BIOINFORMATICS

Vol. 19 no. 15 2003, pages 1901–1908
DOI: 10.1093/bioinformatics/btg344

A comprehensive set of protein complexes in yeast: mining large scale protein–protein interaction screens

Roland Krause[1,*], Christian von Mering[2] and Peer Bork[2]

[1]Cellzome AG, Meyerhofstraße 1, 69117 Heidelberg, Germany and [2]European Molecular Biology Laboratory, 69012 Heidelberg, Germany

Received on May 1, 2003; revised and accepted on July 15, 2003

ABSTRACT

Motivation: The analysis of protein–protein interactions allows for detailed exploration of the cellular machinery. The biochemical purification of protein complexes followed by identification of components by mass spectrometry is currently the method, which delivers the most reliable information—albeit that the data sets are still difficult to interpret.

Consolidating individual experiments into protein complexes, especially for high-throughput screens, is complicated by many contaminants, the occurrence of proteins in otherwise dissimilar purifications due to functional re-use and technical limitations in the detection. A non-redundant collection of protein complexes from experimental data would be useful for biological interpretation, but manual assembly is tedious and often inconsistent.

Results: Here, we introduce a measure to define similarity within collections of purifications and generate a set of minimally redundant, comprehensive complexes using unsupervised clustering.

Availability: Programs and results are freely available from http://www.bork.embl-heidelberg.de/Docu/purclust/

Contact: roland.krause@cellzome.com

INTRODUCTION

Protein–protein interactions provide a wealth of information on fundamental aspects of cellular life, and are also a powerful tool for target selection in drug discovery. The yeast *Saccharomyces cerevisiae* is a model organism for higher eukaryotes and several different methods for protein–protein interaction screens, even proteome wide, can be performed in this organism. The first such screens used yeast two-hybrid (Y2H) technology (Uetz *et al*., 2000; Ito *et al*., 2001). Recently, two studies of biochemical purifications combined with mass spectrometry (MS) were conducted: one using 'high-throughout MS protein complex identification' (HMS–PCI) (Ho *et al*., 2002), the other employing 'Tandem affinity purification (TAP) followed by MS identification' (Gavin *et al*., 2002).

Several attempts to compare, integrate and evaluate data from the high-throughput screens and other data sources of protein–protein interaction have been undertaken and revealed many aspects of their quality, fallacies and novelties (e.g. Bader and Hogue, 2002; Edwards *et al*., 2002; von Mering *et al*., 2002). They further showed that HMS–PCI and TAP are currently the most reliable high-throughput studies of protein–protein interactions.

Both methods employ modification of selected genes to express fusion proteins. The introduced protein or protein fragment (often referred to as bait) can be used to isolate the protein biochemically under mild conditions, so proteins binding the fusion protein are also retrieved. Finally, mass spectrometers are used to identify the purified proteins qualitatively. In the context of this study, we refer to purifications as the results of the identification, which can substantially differ from what is thought to exist in the cell and what is annotated as protein complexes in reference databases. Typically, one needs to perform manual post-processing on the purifications to create bona fide protein complexes.

Such purifications can contain many potential contaminant proteins, most of which are likely non-specifically bound, typically abundant enzymes, ribosomal proteins and chaperones. Often, biochemical function and detection frequency are used as filters to remove such proteins. This strategy is unsatisfactory as the contaminants themselves form complexes, such as the ribosome, and because the same protein can be a bona fide member of a complex in some purifications but a contaminant in others. An example of the latter is actin, a specific and stochiometric interactor in nuclear histone acetylation complexes (Olave *et al*., 2002) and a frequent contaminant due to its abundance in the cytosol. In many purifications the number of unspecific proteins appears to be greater than the number of 'true' interactors. Furthermore, the experiments do not always retrieve the whole complex but sub-complexes only; also protein identification by MS can be of limited sensitivity, especially for components that are of low molecular weight

[*] To whom correspondence should be addressed.

or for those expressed at higher concentrations than others. In small-scale studies, such issues can be addressed by assessing the protein concentration, but this has not been possible for biochemical high-throughput screens so far. Note that quantitative technology is becoming available and future screens might be able to provide the information (Ranish *et al*., 2003).

However, even if we would have the perfect identification tools at our disposal, the interactome is difficult to describe by a simple, static list of its interactions due to the complicated and partially redundant arrangement of the cellular machinery. For instance, several protein complexes exist which are very similar, such as the Sm-like complexes which share six out of seven components (Bouveret *et al*., 2000), but can be considered separate entities with different biological functions.

Another complicating issue is the spatial and temporal context dependency of protein–protein interactions as protein complexes undergo changes throughout their lifecycle. Examples are the assembly of the spliceosome from its protein–RNA-complexes U1 through U6 and various cofactors, the assembly and subsequent activity of the transcription machinery around RNA polymerase II, and the maturation and nuclear export of the ribosomal subunits. Y2H screens and high-throughput biochemical purifications cannot resolve the above issues since the cells in the studies are usually neither synchronized in their cell cycle nor separated into organellar fractions.

The databases for protein–protein interactions employ various approaches to represent information from the different experimental techniques and varying resolutions of spatial and temporal behavior for protein complexes. Databases, such as DIP (Xenarios *et al*., 2002) and BIND (Bader *et al*., 2003) store many details about the experiments, and LiveDIP (Duan *et al*., 2002) can even model most details of biological complexes, if sufficient experimental data are available. Simplified and less redundant databases, such as YPD (Csank *et al*., 2002) and the MIPS collection of protein complexes (Mewes *et al*., 2002) are popular, because they underwent manual curation; they frequently serve as reference sets for bioinformatics comparisons.

The combination of biological redundancy and methodological lack of resolution limits the practical use of the current data sets for molecular biologists to some extent. It would be useful to summarize the results into a concise list of protein complexes for biological interpretation. This reduces the redundancy and creates a platform for in-depth analysis and further experiments.

To this end, Gavin *et al*. (2002) used a combination of data on existing complexes and manual curation to assemble the raw data of 588 purifications of their TAP screen into 232 complexes, thus enabling subsequent in-depth annotation and analysis. Unfortunately, the outcome contains several artifacts such as complexes that are contained fully in others. Some purifications retrieved the same biochemical complex but were grouped into separate annotated complexes. Additionally, some complexes were solely assembled on the use of external information and rather contradict the experimental source data (see for instance the purification of Isw1 and Isw2 in complex 116). The HMS–PCI screen was curated manually only for some purifications (for illustration purposes). A procedure to define protein complexes automatically would be less laborious and easily repeatable when new data sets are made available.

Automated integration, complex prediction and comparison of interaction data have so far relied mostly on unweighted binary protein–protein interactions. However, the graphs resulting from these binary interactions fail to represent some aspects of protein complexes, in particular the sharing of components between protein complexes that joins complexes into a higher order network, the presence of highly similar complexes and the modularity of sub-complexes.

The use of binary interactions is certainly a good choice when comparing data provided by methods which primarily produce this type of information, such as Y2H screens (Uetz *et al*., 2000; Ito *et al*., 2001), genetic screens and *in silico* predictions (Marcotte, 2000; Huynen *et al*., 2003; Osterman and Overbeek, 2003) Also, it is possible to find protein complexes in binary protein–protein interaction graphs, e.g. using the MCODE algorithm (Bader and Hogue, 2003). However, the raw experimental data from biochemical purifications contains more information, which we want to represent and use for defining complexes.

In this study, we explored the strategy of using unsupervised clustering methods (complete clustering, means clustering and single-linkage clustering) on biochemical purifications and successfully assembled biologically meaningful complexes (see Fig. 1 for an overview of the approach). Similar techniques have been used to cluster mRNA expression profiles [see review by Slonim (2002)] and to find superpositions of other data sources, such as expression data and protein–protein interaction data (Steffen *et al*., 2002; Washburn *et al*., 2003).

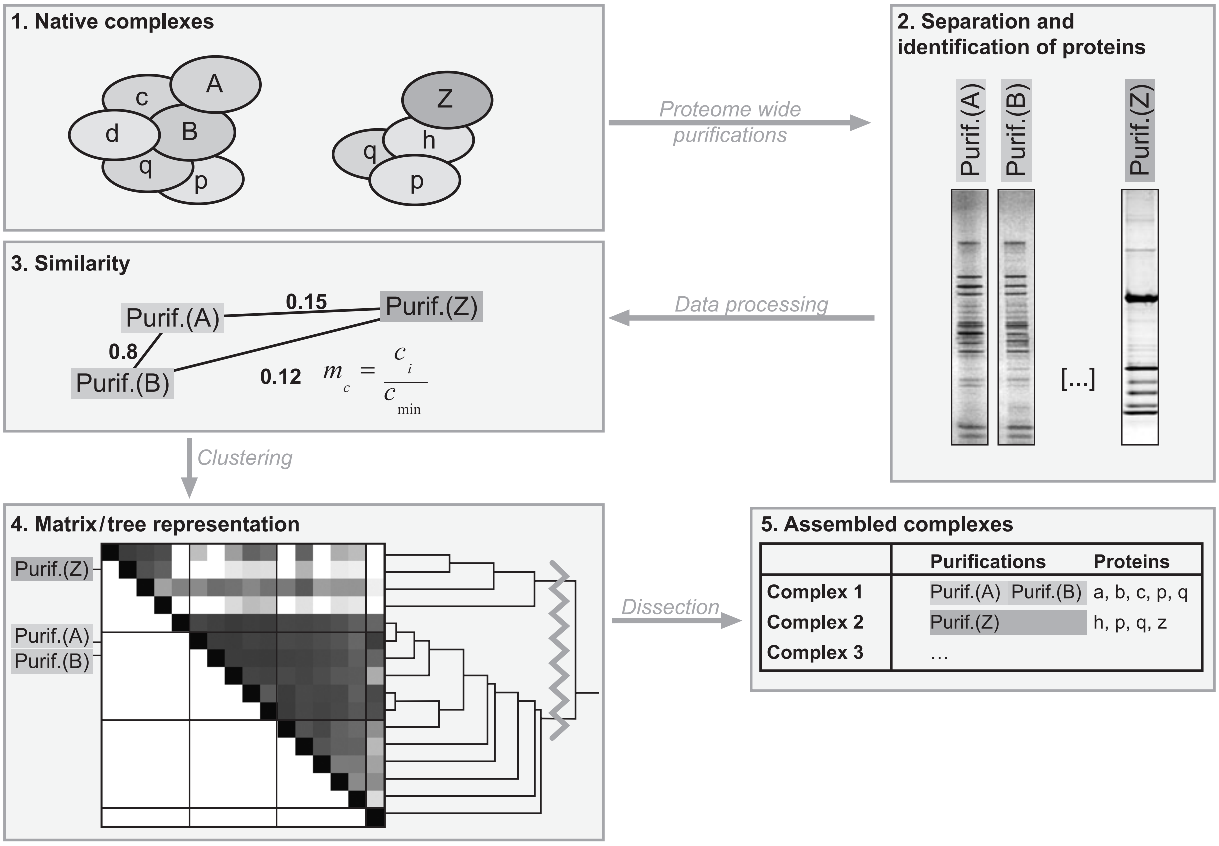
To find the most suitable clustering algorithm and its parameters, we started by assembling the TAP purifications and compared the individual clustering results with the MIPS set of known complexes. Additionally, we applied the procedure to the HMS–PCI data, and produced an inclusive and comprehensive complex annotation from the combined data set.

We compared this set with a prediction of protein complexes by MCODE, which employs a graph-based algorithm on all available protein–protein interaction data in yeast.

METHODS AND DATA SETS

Similarity measures

For the application of clustering techniques a measure of similarity m between samples containing several proteins



	Purifications	Proteins
Complex 1	Purif.(A) Purif.(B)	a, b, c, p, q
Complex 2	Purif.(Z)	h, p, q, z
Complex 3	…	

Fig. 1. Overview of the approach. m_c is the weighted Simpson coefficient (see Methods and Data Sets section).

(e.g. purifications or protein complexes) is needed. The simplest measure is to count proteins contained in both samples. As this does not take the size of the sample into account, it performs poorly on real data sets. This is due to the number of components identified that varies significantly, often largely increased by spurious contaminants.

Bader and Hogue (2003) have suggested using the geometric similarity index

$$m_\omega = \frac{n_i^2}{n_a \cdot n_b} \qquad (1)$$

with n_a and n_b being the number of proteins in the individual complexes A and B, and n_i being the number of proteins in the intersection I.

Several other suitable measures have been proposed in similar contexts such as

$$\text{the Jaccard coefficient} \quad m_j = \frac{n_i}{n_a + n_b - n_i}, \qquad (2)$$

$$\text{the Dice coefficient} \quad m_d = \frac{2n_i}{n_a + n_b}, \qquad (3)$$

and

$$\text{the Simpson coefficient} \quad m_s = \frac{n_i}{\min(n_a, n_b)} \qquad (4)$$

(Goldberg and Roth, 2003).

The performance of the similarity measures was tested on several cases that were studied in the manual assembly of the TAP data set by assessing the rank of a pair of purifications semi-manually (data not shown). In the absence of a ranked reference data set, an evaluation by e.g. a Wilcoxon rank sum test turned out difficult to perform.

We wanted to account for the presence of sub-complexes and stochiometric effects. Complex components expressed higher than others tend to co-purify with fewer proteins, which is strongly demonstrated by purifications of Rvb1 or Ssa1 in the HMS–PCI and the TAP set respectively. Such features are visible in gel images but no assessment of abundance was used in either study.

To reduce the contribution of frequently detected, potentially contaminant proteins in high-throughput sets, we decided to weight the proteins by their inverse detection frequency f in the respective data set to correct for abundance when comparing experimental data and finally used the following measure:

$$m_c = \frac{\sum_{i\in I} 1/f_i}{\sum_{c\in C} 1/f_c} \quad \text{for } C = \left|\begin{array}{ll} |A| & n_a < n_b \\ |B| & \text{otherwise.} \end{array}\right. \qquad (5)$$

The measure m_c can be used to cluster purifications within a large-scale experimental data set. However, for subsequent benchmarking between different data sets, we did not always want to consider complexes of different sizes as similar and for some comparisons instead used a variant Dice similarity, also corrected by the inverse frequency in the combined data sets to yield the following measure:

$$m_r = \frac{2\sum_{i\in I} 1/f_i}{\sum_{a\in A} 1/f_a + \sum_{b\in B} 1/f_b}. \qquad (6)$$

The benchmark—the MIPS collection of protein complexes

To benchmark our results, we compared them with the MIPS collection of protein complexes (Mewes *et al*., 2002). However, the MIPS data set is itself redundant due to several very similar complexes reported in the literature. Consequently, we decided to remove redundancy in the MIPS set and deleted all protein complexes, which appeared as part of larger complexes entirely, and created super-complexes if two complexes differed only slightly. This resulted in 173 entities containing a total of 1059 proteins with 999 unique proteins. No complex displays a m_r of more than 0.8.

Experimental data

The mass spectrometers used for analysis of biochemical purifications can identify proteins in samples which do not give rise to a detectable band in a gel; however, without more information these underrepresented proteins appear as valid interactors. In small-scale experiments these proteins are often removed by manual inspection, usually based on the protein concentration and the resulting stochiometry.

Prior to publication of the high-throughput screens, some obvious contaminants were removed by the original authors and we relied on the filtered data sets from the TAP and the HMS–PCI screens.

For the analysis of the TAP data set we removed purifications that only retrieved the bait (singletons), leaving 454 purifications. These purifications contain 3854 instances of 1361 proteins.

The HMS–PCI results are available in several formats. We considered a data set as downloaded on March 30, 2003 from the MDSP web site (http://www.mdsp.com/yeast/MDSP-Nature10Jan02-YeastComplexes.txt). This set contains information on the individual purifications and varies in part from the data in the original publication. Finally, we used the union of all purifications for any given bait as source for our analysis, resulting in 494 purifications with 4182 identifications of 1472 proteins.

TAP annotation

The complexes defined by manual annotation contain one singleton complex, which was removed; the remaining 231 complexes were used as published (Gavin *et al*., 2002).

MCODE

A large-scale prediction of protein complexes has been generated by the MCODE algorithm (Bader and Hogue, 2003), using all available data as of late 2002. We selected this predicted set for a comparison. As only few redundancies exist in this set of 209 complexes, no processing was necessary.

Clustering

We explored three standard procedures for hierarchical clustering: means (UPGMA) clustering, complete clustering and single-linkage clustering. The cutoff used is the smallest similarity, which would group two or more purifications into one cluster.

IMPLEMENTATION

Data manipulation and computing of similarity measures was implemented in Perl. The procedure was tested on an 800 MHz Intel computer, running Linux and using Perl 5.6.0, but should run on all platforms supporting Perl as no platform-dependent routines are used. Clustering was performed using the program oc (available at http://www.compbio.dundee.ac.uk/Software/OC/oc.html). The clustering is performed within seconds, so running times are not a limiting step.

We used the CASTA format (Bader and Hogue, 2003) to store protein purifications and complexes.

RESULTS

The interactome displays a much greater variability than the genome. One cannot expect any single method to produce one canonical set of protein complexes, as several different solutions can represent the molecular arrangements well. Our goal is not to define 'the true set' but rather to create a guide map, which approximates the complex biological reality of protein–protein interactions, allowing for further detailed inspections of high-throughput purifications.

As an illustration for a case with more than one biologically 'correct' solution consider the polyadenylation complex which contains, amongst other factors, two multiprotein sub-complexes, called cleavage factor I (CFI) and cleavage and polyadenylation factor (CPF). Many of its proteins were tagged in the TAP screen and by other groups (Roguev *et al*., 2001; Vo *et al*., 2001; Proudfoot and O'Sullivan, 2002). Depending on clustering parameters, our method groups CF1 and CPF either into separate complexes or in the same cluster—both solutions can be justified.

As with many other unsupervised clustering procedures, benchmarking the results is not always straightforward.

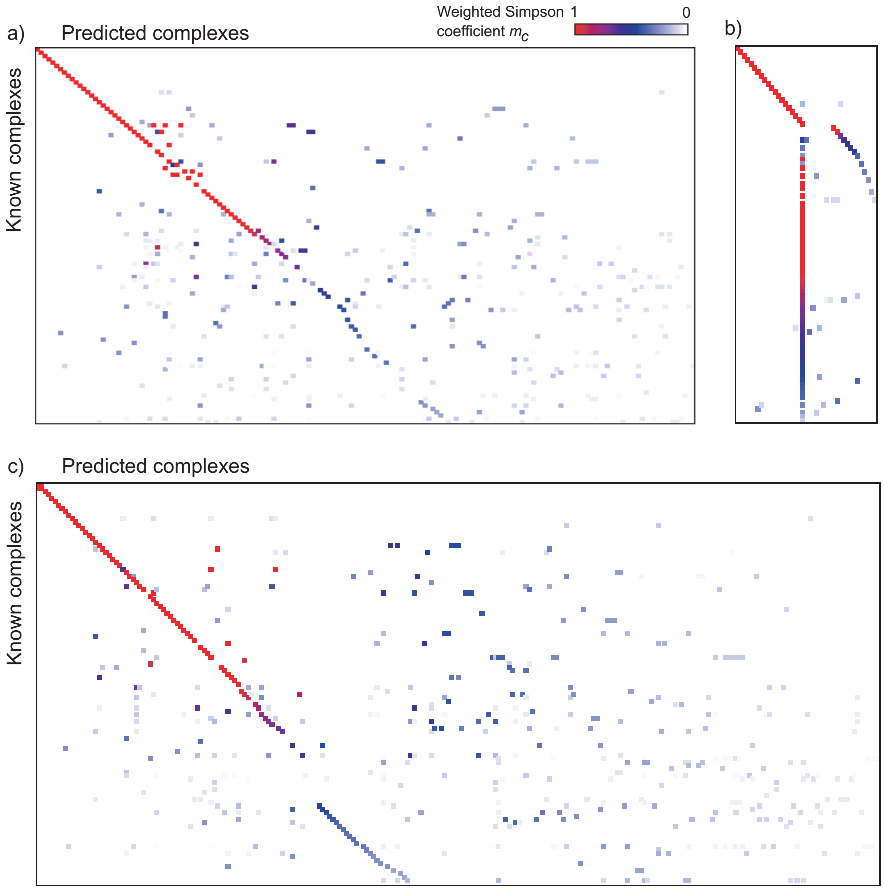


Fig. 2. Similarity between known complexes and three different clustering solutions. **(a)** Solution with a good fit—most predicted complexes match one reference complex. (means clustering, cutoff 0.3). **(b)** Underprediction of clusters—many MIPS complexes are matched by the same cluster (single linkage clustering, 0.125). **(c)** Overprediction. Many predicted complexes match the same MIPS complex (complete clustering, 0.7). The order of the complexes in rows and columns is by quality of best matches. All complexes with a match of at least one protein are shown. Note that the same number of MIPS complexes are matched in (**a**) and (**c**), but the number of predicted complexes matched is increased by about 30%, displayed by the different width.

Simply counting matches when comparing each predicted cluster against each complex in the MIPS data set would not be a useful criterion: the individual purifications already match protein complexes irrespectively of whether they are assigned to a cluster or not. If each cluster corresponds to one purification only, one would generate a maximum of matches; however, this solution is also maximally redundant. Likewise, it is not informative to compare the resulting clusters with features of their functional annotation to assess the quality of the solution.

Instead, we defined the following criteria to assess the fit of our prediction to the benchmark data set:

(1) The number of clusters matching MIPS complexes should be maximal.

(2) The number of clusters matching an *individual* MIPS complex should be one.

(3) Each cluster should map to one MIPS complex only. Clusters matching more than one complex are possibly predicted too inclusive.

(4) Complexes should have a similar average size and size distribution to the benchmark data set.

Taken together, the criteria 1, 2 and 3 would require a one-to-one correspondence between predicted clusters and MIPS complexes to be fulfilled completely. As real data contains false negatives and false positives and more than one reference complex can be contained in a purification, it will not be possible to meet all requirements simultaneously.

We performed a parameter exploration using the TAP data set and different clustering algorithms (complete, means and single-linkage clustering), comparing the results to the MIPS complex set, to find the combination of algorithm and cutoff that performed best. Figure 2 displays clustering solutions of different quality.

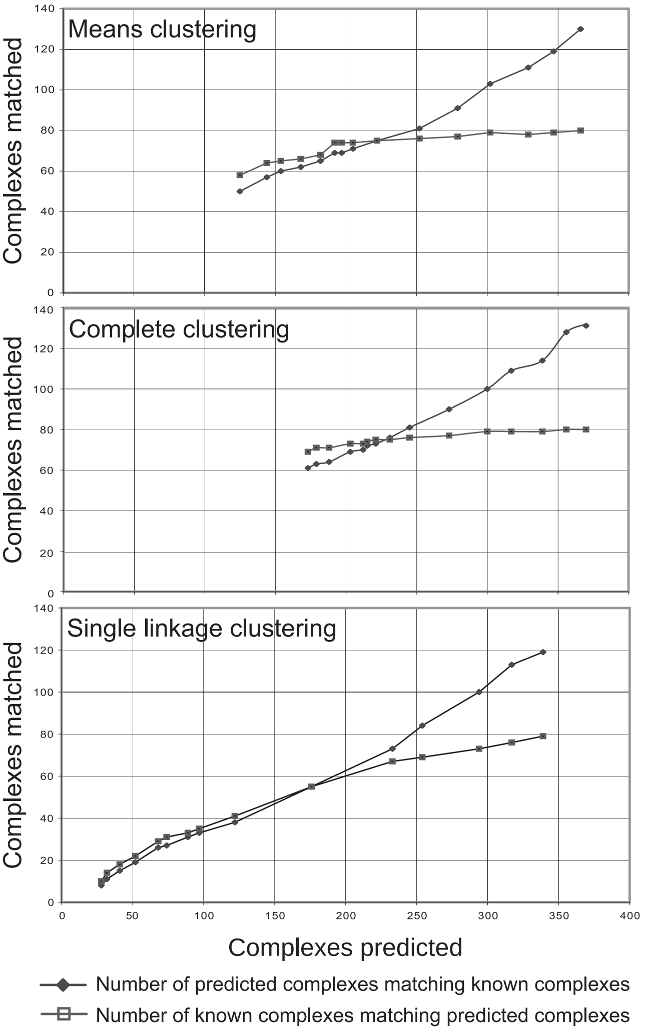


Fig. 3. Parameter exploration. Comparison of predicted and known protein complexes. Complexes were considered to match if their m_r is greater than 0.2 and they share at least two proteins. The weighted dice coefficient m_r was used since we require the complexes under consideration to be similar in content and size.

Table 1. Size distribution of predicted complexes from different data sets

	No. of purifications	Clusters	Average	Median
TAP	454	252	10.12	6
HMS–PCI	494	370	10.34	6
Combined	948	545	11.26	7

Settings that gave rise to more than ∼250 predicted complexes resulted in several predicted complexes matching the same MIPS complex with no further rise in the total number of MIPS complexes matched (Fig. 3). This range contained many unclustered purifications which should be grouped together with others into a cluster.

Single-linkage clustering was found to perform poorly. The predicted clusters tend to contain unrelated purifications because joining is done by closest similarity. Means clustering and complete clustering produced better and similar results. After manual inspection of solutions for individual complexes and overall behavior (Fig. 2), we selected means clustering with a cutoff 0.3. We believe that this solution is more representative than the initial manual annotation of the TAP purifications.

Using the parameters, several complexes containing multiprotein sub-complexes, such as the polyadenylation and the RNA cleavage factors or the large and the small subunits of the mitochondrial ribosome are resolved as distinct complexes. However, these sub-complexes are merged at lower cutoff levels. Note that these complexes have been merged in the original TAP annotation.

The comparison with the original, manual TAP assembly showed that 130 predicted assemblies were identical to complexes in the manual annotation; another 49 had two or more baits in common. The manual annotation, however, assigned some purifications to more than one complex.

We subsequently performed the clustering on the HMS–PCI data set and on the combined data sets. The results of the HMS–PCI screen and the TAP screen have been compared with each other (Bader and Hogue, 2002; von Mering *et al.*, 2002). We can expect a few discrepancies from stochiometric effects due to overexpression (in the HMS–PCI screen) or lack thereof (in the TAP screen) and due to different sensitivities of the mass spectrometers used. Still, it was found that both data sets display a surprisingly poor fit when the 94 purifications of the same bait protein were compared against each other.

We found 46% of equivalent purification using the same bait in the two data sets in the same cluster using the above settings—more than one would expect from the initial comparisons.

The program MCODE is a felicitous approach for using a combined set of data sources to predict protein complexes. When comparing the results from an MCODE prediction utilizing all available yeast data, we find 126 MCODE complexes that overlapped with 157 complexes from our clustering by at least two proteins and a m_r of at least 0.2 and 61 MCODE complexes with 61 from our prediction when applying a threshold of 0.4. This discrepancy appears to be sizeable; however, the MCODE prediction takes a very different approach to find complexes and while the data resulting from the biochemical purifications are included, the data representation is different. Moreover, our approach can map proteins into many protein complexes easily while the MCODE algorithm is designed to do so on a limited level only. Both approaches have their strengths and weaknesses: MCODE deals well with data from several sources, but also requires several individual interactions to predict a complex and it is difficult to match the results back to the experimental source data. Our approach works with biochemical purifications only and neglects the data from Y2H and other sources so far. However, it delivers a connection between individual complexes and the original experimental data, which is helpful for subsequent assessment and interpretation. The strong matches between the two sets of predicted complexes are often well described in the literature, suggesting that both algorithms perform well if sufficient experimental data are available.

Generally, complexes from the high-throughput screens appeared larger than ones from individual experiments. This was expected for several reasons, namely that contaminant proteins, which appeared in many purifications, would raise the background and because a single bait can purify more than one complex.

When examining the results we find that the complexes predicted for the HMS–PCI data are generally larger than the ones obtained from the TAP set. This can be explained by more sensitive MS procedures in HMS–PCI study, which has been noted before (Bader and Hogue, 2002; von Mering *et al*., 2002). The different background proteins generate an increased complex size when both data sets are joined.

CONCLUSION AND OUTLOOK

We describe a method for clustering biochemical purifications from high-throughout screens to create a concise list of protein complexes quickly, which correspond well to complexes described in the literature. We introduce a new similarity measure for protein purifications that reduces effects of contaminants in the raw experimental data. The initial assembly of the TAP set was done manually and took considerable time, whereas our method can be run in a few seconds. We also provide the first assembly of the HMS–PCI set and the joined set, which could help both individual assessments as well as large-scale comparisons with other data sets. For future screens, the method can be used to assess the novelty of particular purifications and to provide in-process information on the complexes discovered.

A further improvement could be the incorporation of the information about the protein that was used as a bait, as the interaction signal between bait and retrieved protein is more reliable than the signal between two retrieved proteins (Bader and Hogue, 2002; von Mering *et al*., 2002).

When comparing these approaches with clustering in mRNA expression analysis, it is clear that to predict protein complexes one needs to identify more clusters from a lot less data, thus making the results less reliable by comparison. Fortunately, more protein–protein interaction data is quickly becoming available and the method we describe here will allow a quick and comprehensive survey for future high-throughput screens for protein complexes.

ACKNOWLEDGEMENTS

We thank Jens M. Rick and Jörg Schultz for pioneer work in clustering purifications, and Evgeny Zdobnov, Julien Gagneur and Matthieu Pichaud for advice with clustering and comparisons.

REFERENCES

Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.

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

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Bouveret,E., Rigaut,G., Shevchenko,A., Wilm,M. and Seraphin,B. (2000) A Sm-like protein complex that participates in mRNA degradation. *EMBO J.*, **19**, 1661–1671.

Csank,C., Costanzo,M.C., Hirschman,J., Hodges,P., Kranz,J.E., Mangan,M. *et al.* (2002) Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol.*, **350**, 347–373.

Duan,X.J., Xenarios,I. and Eisenberg,D. (2002) Describing biological protein interactions in terms of protein states and state transitions: the LiveDIP database. *Mol. Cell. Proteomics*, **1**, 104–116.

Edwards,A.M., Kus,B., Jansen,R., Greenbaum,D., Greenblatt,J. and Gerstein,M. (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.*, **18**, 529–536.

Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Goldberg,D.S. and Roth,F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci. USA*, **100**, 4372–4376.

Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Huynen,M.A., Snel,B., von Mering,C. and Bork,P. (2003) Function prediction and protein networks. *Curr. Opin. Cell Biol.*, **15**, 191–198.

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

Marcotte,E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.*, **10**, 359–365.

Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

Olave,I.A., Reck-Peterson,S.L. and Crabtree,G.R. (2002) Nuclear actin and actin-related proteins in chromatin remodeling. *Annu. Rev. Biochem.*, **71**, 755–781.

Osterman,A. and Overbeek,R. (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.*, **7**, 238–251.

Proudfoot,N. and O'Sullivan,J. (2002) Polyadenylation: a tail of two complexes. *Curr. Biol.*, **12**, R855–R857.

Ranish,J.A., Yi,E.C., Leslie,D.M., Purvine,S.O., Goodlett,D.R., Eng,J. and Aebersold,R. (2003) The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.*, **33**, 349–355.

Roguev,A., Schaft,D., Shevchenko,A., Pijnappel,W.W., Wilm,M., Aasland,R. *et al.* (2001) The *Saccharomyces cerevisiae* Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *EMBO J.*, **20**, 7137–7148.

Slonim,D.K. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.*, **32**(suppl.), 502–508.

Steffen,M., Petti,A., Aach,J., D'Haeseleer,P. and Church,G. (2002) Automated modelling of signal transduction networks. *BMC Bioinformatics*, **3**, 34.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R. *et al.* (2000) A comprehensive analysis ofprotein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Vo,L.T., Minet,M., Schmitter,J.M., Lacroute,F. and Wyers,F. (2001) Mpe1, a zinc knuckle protein, is an essential component of yeast cleavage and polyadenylation factor required for the cleavage and polyadenylation of mRNA. *Mol. Cell. Biol.*, **21**, 8346–8356.

von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.

Washburn,M.P., Koller,A., Oshiro,G., Ulaszek,R.R., Plouffe,D., Deciu,C. *et al.* (2003) Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **100**, 3107–3112.

Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.