellular processes as members of protein complexes of varying size. It is believed that combinatorial interactions among proteins serve as an important basis for the biological complexity of higher organisms [1]. Therefore, increased knowledge about protein-protein interactions and protein complexes will greatly aid our understanding of protein function.

In recent years, there have been several large-scale efforts to map protein-protein interactions in yeast. The yeast two-hybrid (Y2H) system [2, 3] detects both transient and stable interactions. However, it suffers from high false-positive rate due to a number of factors such as fortuitous activation of reporter genes and self-activating "bait" proteins. False negatives are also inherent in the yeast two-hybrid system because of insufficient depth of screening, misfolding in the fusion proteins that abrogate the interactions, and use of full-length proteins that may mask interactions [3, 4]. In addition, both "bait" and "prey" proteins are over-expressed in the nucleus, so interactions detected may not be physiologically relevant [5], while certain interactions, for example, those involving membrane proteins and those requiring ancillary non-nuclear factors, may be undetectable [4]. Affinity purification coupled with mass spectrometry (APMS) has also been used to identify components of protein complexes on a large scale [6, 7]. Protein interactions identified in this way are more likely to be physiological, especially when tagged "bait" proteins are expressed under endogenous promoters [6]. Yet APMS is also subject to experimental error. Epitope tags may disable some protein interactions. Weakly associated components may dissociate and escape detection. Complexes containing transmembrane proteins are poorly detected while other condition-specific interactions may be missed [5]. Considering only interactions supported by more than one type of high-throughput evidence improves accuracy, but sacrifices sensitivity [5]. Therefore, more sophisticated methods are required to appropriately combine different high-throughput experimental datasets.

Integrating information beyond direct measurement of protein interactions could potentially improve the quality of protein interaction data as well. It has been shown that two proteins with similar mRNA expression profiles are more likely to interact with each other [8-12] (reviewed in [13]). Subcellular localization of proteins also provides information, since two interacting proteins usually reside in the same subcellular compartment [5, 14, 15]. Many other characteristics of a gene or protein pair might also have predictive value [16]. Although each characteristic alone may contain only limited information about whether a protein pair is co-complexed, many characteristics considered in combination may be more predictive.

Previously, there have been several efforts in integrating heterogeneous biological data types. Earlier studies addressed the question in a semi-manual and heuristic manner [17, 18]. More recently, the Support Vector Machine (SVM) algorithm has been applied to learning gene functions from two data types [19], which performs the task in an automated fashion. Bayesian networks have also been used to combine heterogeneous data sources [20, 21], and King *et al.* predicted gene function and knockout phenotype from patterns of annotation using a probabilistic decision tree approach [22, 23]. Probabilistic decision trees provide confidence levels of the predictions, as does Bayesian networks. In addition, the decision tree presents all the rules used in the prediction, making it easily interpretable which attribute combinations are most informative. When combining multiple biological data sources, learning the contributions of different attribute combinations can greatly help us to gain insight of the underlying biological

relationships, and therefore probabilistic decision trees represent an appropriate approach for this task.

Here we focused on the prediction of co-complexed protein (CCP) pairs in *Saccharomyces cerevisiae* and employed a probabilistic decision tree approach to integrate many gene- and protein-pair characteristics (see Table 1 and 2 for a summary and Additional file 1 for a complete list). A CCP pair is defined as a pair of proteins that belong to the same protein complex. Based on a training set, a probabilistic decision tree was generated and used to score protein pairs in a test set. High-scoring protein pairs by this approach represent predicted CCPs. Predictions were assessed by cross-validation according to a reference set based on the MIPS (Munich Information center for Protein Sequences) complex catalogue [24, 25]. Furthermore, top-scoring protein pairs not listed in MIPS as being co-complexed were validated by another database, YPD (Yeast Proteome Database) [26], at a significantly higher rate than expected by chance.

Results

We sought to combine a wide range of gene- and protein-pair characteristics using probabilistic decision trees to predict which protein pairs belong to the same complex. The approach was tested on the budding yeast *Saccharomyces cerevisiae*, for which extensive genomic and proteomic information is available. Data were obtained for a total of 467 gene- or protein-pair attributes, which were organized hierarchically and fell into 9 major categories (see Table 1 for a summary and Additional file 1 for more details). A reference set of 8707 CCPs was obtained from annotated protein complexes in MIPS [25]. We chose this literature-derived reference set as our "gold standard" because of its high reliability, but we note that this reference set is still imperfect since it reflects investigational bias that may lead us to predict fewer CCPs between uncharacterized proteins.

Probabilistic decision tree

To model the conditional probability that a protein pair is co-complexed given its other known attributes, we constructed a probabilistic decision tree using all protein pairs in *Saccharomyces cerevisiae* and all attributes listed in Table 1. The decision tree successively partitioned protein pairs according to the values (0 or 1) of their particular attributes. The structure of the tree was learned automatically, and the attribute used to define each successive partition was the attribute providing the greatest reduction of entropy with respect to the CCP attribute (see Methods section). Figure 1 shows the decision tree constructed using all attributes described in Table 1. Some of the rules specified in the decision tree capture biological knowledge about co-complexed proteins. For example, protein pairs in one high-scoring node (Figure 1, green arrowhead) are annotated with the attributes "TAP, 'spoke' model (I_APMS.TAP.spoke)" and "gene neighborhood (N)", which is consistent with the fact that the TAP study screens for protein complexes at a large scale [6], and the observation that proteins with conserved gene neighborhood are more likely to interact [5].

The attribute "bound by Fhl1p, $p<0.001$ (R_p001.FHL1)", describing putative regulation of genes by the transcription factor Fhl1p according to chromatin immunoprecipitation experiments, was chosen to make the first partition (shown as the root node in Figure 1), since this attribute yielded the greatest reduction in entropy. One might wonder why it is more informative than

high-throughput screens designed to assess protein-protein interactions. Note that our attribute selection criterion – conditional information gain – takes into consideration both accuracy and coverage. Although binding of Fhl1p does not provide information comprehensive enough to cover most of the yeast proteome, no existing evidence type is both very accurate and very comprehensive. Therefore it is not surprising that a relatively accurate attribute with a fair coverage becomes the winner. Fhl1p binds to the promoters of 194 genes at a p-value threshold of 0.001 [27], which translates to 18,721 protein pairs. This number is comparable to those of the APMS studies (26,742 for HMS-PCI [7] and 17,314 for TAP [6], and is significantly higher than those of the Y2H studies (4,475 and 948 [2, 3]). A significant portion (3,590 pairs) of the 18,721 protein pairs bound by Fhl1p are annotated as CCPs in our reference set, which should be regarded as relatively accurate considering the noisiness of the high-throughput interaction datasets. In addition, Fhl1p is believed to regulate the transcription of genes involved in rRNA processing [28], and many rRNA processing proteins, together with small nucleolar RNA's (snoRNA's), form a large RNP complex — the processome [29]. Many of the genes regulated by Fhl1p are likely to be actually members of the processome complex, therefore it is reasonable that the attribute "bound by Fhl1p, $p<0.001$ (R_p001.FHL1)" came out to be the variable most informative of co-complexed relationships.

Among attributes listed in Table 1, those that individually provide the greatest reduction in entropy at the root node are shown in Table 3. To compare this reduction with the entropy of the node before it is partitioned, we also describe relative reduction in entropy (defined as the conditional information gain divided by the entropy of the root node) for the top attributes. Relative reduction in entropy among the top 20 attributes ranges from 2.0% to 25.7%. Each of the 20 top-scoring protein-pair attributes shows significant positive correlation with CCP ($p<10^{-300}$ by Fisher's Exact Test, with multiple hypotheses adjusted for using the conservative Bonferroni correction). Most of these top attributes are from the categories "same transcriptional regulator," "correlated mRNA expression," and "high-throughput screens of interaction." This supports previous observations that co-complexed proteins are more likely to have correlated expression profiles and to have been identified in previous high-throughput interaction screens [5, 8, 9]. Yet it is worth noting that even attributes with low relative reduction in entropy at the root node could potentially be useful when combined with other attributes. For example, the relative entropy reduction provided by the attribute "bound by Grf10p, $p<0.005$" at the root node is only 0.0025%, but it is nevertheless used in the decision tree for an informative partition (Figure 1).

Table 4 lists the 61 attributes used in the decision tree shown in Figure 1. This list includes attributes from 8 of the 9 categories. Although some attributes never appear in the decision trees, this does not necessarily mean that they are not informative with regard to CCPs. Absence of an attribute may simply indicate that the information it provides is at least partially redundant with other attributes that are used in the tree.

Assessment using cross-validation

We used four-fold cross-validation to score each protein pair according to its estimated probability of being a CCP pair. Successively omitting one quarter of all protein pairs, we generated four decision trees, each very similar to the one generated using all protein pairs (data not shown). In the scoring procedure, a protein pair is mapped to a terminal or "leaf" node in the decision tree, whereupon it is assigned a probability of CCP calculated from the numbers of CCP

and non-CCP pairs in the training set that map to the same leaf node (see Methods section). True-positive rates (defined as the number of true positives divided by the total number of trues) and false-positive rates (defined as the number of false positives divided by the total number of falses) of the predictions were calculated at a series of score thresholds, and these values were used to plot a Receiver Operating Characteristic (ROC) curve, shown in Figure 2 at different resolutions. Note that a method making random guesses will have an expected ROC curve on the diagonal (*i.e.*, true-positive rate equals false-positive rate). Using our probabilistic decision tree approach, over 78.9% of CCPs are correctly predicted at a false-positive rate of 1% (Fig. 2B). Because experimentally testing a large number of protein pairs for CCP is both time-consuming and costly, predictions with many false-positives are not practically very useful. Given the ~20 million possibly interacting protein pairs in yeast, even a false-positive rate of 0.01 is likely to be unacceptable. Therefore, we focused on the part of our ROC curve where the false-positive rate is very low (~10^{-5}) (Fig. 2C). Among the top 83 predictions, 74 are known CCP pairs. At a false-positive rate of 5.4×10^{-5} (1125 false positives), the true-positive rate is 0.12 (1005 true positives). Different users of our predictions may have different levels of acceptable true-positive or false-positive rate. Our ROC curve allows users to tune predictions to suit their applications.

To assess the contribution of different datasets, we repeated the training and cross-validation procedures, successively omitting one category of attributes when constructing the decision trees (Fig. 2 and data not shown). Judging from the ROC curves, five out of the nine categories have little observable effect on the predictions when excluded (data not shown), and omission of each of the remaining four categories – "high-throughput screens (HTS) of interaction", "correlated mRNA expression", "same transcriptional regulator" and "sequence homology" – shows modest decrease in performance (Fig. 2). This indicates that most attributes are at least partially redundant with one or more attributes in another category. It also suggests that many strong predictions of CCP relationships can be made without direct evidence of physical interaction.

The MIPS database contains other types of information, such as protein function, protein class and subcellular localization, which may also be informative of CCP relationships. However, some of these annotations may be derived solely from physical interaction evidence, thereby resulting in circularity. With this substantial caveat in mind, we repeated training and cross-validation using attributes from three additional categories — "same subcellular localization (MIPS)", "same function (MIPS)" and "same protein class (MIPS)" (Table 2). The performance improves considerably with the addition of these attributes, with only 108 false positives (false-positive rate 5.2×10^{-6}) when 1015 true CCP pairs are predicted (true-positive rate 0.117) (Fig. 2, grey curve). At least part of the improvement came from non-circular evidence because not all of these annotations are derived from physical interactions. In addition, since these attributes can be used without risk of circularity for protein pairs not known to physically interact, this all-inclusive tree should be used to make predictions for such pairs.

To compare decision tree predictions with those of high-throughput experiments, we calculated true-positive and false-positive rate for predictions made by high-throughput interaction screens (two high-throughput APMS and two Y2H studies) (Fig. 3: A, B, C). Because APMS experiments use only a subset of genes as baits and therefore have not examined all possible protein pairs in the yeast proteome, we made two separate comparisons considering only protein pairs covered by each of the two APMS studies (using the "spoke" model, in which only bait-prey protein pairs are considered [30]) (Fig. 3: B, C). Comparison of the ROC curves shows that the decision tree approach based on a wide variety of evidence types is superior to any single

high-throughput method (Fig. 3: A, B, C). In addition, we compared our predictions with simple combinations of experimental evidence types. Since we are more concerned about predictions with low false-positive rates, we then focused on predictions supported by at least two high-throughput studies (Fig. 3A). Two other ROC curves are also plotted, one for decision tree predictions using only the four high-throughput interaction datasets and the other for predictions using all attributes together with attributes from the three additional categories "same function (MIPS)" and "same protein class (MIPS)" and "same subcellular localization (MIPS)" (Fig. 3: A, B, C). The decision tree approach using only high-throughput interaction datasets yields slightly better predictions than those generated by simple combinations of the same four datasets, and furthermore is more "tunable" to a desired true-positive or false-positive rate. Prediction success of the decision tree approach improves considerably after adding other genomic and proteomic information.

Assessment based on the Yeast Proteome Database (YPD)

Having demonstrated the success of our approach using cross-validation, we went further to see if we could predict CCPs not in the MIPS reference set. Among protein pairs not known to be CCP in the reference set, the top-scoring ones (predicted using all attributes in Table 1) were further examined. Since our reference set may not contain all known CCPs, especially the recently identified ones, some of these "false positives" might have already been tested and shown to be true CCPs. We searched for evidence of co-complexed relationships for these 50 "false positives" in a separate database, YPD [26]. YPD contains literature-based protein complex annotations and was not used as a data source in building our decision trees. We excluded YPD complexes for which interaction evidence comes solely from the high-throughput experiments used in our decision tree. Out of the top 50 "false positives," 15 are annotated in YPD as members of the same complex and are therefore true CCPs (Table 5, also see Table 1S in Additional file 2 for a longer list). This cannot be solely accounted for by the additional CCP annotations in YPD, because if the 50 protein pairs are randomly chosen among non-CCP pairs according to MIPS, the probability of seeing 15 or more pairs annotated with CCP in YPD is very low ($p<10^{-35}$ by Fisher's Exact Test). We also compared this result with two datasets: the TAP (tandem-affinity purification) APMS study [6] and the HMS-PCI (high-throughput mass spectrometric protein complex identification) APMS study [7]. For each dataset, we calculated the probability of finding 15 or more CCP pairs in YPD among protein pairs that show interaction according to the dataset of interest but are non-CCP in MIPS. By this measure, our approach showed slightly better performance than the TAP study alone ($p=0.2$), and significantly outperformed the HMS-PCI study alone ($p=2\times10^{-11}$).

As a comparison, we also performed the opposite experiment — using CCPs annotated in YPD as the gold standard in decision tree prediction. Cross-validation performance was comparable to that obtained using the MIPS reference set (Figure 1S in Additional file 3). Using the same false-positive rate threshold of 5×10^{-5}, predictions based on MIPS and YPD overlap by more than one third. Such an overlap is highly significant considering the size of the yeast proteome ($p<10^{-269}$ by Fisher's Exact Test), indicating that our approach is robust with regard to the gold standard used. Among the top-scoring 50 protein pairs not in the YPD reference set, 11 of them are annotated as CCPs in MIPS, comparable to the results shown earlier (15 out of 50). This is again highly significant ($p<10^{-28}$ by Fisher's Exact Test) given the null hypothesis that the

50 protein pairs are randomly chosen from non-CCP pairs according to YPD.

Discussion

Using a probabilistic decision tree approach, we were able to integrate a large number of gene- or protein-pair characteristics to predict co-complexed pairs of proteins. When evaluated by cross-validation, our method yielded more sensitive and specific predictions than the high-throughput interaction screens alone or in combination. However, we note that APMS experiments are not designed to examine pairwise interactions, and provide additional information about protein complexes that is not directly available from our approach. Furthermore, we do not suggest that interaction screens could be replaced by our approach. On the contrary, the success of our approach depends on the integration of such protein interaction datasets as well as other genomic and proteomic data types.

The reference set of CCPs used in this study derives from the MIPS complex catalogue [24] and may present a bias towards well-known proteins. Such a bias, if combined with attribute data with the same bias, may artificially inflate the performance in cross-validation. Since all attributes in Table 1 are from high-throughput or genome-wide studies, they contain little bias against unknown proteins. Therefore we expect our results using only these attributes (Figure 2 and 3, solid black lines, and Table 5) to accurately reflect the real method performance. The additional attributes listed in Table 2 are from collections of individual studies, and hence may be biased towards well-known proteins. As a consequence of such bias, as well as the potential circularity noted earlier, results obtained when the additional attributes in Table 2 were included (Figure 2 and 3, grey lines) may be artificially inflated.

One of the merits of the probabilistic decision tree approach is that for each protein pair, it provides a score which corresponds to the estimated probability that the protein pair is co-complexed. The collection of CCP probabilities for all protein pairs constitutes a weighted network of interactions in which the weight of each edge is the probability of interaction. Such a probabilistic interaction network presents a starting point for improved *ab initio* complex prediction [31].

The probabilistic interaction network can also be used to identify additional members of existing complexes. For example, according to the MIPS complex catalogue, the rRNA processing complex contains 18 proteins (Figure 4). Six additional proteins were found by our decision tree to be co-complexed with one or more of these 18 members with a score threshold of 0.5 (Figure 4). Three of them (Lcp5p, Mtr3p and Rrp40p) are verified in YPD. For the other three (Rrp1p, Srp1p and Cbf5p), each of them has been found to be associated with members of the rRNA processing complex in multiple affinity purifications in the high-throughput studies [6, 7]. Srp1p binds to nuclear localization sequences (NLS) in nuclear proteins to bring them to the nuclear pore complex [32], and therefore its association with proteins in the complex is more likely to be transient rather than stable. Cbf5p is involved in multiple uridine to pseudouridine conversions in rRNA [33] and Rrp1p is involved in maturation of rRNA [34]. Both of them are likely to be actual members of the rRNA processing complex. We expect that the probabilities generated here could be used to improve previously-described methods for discovering new members of partially-known protein complexes [35, 36].

Decision tree predictions can also be used to stratify individual interactions derived from the

high-throughput datasets by confidence. For each of the four APMS datasets (TAP spoke, TAP matrix, HMS-PCI spoke and HMS-PCI matrix), we partitioned the protein pairs based on scores from decision tree predictions. We found that the fraction of protein pairs in each subset that are annotated in YPD is correlated with the score (Figure 5). In general, a higher percentage of protein pairs are verified in a high-scoring subset than in a subset with low scores. Hence the score from decision tree prediction can serve as a good indicator of our confidence in the interaction and be used to further discriminate candidate CCP pairs resulting from high-throughput studies.

Integrating error-prone datasets and extracting useful information is an enormous challenge. For multiple evidence types with high false-positive and low false-negative rates, an obvious approach is to predict according to the intersection of all datasets. On the other hand, one might want to take the union if the evidence types have low false-positive rates but high false-negative rates. These two simple methods will be most effective if the evidence types are "orthogonal" [37], or more precisely, conditionally independent given the truth. However, these two extremes are not generally applicable in integrating multiple datasets related to protein interactions. Furthermore, most such datasets are not independent. Given the heterogeneous nature of various genomic data, it is desirable to develop more effective rules of data integration that can take into account the different predictive value of every data source and their combinations. One way to combine the different features of the datasets is to model the conditional probability of CCP given all gene- or protein-pair characteristics. A recent study combined evidence from six datasets by dividing protein pairs into 2^6 subsets according to combinations of evidence types and estimated error rate for each of them as the fraction of false positives in the subset [38]. However, such a method scales poorly as the number of datasets increases because the number of parameters (*i.e.* error rates) grows exponentially with the number of attributes, and is therefore highly prone to over-fitting. Here we took a probabilistic decision tree approach to tackle the problem. By post-pruning the decision trees, we were able to choose features informative of CCP and avoid over-fitting, and were therefore able to integrate a much larger number of gene- and protein-pair characteristics. Our method substantially outperformed the Jansen *et al.* 2002 approach. (There are 46 true positives and 37 false positives among the top 83 predictions in [38], evaluated on the training set, while our method, evaluated by cross-validation, predicted 74 true positives among the top 83 predictions.) This improvement demonstrates the benefit of integrating diverse data types to predict CCPs.

During the preparation of this manuscript, Jansen *et al.* published another related study using naïve Bayes and a fully-connected Bayesian network to combine multiple evidence types [20]. The naïve Bayes approach allows them to incorporate more evidence types than in their previous study [38], but assumes conditional independency between the attributes, which they justify by showing the lack of linear correlation between most of the attributes used. (But note that conditional independency does not follow the absence of linear correlation.) The results, however, are not directly comparable for at least three reasons. First, they use a "gold-standard" in which positives are defined by the MIPS complex catalogue (the same as in our study), but negatives are non-positive protein pairs with different subcellular localizations. This largely recasts the problem of CCP prediction as the problem of predicting protein pairs that either are co-complexed or share the same subcellular localization, which over-simplifies the task. Second, due to their choice of gold-standard negatives, their training set used in cross-validation is enriched with protein pairs for which both members have known subcellular localization and in

consequence the result does not represent their performance on the entire yeast proteome. Third, they use functional annotation to make their predictions, which has the potential for circularity (*e.g.*, if the function is actually assigned on the basis of CCP annotation in the "gold standard") and introduces a strong bias towards well-studied proteins, both of which may artificially inflate the performance.

Conclusions

A probabilistic decision tree approach has been previously used to predict some characteristics of genes or proteins (*e.g.*, knockout phenotype and protein function) [22, 23, 39]. Here we showed that a similar approach can also be used to predict a characteristic of protein pairs (*i.e.* co-complexed relationship) from other characteristics. CCP predictions provide testable hypotheses for experimental validation. The estimated CCP probabilities provided by integrating heterogeneous data with probabilistic decision trees may lead to improved *ab initio* complex discovery from interaction data [31] or to more accurate addition of proteins to partially-known protein complexes. Predicted CCP membership may also represent functional links between proteins, and therefore aid in the prediction of protein function. This general approach can be readily applied to other characteristics of gene or protein pairs and in other organisms as large-scale genomic and proteomic data becomes available.

Methods

Collecting datasets

We collected 12 major categories of gene- and protein-pair characteristics for all protein pairs in *Saccharomyces cerevisiae*. A summary with references to the data sources is shown in Table 1 and 2. Each evidence type was mapped to one or more binary variables ("attributes"). For an evidence type with continuous values (*e.g.*, expression correlation coefficient), a series of alternative thresholds were used to convert it into several binary attributes. All attributes were hierarchically organized into a directed acyclic graph (DAG), with an edge from attribute *i* to attribute *j* indicating that any protein pair annotated with attribute *j* is, by logical necessity, also annotated with attribute *i*.

A reference set of co-complexed protein pairs was obtained from the MIPS complex catalogue [24, 25] which provides a relatively complete list of currently known protein complexes in yeast. All protein pairs within the same complex were recorded as CCPs. Since the MIPS complex catalogue is organized into a hierarchy of complexes, we only considered complexes with no annotated sub-complexes. Altogether, our MIPS-derived reference set contains 8707 CCPs collected from a total of 250 complexes.

If a protein pair is not annotated with a particular attribute, it could be because previous study showed that it does not have the attribute (negative evidence), or because it has not been examined (absence of evidence). We did not make any distinction between these two scenarios since this information is typically unavailable. Similarly, no distinction was made between negative evidence and absence of evidence for CCP annotations.

Cross-validation:

All protein pairs were randomly partitioned into four subsets. In each of the four iterations, a probabilistic decision tree was constructed using training data composed of three out of the four subsets, successively leaving one out as the test set. Protein pairs in the test set were then scored according to the decision tree generated from the corresponding training data.

Generating decision trees

A detailed overview of decision trees and their applications can be found in [40, 41]. In our case, we started with all protein pairs of the training set *R* in a single root node, and constructed the decision tree greedily by recursively partitioning each node *N* into two daughter nodes based on the attribute *k* that gives the greatest reduction in entropy or, equivalently, the maximal conditional information gain. Let $Y_k(m)$ denote whether protein pair *m* is annotated with attribute *k*, and *X* be the random variable indicating whether a protein pair is annotated as a CCP. If node *N* is partitioned into two nodes N_0 and N_1 where $N_t = \{m \in N, Y_k(m) = t\}$, the conditional information gain is defined as:

$$H_N(X) - \sum_{t=0,1} \frac{|N_t|}{|N|} H_{N_t}(X).$$

Here $|N|$ represents the number of protein pairs within node *N*, and $H_N(X)$ is the entropy of *X* at node *N*, defined as $-p_N \log(p_N) - (1-p_N)\log(1-p_N)$, where p_N is the probability that a protein pair $m \in N$ is annotated as a CCP. We estimated p_N as the fraction of CCPs in node *N*, using one pseudocount (with the same CCP distribution as the entire training set *R*) for small-sample-size regularization.

A tree generated in the above fashion risks over-fitting the training data. The standard approach to combat this is post-pruning — pruning away some of the branches after the tree is grown [41]. We used the Bayesian Information Criterion (BIC) for model selection during pruning, as previously described [22]. After the tree was fully grown, we started from the leaves and pruned away any branch whose removal decreased the tree's BIC score. Such pruning dramatically reduced the size of the tree, hence the number of parameters, and avoided over-fitting the training data.

Scoring for co-complexed protein pairs

Protein pairs in each test set were scored according to the decision tree generated from the corresponding training set. Starting from the root node, the decision tree prescribes a series of binary questions for any given protein pair. All questions are of the form "Does the protein pair have attribute *j*?" Which question is asked depends on the answer to the previous question. After each question, the protein pair is assigned to one of the two daughter nodes, based upon whether or not it is annotated with attribute *j*. In the end, the protein pair is located to a leaf node *N*. The score of the protein pair is then the estimated probability p_N that a protein pair $m \in N$ is annotated with CCP, as described above.

List of abbreviations

CCP, co-complexed protein; ROC, Receiver Operating Characteristic; Y2H, yeast two-hybrid; APMS, affinity purification coupled with mass spectrometry; YPD, Yeast Proteome Database; MIPS, Munich Information center for Protein Sequences.

Authors' contributions

LVZ originated the idea of integrating multiple evidence types to predict protein interactions, conducted data collection, algorithm implementation and method assessment, and drafted the manuscript. SLW participated in data collection and implementation of the algorithm. ODK provided critical input on the methodology. FPR conceived of the study and directed the entire project. All authors participated in revising the manuscript, read and approved the final manuscript.

Acknowledgements

We thank F. Gibbons and G. Berriz for programming assistance and D. Goldberg and M. Vidal for helpful discussions. This work was sponsored in part by an institutional grant from the HHMI Biomedical Research Support Program for Medical Schools. L.V.Z. was supported in part by a Fu Fellowship. S.L.W. was supported in part by a Ryan Fellowship and by the Milton Fund of Harvard University. O.D.K was supported by an NRSA Fellowship from NHGRI.

References

1. Claverie JM: **Gene number. What if there are only 30,000 human genes?** *Science* 2001, **291**(5507):1255-1257.

2. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403**(6770):623-627.

3. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *Proc Natl Acad Sci U S A* 2001, **98**(8):4569-4574.

4. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y: **Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins**. *Proc Natl Acad Sci U S A* 2000, **97**(3):1143-1147.

5. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**(6887):399-403.

6. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**(6868):141-147.

7. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M: **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry**. *Nature* 2002, **415**(6868):180-183.

8. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions**. *Genome Res* 2002, **12**(1):37-46.

9. Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae**. *Nat Genet* 2001, **29**(4):482-486.

10. Deane CM, Salwinski L, Xenarios I, Eisenberg D: **Protein interactions: two methods for assessment of the reliability of high throughput observations**. *Mol Cell Proteomics* 2002, **1**(5):349-356.

11. Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, Holstege FC: **Protein interaction verification and functional annotation by integrated analysis of genome-scale data**. *Mol Cell* 2002, **9**(5):1133-1143.

12. Grigoriev A: **A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae**. *Nucleic Acids Res* 2001, **29**(17):3513-3519.

13. Ge H, Walhout AJ, Vidal M: **Integrating 'omic' information: a bridge between genomics and systems biology**. *Trends Genet* 2003, **19**(10):551-560.

14. Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast**. *Nat Biotechnol* 2000, **18**(12):1257-1261.

15. Xenarios I, Eisenberg D: **Protein interaction databases**. *Curr Opin Biotechnol* 2001, **12**(4):334-339.

16. Hazbun TR, Fields S: **Networking proteins in yeast**. *Proc Natl Acad Sci U S A* 2001, **98**(8):4277-4278.

17. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences**. *Science* 1999, **285**(5428):751-753.

18. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function**. *Nature* 1999, **402**(6757):83-86.

19. Pavlidis P, Weston J, Cai J, Noble WS: **Learning gene functional classifications from multiple data types**. *J Comput Biol* 2002, **9**(2):401-411.

20. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data**. *Science* 2003, **302**(5644):449-453.

21. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D: **A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8348-8353.

22. King OD, Foulger RE, Dwight SS, White JV, Roth FP: **Predicting gene function from patterns of annotation**. *Genome Res* 2003, **13**(5):896-904.

23. King OD, Lee JC, Dudley AM, Janse DM, Church GM, Roth FP: **Predicting phenotype from patterns of annotation**. *Bioinformatics* 2003, **19 Suppl 1**:I183-I189.

24. **MIPS complex catalogue** [http://mips.gsf.de/proj/yeast/catalogues/complexes/index.html]

25. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B: **MIPS: a database for genomes and protein sequences**. *Nucleic Acids Res* 2002, **30**(1):31-34.

26. Csank C, Costanzo MC, Hirschman J, Hodges P, Kranz JE, Mangan M, O'Neill K, Robertson LS, Skrzypek MS, Brooks J, Garrels JI: **Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD)**. *Methods Enzymol* 2002, **350**:347-373.

27. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *Science* 2002, **298**(5594):799-804.

28. Hermann-Le Denmat S, Werner M, Sentenac A, Thuriaux P: **Suppression of yeast RNA polymerase III mutations by FHL1, a gene coding for a fork head protein involved in rRNA processing**. *Mol Cell Biol* 1994, **14**(5):2905-2913.

29. Dragon F, Gallagher JE, Compagnone-Post PA, Mitchell BM, Porwancher KA, Wehner KA, Wormsley S, Settlage RE, Shabanowitz J, Osheim Y, Beyer AL, Hunt DF, Baserga SJ: **A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis**. *Nature* 2002, **417**(6892):967-970.

30. Bader GD, Hogue CW: **Analyzing yeast protein-protein interaction data obtained from different sources**. *Nat Biotechnol* 2002, **20**(10):991-997.

31. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4**(1):2.

32. Conti E, Uy M, Leighton L, Blobel G, Kuriyan J: **Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha**. *Cell* 1998, **94**(2):193-204.

33. Ni J, Tien AL, Fournier MJ: **Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA**. *Cell* 1997, **89**(4):565-573.

34. Hess SM, Stanford DR, Hopper AK: **SRD1, a S. cerevisiae gene affecting pre-rRNA processing contains a C2/C2 zinc finger motif**. *Nucleic Acids Res* 1994, **22**(7):1265-1271.

35. Bader JS: **Greedily building protein networks with confidence**. *Bioinformatics* 2003, **19**(15):1869-1874.

36. Asthana S, King OD, Roth FP: **Predicting protein complex membership using probabilistic network reliability**. *Genome Res* in press.

37. Gerstein M, Lan N, Jansen R: **Proteomics. Integrating interactomes**. *Science* 2002, **295**(5553):284-287.

38. Jansen R, Lan N, Qian J, Gerstein M: **Integration of genomic datasets to predict protein complexes in yeast**. *J Structural and Functional Genomics* 2002(2):71-81.

39. Vogel DS, Axelrod RC: **Predicting the effects of gene deletion**. *SIGKDD Explorations* 2002, **4**(2):101.

40. Quinlan JR: **C4.5 : programs for machine learning**. San Mateo, Calif.: Morgan Kaufmann Publishers; 1993.

41. Breiman L, Friedman JH, Olshen RA, Stone CJ: **Classification and regression trees**. Belmont, Calif.: Wadsworth International Group; 1984.

42. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH: **Functional discovery via a compendium of expression profiles**. *Cell* 2000, **102**(1):109-126.

43. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle**. *Mol Cell* 1998, **2**(1):65-73.

44. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y, Cheung KH, Miller P, Gerstein M, Roeder GS, Snyder M: **Subcellular localization of the yeast proteome**. *Genes Dev* 2002, **16**(6):707-719.

45. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

Figure legends

Figure 1:

Decision tree constructed using all protein pairs. Each leaf node is labeled with the numbers of CCPs and non-CCPs associated with it, while each internal node is labeled with the attribute (*j*) used for subsequent partitioning (see Table 4 or Supplementary Information for descriptions of the attributes). Two edges originate from each internal node, labeled "+" or "-," corresponding to the daughter nodes that have or do not have attribute *j*, respectively. Nodes with percentages of CCPs higher than that of the root node are colored red, while those with lower CCP percentages are blue. The color saturation depends on the relative entropy compared with the root node. The arrowhead size of an edge from a given node approximately represents the fraction of protein pairs in the parent node assigned to the corresponding daughter node.

Figure 2:

ROC curves for predictions based on: all attributes (black), all attributes except the category "high-throughput screens of interaction" (yellow), all attributes except the category "correlated mRNA expression" (green), all attributes except the category "same transcriptional regulator" (red), all attributes except the category "sequence homology" (blue) and all attributes together with the categories "same subcellular localization (MIPS)", "same function (MIPS)" and "same protein class (MIPS)" (grey). The expected ROC curve for random guesses is the diagonal where true-positive rate equals false-positive rate (black dotted line). A-C show the same ROC curve at different resolutions.

Figure 3:

A: Decision tree predictions compared with four high-throughput datasets and their simple combinations. B and C: Decision tree predictions compared with two APMS studies: HMS-PCI (B) and TAP (C), respectively. Only protein pairs covered by each respective study (using the "spoke" model [30]) were considered. Black solid line: decision tree predictions using all attributes; blue solid line: decision tree predictions using only high-throughput interaction datasets; grey solid line: decision tree predictions using all attributes together with the categories "same function" and "same protein class"; black dotted line: expected performance of random guesses.

Figure 4:

The rRNA processing complex with candidate members predicted by the decision tree. Red circles represent members of the complex annotated in MIPS. Green and yellow circles are proteins found to be co-complexed with the MIPS complex members by the decision tree with a score higher than 0.5. The yellow ones are verified in YPD while the green ones are not. The width of each edge is proportional to the decision tree score of the corresponding protein pair. Edges with scores lower than 0.1 as well as edges between the MIPS complex members are not shown.

Figure 5:

Correlation between scores from decision tree predictions and the fractions verified by YPD. For each of the four datasets (TAP spoke, TAP matrix, HMS-PCI spoke and HMS-PCI matrix), we plotted the fractions of its protein pairs at different score intervals that are also annotated in YPD.

Tables

Table 1. Categories of gene- and protein-pair attributes used

Attribute ID	Description	Number of Attributes	References
I.	High-throughput screens (HTS) of interactions	11	[2, 3, 6, 7]
X.	Correlated mRNA expression	23	[42, 43]
R.	Same transcriptional regulator	229	[27]
L.	Same subcellular localization (high-throughput)	16	[44]
P.	Same knockout phenotype	181	[25]
H.	Sequence homology	4	[45]
U.	Gene fusion	1	[5]
N.	Gene neighborhood	1	[5]
O.	Gene co-occurrence in phylogenetic profiles	1	[5]

Tabel 2: Additional categories of gene- and protein-pair attributes

Attribute ID	Description	Number of Attributes	References
S.	Same subcellular localization (MIPS)	43	[25]
F.	Same function (MIPS)	258	[25]
C.	Same protein class (MIPS)	191	[25]

Table 3: Top 20 attributes ranked by reduction in entropy provided by partitioning the root node

Attribute ID	Entropy Reduction	Relative Entropy Reduction	Attribute Description
R_p001.FHL1	9.5e-4	25.7%	Bound by Fhl1p, p<0.001
R_p005.FHL1	9.3e-4	25.3%	Bound by Fhl1p, p<0.005
X_cc.p.8	7.6e-4	20.7%	Correlated mRNA expression, cell cycle dataset, cc>0.8
X_cc.p.7	7.4e-4	20.0%	Correlated mRNA expression, cell cycle dataset, cc>0.7
X_cc.p.6	6.0e-4	16.2%	Correlated mRNA expression, cell cycle dataset, cc>0.6
R_p001	6.0e-4	15.9%	Same transcriptional regulator, p<0.001
R_p005.RAP1	5.0e-4	13.6%	Bound by Rap1p, p<0.005
X_cc	5.0e-4	13.4%	Correlated mRNA expression, cell cycle dataset
X	5.0e-4	13.4%	Correlated mRNA expression
R_p005	4.3e-4	11.6%	Same transcriptional regulator, p<0.005
I_APMS.TAP	3.0e-4	8.2%	TAP
R_p001.RAP1	3.0e-4	8.2%	Bound by Rap1p, p<0.001
I_APMS	2.7e-4	7.3%	APMS
I	2.7e-4	7.3%	High-throughput screens (HTS) of interactions
I_APMS.TAP.spoke	1.5e-4	4.1%	TAP, "spoke" model
X_cc.p.9	1.4e-4	3.7%	Correlated mRNA expression, cell cycle dataset, cc>0.9
X_Rst.p.6	1.2e-4	3.3%	Correlated mRNA expression, Rosetta compendium, cc>0.6
N	1.2e-4	3.2%	Gene neighborhood
X_Rst	1.1e-4	2.8%	Correlated mRNA expression, Rosetta compendium
I_APMS.HMS-PCI	7.3e-5	2.0%	HMS-PCI

Table 4: Attributes used in the decision tree

Attribute ID	Attribute Description
I	High-throughput screens (HTS) of interaction
I_APMS.TAP	Tandem-affinity purification (TAP)
I_APMS.TAP.spoke	Tandem-affinity purification (TAP), "spoke" model
I_APMS.HMS-PCI	High-throughput mass spectrometric protein complex identification (HMS-PCI)
I_APMS.HMS-PCI.spoke	High-throughput mass spectrometric protein complex identification (HMS-PCI), "spoke" model
I_Y2H	Yeast two-hybrid (Y2H)
I_Y2H.Uetz	Yeast two-hybrid (Y2H), Uetz *et al.*
X	Correlated mRNA expression
X_Rst	Correlated mRNA expression, Rosetta compendium
X_Rst.p	Positively correlated mRNA expression, Rosetta compendium
X_Rst.p.8	Correlated mRNA expression, Rosetta compendium, cc>0.8
X_cc.p	Positively correlated mRNA expression, cell cycle dataset
X_cc.p.7	Correlated mRNA expression, cell cycle dataset, cc>0.7
X_cc.p.8	Correlated mRNA expression, cell cycle dataset, cc>0.8
X_cc.p.9	Correlated mRNA expression, cell cycle dataset, cc>0.9
R	Same transcriptional regulator
R_p005.ABF1	Bound by Abf1p, p<0.005
R_p005.GRF10	Bound by Grf10p, p<0.005
R_p005.HAP4	Bound by Hap4p, p<0.005
R_p005.RAP1	Bound by Rap1p, p<0.005
R_p005.RME1	Bound by Rme1p, p<0.005
R_p005.SFP1	Bound by Sfp1p, p<0.005
R_p005.SWI4	Bound by Swi4p, p<0.005
R_p005.YAP5	Bound by Yap5p, p<0.005
R_p001.FHL1	Bound by Fhl1p, p<0.001
R_p001.HAP4	Bound by Hap4p, p<0.001
R_p001.HIR2	Bound by Hir2p, p<0.001
R_p001.RAP1	Bound by Rap1p, p<0.001
R_p001.REB1	Bound by Reb1p, p<0.001
L	Same subcellular localization (high-throughput)
L_05	ER
L_08	Mitochondrial
L_10	Nucleus
L_04	Cytoplasm
P	Same Phenotype
P_1	Conditional phenotypes
P_1.1	Heat-sensitivity
P_1.3	Slow-growth
P_2	Cell cycle defects
P_2.4	Other cell cycle defects
P_4.2	Methionine auxotrophy
P_4.5.4	Respiratory deficiency
P_5	Cell morphology and organelle mutants
P_5.2.5	Other budding mutants
P_5.3	Cell wall mutants
P_5.6.1	Tubulin cytoskeleton mutants
P_5.6.1.5	Other tubulin cytoskeleton mutants
P_5.6.2	Actin cytoskeleton mutants
P_5.9	Secretory mutants
P_5.11	Mitochondrial mutants
P_5.13.2	Other vacuolar mutants
P_5.14	Other cell morphology mutants
P_8	Nucleic acid metabolism defects
P_8.1	DNA repair mutants
P_8.1.1	UV light sensitivity
P_8.2	DNA replication mutants
P_9.9	Staurosporine sensitivity
H	Sequence homology, E<e-6
H.e-12	Sequence homology, E<e-12
N	Gene neighborhood
O	Gene co-occurrence

Table 5: Top predictions not annotated as CCPs in the reference set

The 50 top-scoring protein pairs not annotated in our reference set (so-called “false positives”) with results of a further search for pre-existing evidence of CCP. 15 of them are shown to be true CCPs according to YPD.

Rank	Protein 1	Protein 2	Score	YPD Complex Annotation
1	Rpl40Bp	Rps31p	0.943	
2	Rps31p	Rpl40Ap	0.938	
3	Smc1p	Smc3p	0.864	Cohesin
4	Gpt2p	Sec28p	0.857	
5	Pwp2p	Utp13p	0.844	Small subunit processome
5	Sgn1p	Pub1p	0.844	
7	Rdh54p	Rad5p	0.833	
7	Arp3p	Rvs167p	0.833	
7	Arp3p	Srv2p	0.833	
10	Spt5p	Rpb3p	0.800	Paf1p complex
10	Spt5p	Rpo21p	0.800	Paf1p complex
12	Pwp2p	Dip2p	0.776	Small subunit processome
12	Pwp2p	Ylr409C	0.776	
12	Sap190p	Sap155p	0.776	
12	Sap190p	Sap185p	0.776	
12	Pph21p	Pph22p	0.776	
12	Nop7p	Fpr4p	0.776	
12	Sap185p	Sap155p	0.776	
12	Sik1p	Cbf5p	0.776	
12	Nop2p	Ebp2p	0.776	Pre-60S ribosomal particle
12	Rpa135p	Ret1p	0.776	
22	Pwp2p	Asc1p	0.750	
22	Drs1p	Spb4p	0.750	
24	Rsm10p	Mrps5p	0.744	Mrp4p-associated complex (mitochondrial ribosome)
24	Mtr3p	Rrp45p	0.744	Exosome 3'-5' exoribonuclease complex
24	Rrp40p	Rrp46p	0.744	Exosome 3'-5' exoribonuclease complex
24	Rrp40p	Ski6p	0.744	Exosome 3'-5' exoribonuclease complex
28	Fun12p	Cbf5p	0.743	
28	Mrpl16p	Yml025Cp	0.743	
28	Mrpl1p	Mrpl9p	0.743	
28	Mrpl9p	Ypl183C-Ap	0.743	
28	Rrp40p	Rrp45p	0.743	Exosome 3'-5' exoribonuclease complex
33	Gin4p	Kcc4p	0.727	
33	Ecm16p	Prp43p	0.727	
35	Rps27Ap	Rpl42Bp	0.714	
35	Rps17Bp	Rpl36Ap	0.714	
35	Rps4Ap	Rpp2Ap	0.714	
35	Dur1,2p	Pdb1p	0.714	
35	Rsm7p	Mrps5p	0.714	Mrp4p-associated complex (mitochondrial ribosome)
40	Pat1p	Lsm2p	0.692	mRNA decay complex
40	Hrp1p	Nab2p	0.692	
42	Mrpl1p	Mrpl10p	0.684	
42	Mrpl9p	Yml025Cp	0.684	
44	Lsm2p	Dhh1p	0.667	45S penta-snRNP
44	Pat1p	Dhh1p	0.667	45S penta-snRNP
46	Dyn1p	Cdc55p	0.667	
46	Emp24p	Fks1p	0.667	
46	Yef3p	Act1p	0.667	
46	Yef3p	Pph22p	0.667	
46	Asc1p	Tfp1p	0.667	

Description of Additional files

1. ccp-supp1.doc
   Microsoft Word document
   List of data sources and attributes used.

2. ccp-supp2.doc
   Microsoft Word document
   Table 1S: A list of 1000 top-scoring protein pairs not found in the MIPS reference set, together with YPD annotations (where available).

3. ccp-supp3.doc
   Microsoft Word document
   Figure 1S: Comparison of cross-validation performance using MIPS or YPD as the gold standard.

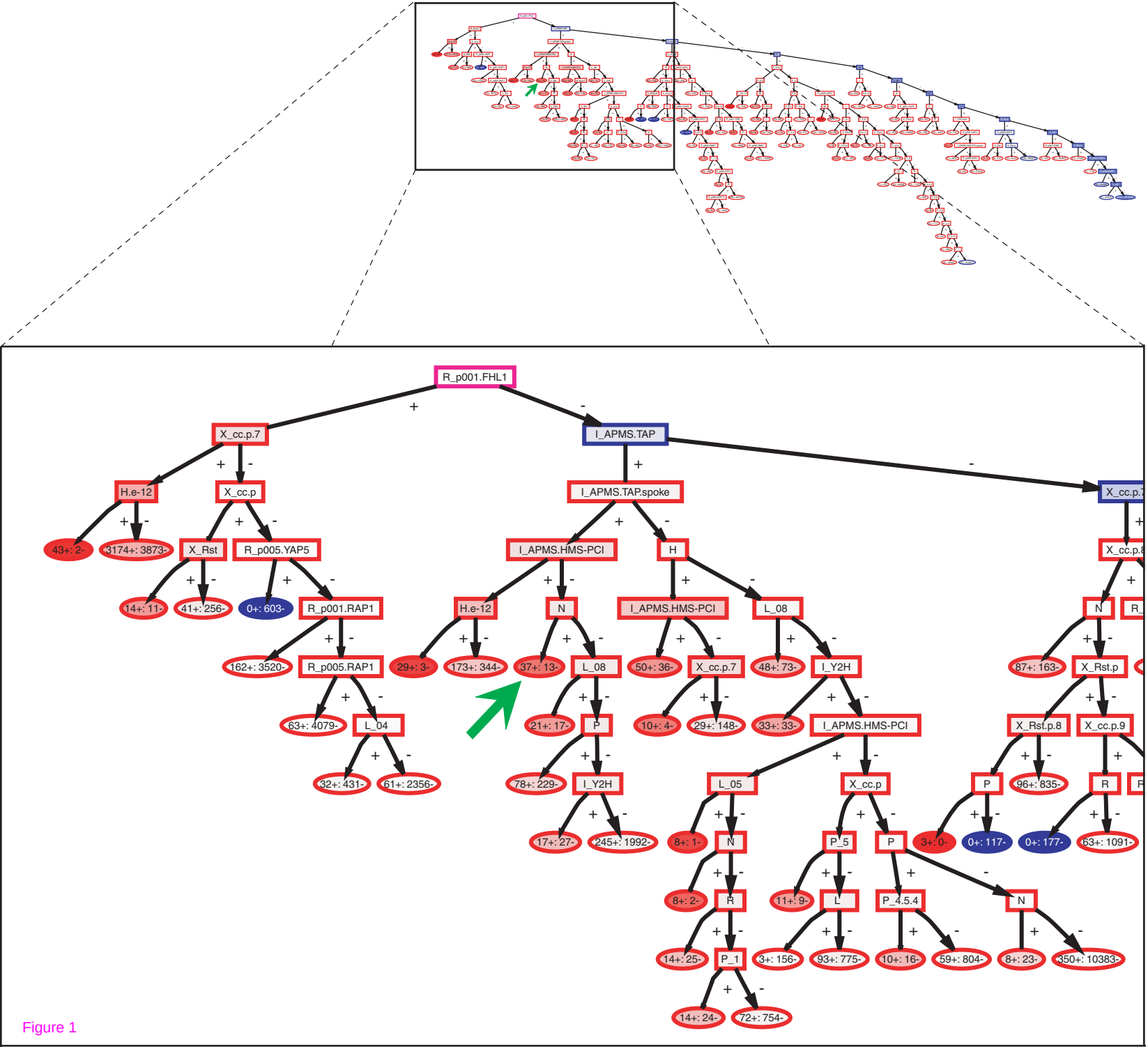

R_p001.FHL1
X_cc.p.7
I_APMS.TAP
H.e-12
X_cc.p
I_APMS.TAP.spoke
43+: 2-
3174+: 3873-
X_Rst
R_p005.YAP5
I_APMS.HMS-PCI
H
14+: 11-
41+: 256-
0+: 603-
R_p001.RAP1
H.e-12
N
I_APMS.HMS-PCI
L_08
162+: 3520-
R_p005.RAP1
29+: 3-
173+: 344-
37+: 13-
L_08
50+: 36-
X_cc.p.7
48+: 73-
I_Y2H
63+: 4079-
L_04
21+: 17-
P
10+: 4-
29+: 148-
33+: 33-
I_APMS.HMS-PCI
32+: 431-
61+: 2356-
78+: 229-
I_Y2H
L_05
X_cc.p
17+: 27-
245+: 1992-
8+: 1-
N
P_5
P
8+: 2-
R
11+: 9-
L
P_4.5.4
N
14+: 25-
P_1
3+: 156-
93+: 775-
10+: 16-
59+: 804-
8+: 23-
350+: 10383-
14+: 24-
72+: 754-

Figure 1

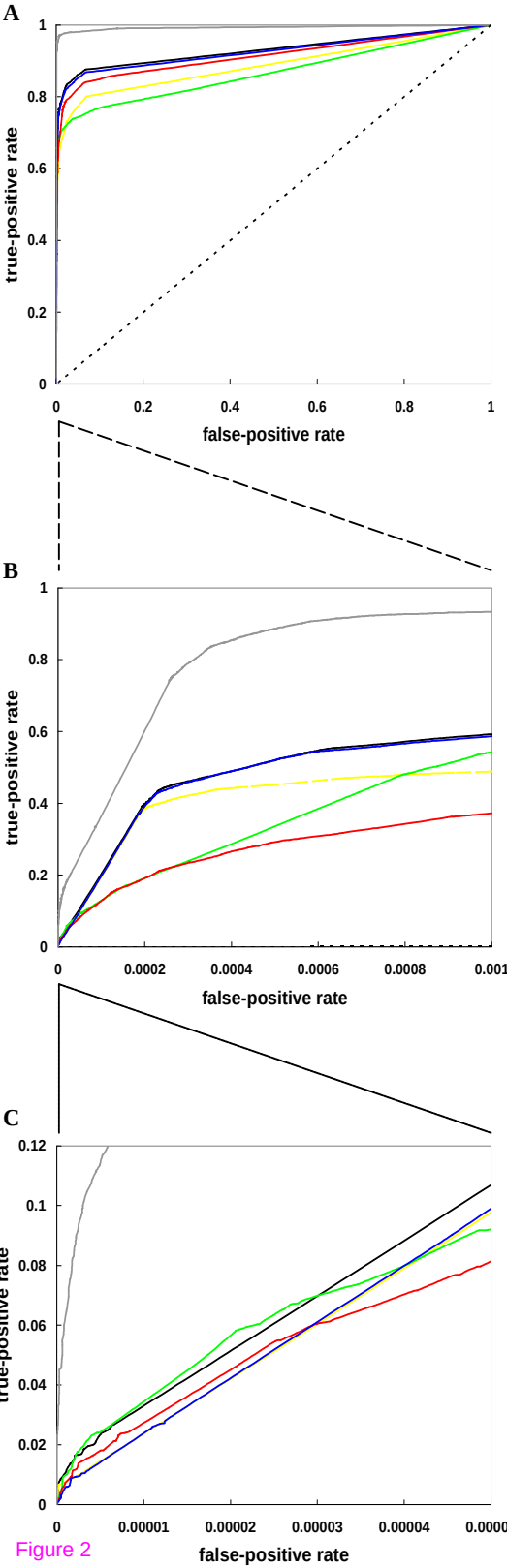
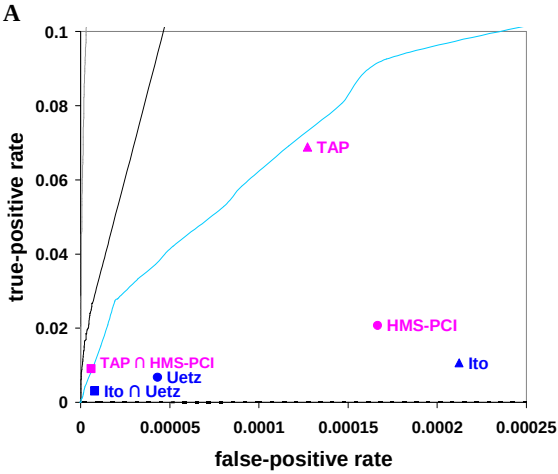
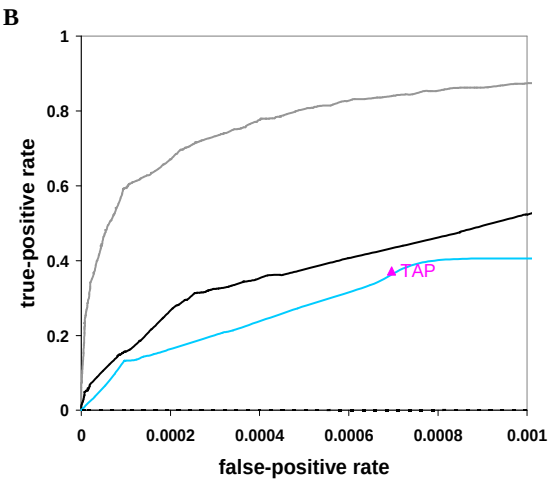

A
true-positive rate
0
0.2
0.4
0.6
0.8
1
false-positive rate
B
true-positive rate
0
0.2
0.4
0.6
0.8
1
0.0002
0.0004
0.0006
0.0008
0.001
false-positive rate
C
true-positive rate
0
0.02
0.04
0.06
0.08
0.1
0.12
0.00001
0.00002
0.00003
0.00004
0.00005
false-positive rate

Figure 2


A
true-positive rate
0.1
0.08
0.06
0.04
0.02
0
TAP
HMS-PCI
Ito
TAP ∩ HMS-PCI
Uetz
Ito ∩ Uetz
0
0.00005
0.0001
0.00015
0.0002
0.00025
false-positive rate


B
true-positive rate
1
0.8
0.6
0.4
0.2
0
TAP
0
0.0002
0.0004
0.0006
0.0008
0.001
false-positive rate

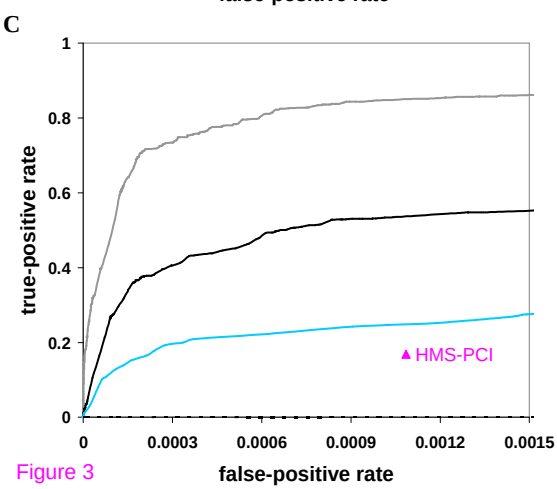

C
true-positive rate
1
0.8
0.6
0.4
0.2
0
HMS-PCI
0
0.0003
0.0006
0.0009
0.0012
0.0015
false-positive rate

Figure 3

Figure 4

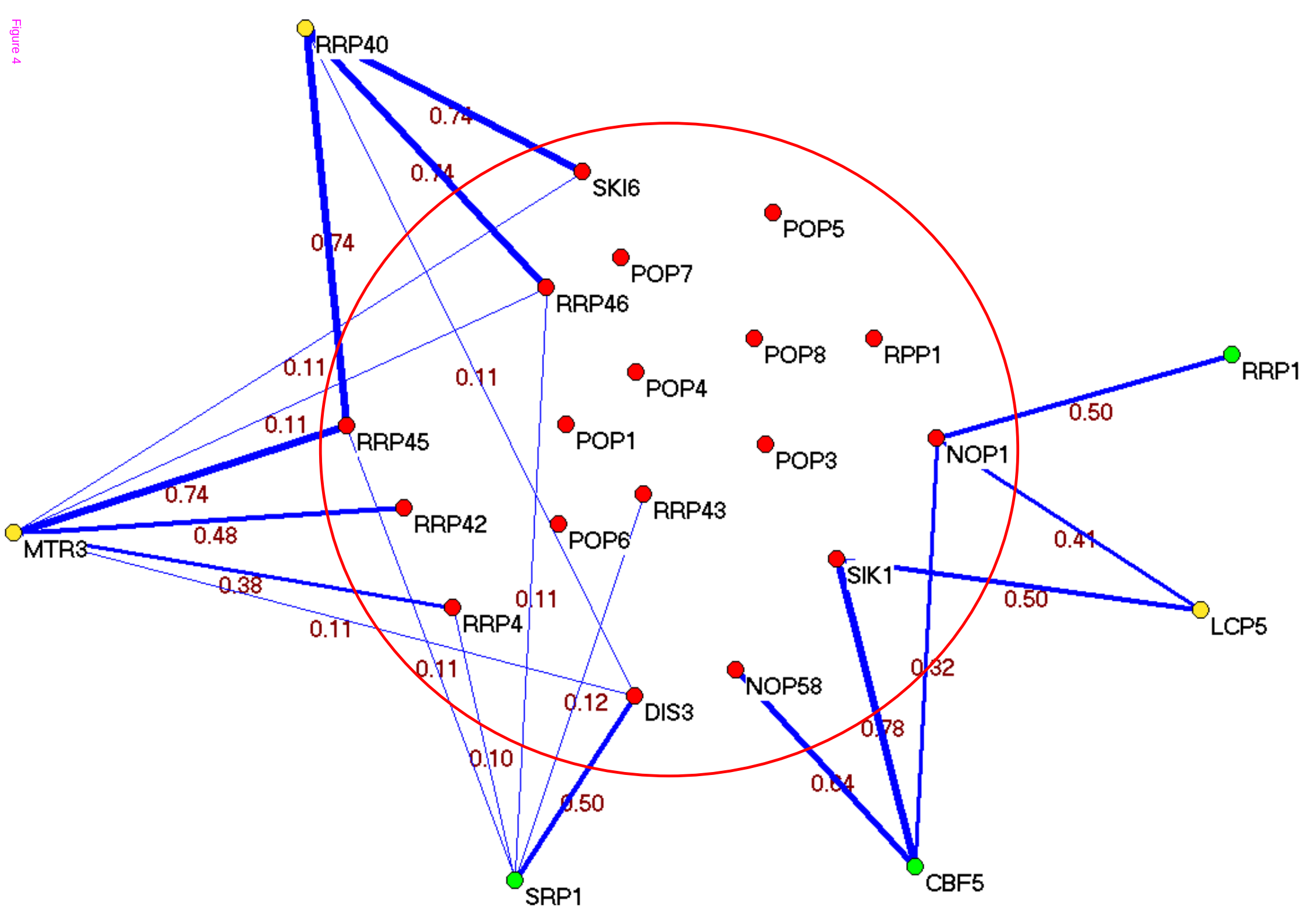
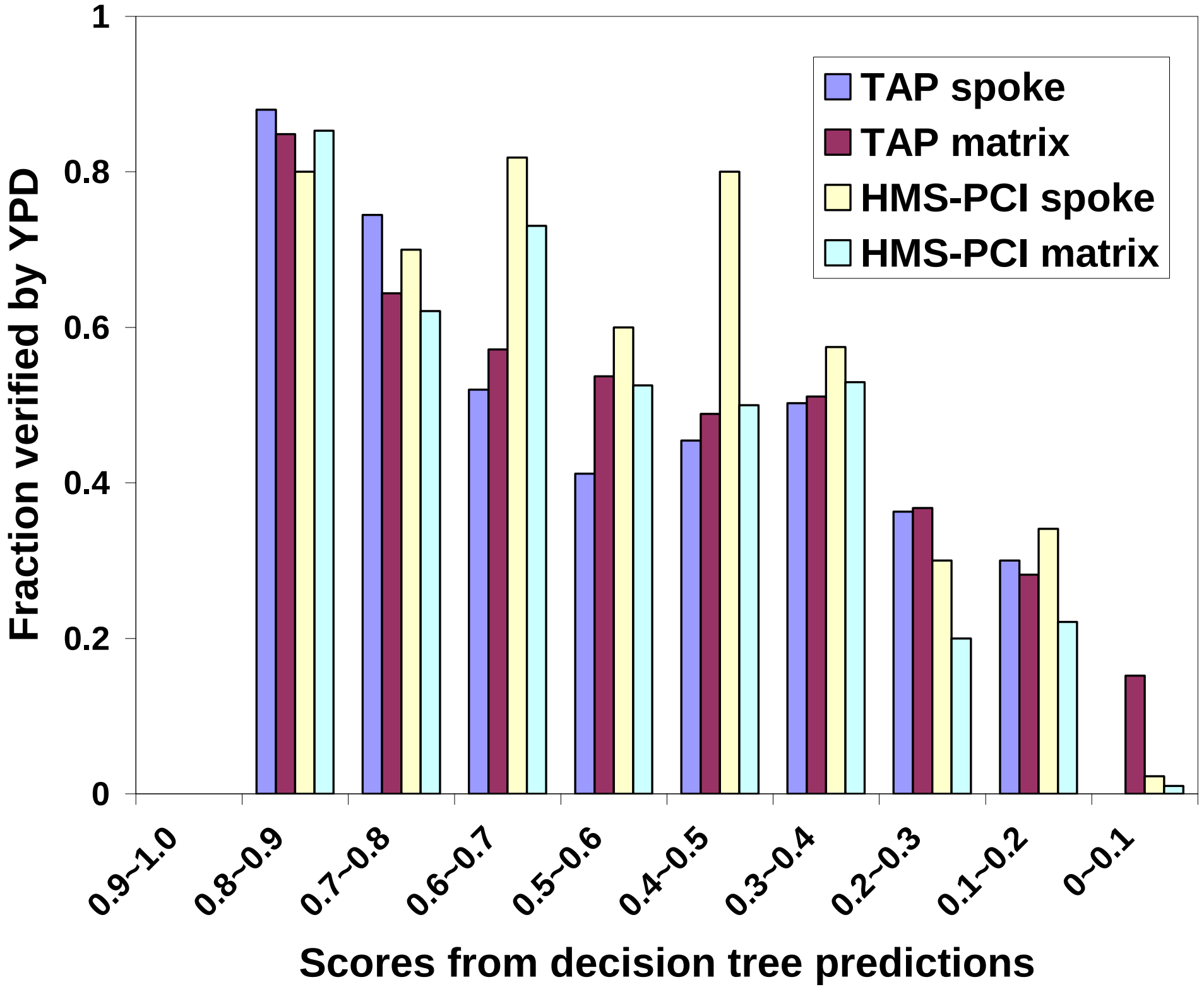

RRP40
SKI6
POP5
POP7
RRP46
POP8
RPP1
RRP1
POP4
POP1
POP3
RRP45
NOP1
RRP43
RRP42
POP6
MTR3
SIK1
LCP5
RRP4
NOP58
DIS3
CBF5
SRP1
0.74
0.74
0.74
0.11
0.11
0.11
0.74
0.48
0.38
0.11
0.11
0.12
0.10
0.50
0.50
0.41
0.50
0.32
0.78
0.64

Figure 5


TAP spoke
TAP matrix
HMS-PCI spoke
HMS-PCI matrix
Fraction verified by YPD
1
0.8
0.6
0.4
0.2
0
0.9–1.0
0.8–0.9
0.7–0.8
0.6–0.7
0.5–0.6
0.4–0.5
0.3–0.4
0.2–0.3
0.1–0.2
0–0.1
Scores from decision tree predictions

Additional files provided with this submission:

Additional file 1: ccp-supp1.doc : 158KB
http://www.biomedcentral.com/imedia/1810026314292040/sup1.doc

Additional file 2: ccp-supp2.doc : 1050KB
http://www.biomedcentral.com/imedia/7821935562920480/sup2.doc

Additional file 3: ccp-supp3.doc : 24KB
http://www.biomedcentral.com/imedia/1392316741309459/sup3.doc