Challenges and Rewards of Interaction Proteomics*

Shoshana J. Wodak‡§¶‖, Shuye Pu‡**, James Vlasblom‡, and Bertrand Séraphin‡‡§§

The recent explosion of high throughput experimental technologies for characterizing protein interactions has generated large amounts of data describing interactions between thousands of proteins and producing genome scale views of protein assemblies. The systems level views afforded by these data hold great promise of leading to new knowledge but also involve many challenges. Deriving meaningful biological conclusions from these views crucially depends on our understanding of the approximation and biases that enter into deriving and interpreting the data. The challenges and rewards of interaction proteomics are reviewed here using as an example the latest comprehensive high throughput analyses of protein interactions in yeast. ***Molecular & Cellular Proteomics 8:3–18, 2009.***

Proteins are undeniably some of the most fascinating and complex macromolecules in living systems. Their molecular and cellular functions are determined by an intricate interplay between the laws of physics and chemistry that underlie their structural and dynamic properties and evolutionary forces, which have tailored these properties for the extraordinarily diverse roles that proteins play in sustaining life. Decades of classical cell biology, biochemistry, structural biology, biophysical methods, and mutagenesis techniques have produced a remarkable body of knowledge on the function and molecular properties of individual proteins, but proteins rarely act alone. They interact with one another, often forming large edifices that act as complex molecular machines (1, 2). Although this has long been realized, the prevalence of such interactions and complexes in living cells only became apparent less than 10 years ago thanks to technological developments enabling large scale studies of protein-protein interactions and complexes in the yeast *Saccaromyces cerevisiae* (3–6). These studies and more recent work in yeast (7, 8) and other model organisms, including the bacteria *Escherichia coli* (9), the fly *Drosophila melanogaster* (10), the worm *Caenorhabditis elegans* (11), and human (12–14), produced data describing thousands of protein-protein interactions grouped into hundreds of complexes. This new and exciting system level view is having an enormous impact as it comes against the backdrop of many new developments in biology and other fields. Of particular relevance are the discoveries made by genome sequencing efforts and related research notably on the conservation of key cellular processes across genomes (15, 16) and on the pleiotropic function of proteins (17, 18). Also important is the progress in analytical methods and imaging techniques for quantitative characterization of structural and dynamic properties of proteins and their assemblies (19). Lastly ready access to information through the World Wide Web and the ever increasing power of computers have been essential for enabling the analysis and visualization of large and complex biological data.

Undeniably this new landscape of protein science offers exciting perspectives for generating new knowledge. But navigating it involves many pitfalls and challenges that are not always recognized or are sometimes overlooked in the excitement of the discovery process. This may be of some consequence as system level proteomics has become so multidisciplinary that few if any researchers in the field, let alone outside, can critically evaluate all the aspects.

An important aim of this review is to clearly document some of these challenges. Taking as an example the latest high throughput analyses by tandem affinity purification of the yeast interaction proteome (7, 8, 20, 21), we review the complex procedures of generating the raw experimental data and translating these data into meaningful descriptions of the biological reality. In the process we highlight the biases that may affect the raw data, the crucial role of computational procedures in interpreting the data, the problems encountered, and the efforts made to tackle them. In addition, we examine outstanding issues in assessing the quality and coverage of interaction data sets, and finally, commenting on the current most comprehensive descriptions of the yeast interactome we outline some of the exciting future research directions that such descriptions enable. The possibility of interactively visualizing and comparing several of these descriptions is also available at the Wodak laboratory website.

HIGH THROUGHPUT ANALYSES OF PROTEIN INTERACTIONS IN YEAST: THE DIVERGENT VIEWS

The tandem affinity purification (TAP)[1] method (22) is perhaps one of the most powerful methods for systematically

From the ‡Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada, Departments of §Medical Genetics and Microbiology and ¶Biochemistry, University of Toronto, 1 Kings College Circle, Toronto, Ontario M5S 1A8, Canada, and ‡‡CNRS-UPR-2167, Centre de Génétique Moléculaire, Avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France

Received, August 7, 2008

Published, MCP Papers in Press, September 17, 2008, DOI 10.1074/mcp.R800014-MCP200

[1] The abbreviations used are: TAP, tandem affinity purification; CBC, cap binding complex; BioGRID, General Repository for Interaction Datasets; HC, high confidence; MCL, Markov cluster algo-

identifying protein-protein interactions. It involves expressing a tagged protein at its normal concentration *in vivo* and applying a double purification protocol yielding virtually homogeneous material for all but the least abundant protein complexes. Analysis of the purified material by MS techniques (23) is then able to identify stoichiometric, stably associated subunits and substoichiometric, often weakly interacting proteins.

Using this TAP-MS methodology in a high throughput mode, two groups working independently, one in Heidelberg (7) and the other in Toronto (8), recently produced the most comprehensive set of protein interactions and complexes in the yeast *S. cerevisiae*, involving about 70% of the predicted proteome of this organism. Although there was substantial overlap between the sets of tagged proteins produced by both groups and between their co-purified partners, the final sets of complexes were very different. The total number and size distribution of these complexes was similar, but the composition of individual complexes and other properties differed significantly. A detailed one to one comparison between the complexes derived in the two studies (21) showed that the majority (86%) differed by more than 50% of their components, whereas only 4% were virtually identical. Furthermore the Heidelberg study produced a large number of "complex isoforms," representing modules differing from one another by only a subset of the components. Their final set of 491 complexes displayed only a partial overlap with known complexes stored in databases (see Fig. 3) but very extensive overlap among themselves with 97% of the complexes sharing components with 40 others on average. In stark contrast, the Toronto study (8) reported 547 non-overlapping complexes, which displayed more extensive, although still partial, overlap with known complexes. Judged by the composition of known yeast complexes, which share on average 2.3 proteins with one other complex (21), the complete lack of shared components between different complexes is clearly not realistic. If this is indeed the case, is the high modularity and extensive overlap between the Heidelberg complexes a better reflection of biological reality (7)? To address these issues and gain insight into the possible origins of these surprising differences (24), it is helpful to highlight the biases that may arise in preparing the purified material and characterizing its protein composition using the TAP-MS procedure and to examine in some detail the analytical procedures used to translate the "raw data" on protein composition into information on complexes.

rithm; MIPS, Munich Information Center for Protein Sequences; SGD, *Saccharomyces* Genome Database; 3D, three-dimensional; GSP, gold standard positive; GSN, gold standard negative; FP, false positive; TP, true positive; DIP, Database of Interacting Proteins; GO, Gene Ontology; STRING, Search Tool for the Retrieval of Interacting Genes/Proteins; PSI-MI, proteomics standards initiative-molecular interaction; BIND, Biomolecular Interaction Network Database.

FROM TAGGED YEAST STRAINS TO DATA ON CO-PURIFIED PROTEINS

The TAP-MS procedure involves a series of steps outlined in Fig. 1. The information output by the procedure consists of lists of proteins that co-purify with a specific tagged target protein, but the relationship of these co-purified polypeptides to the protein complexes actually formed in the cell is not straightforward.

Protein complexes differ from aggregates in that they represent assemblies of polypeptides that interact in a specific manner with a defined stoichiometry. They may in addition contain non-protein components, including small molecule cofactors as well as nucleic acids of various sizes. Furthermore protein complexes tend to be dynamic entities: they may assemble or disassemble as conditions change. Larger complexes may form through the assembly of several smaller pre-existing complexes as seen in the formation of the spliceosome during the splicing process (25). In addition, a given protein may participate in several complexes by forming either simultaneous or time- and condition-dependent interactions with several partners. Proteins recovered in a given purification hence rarely correspond to a single complex but often to a heterogeneous mixture of related complexes (Fig. 1*d*) with some particles representing naturally occurring subcomplexes of the largest assembly, whereas others correspond to partial complexes having lost peripheral component(s) during purification. In all cases, the particles amenable to characterization must be abundant enough and sufficiently stable to withstand the purification procedure, and thus by design stable and abundant complexes are more readily identified by the TAP-MS procedure than transient or low abundance ones.

Some typical problems may bias the outcome of TAP-MS analyses performed on individual complexes (low throughput), and those are often amplified when the procedure is applied in a high throughput mode. A major problem is the presence of contaminants. Those are polypeptides that tend to either bind to the column matrices used for affinity purification or form nonspecific interactions with many different proteins. They therefore do not represent biologically significant partners and need to be filtered out of the raw data (Fig. 1*g*). Typical contaminants are highly abundant proteins (*e.g.* ribosomal proteins, subunits of the cytoskeleton, and chaperones) or polypeptides particularly prone to such interactions (often called "sticky" proteins). Even under optimal conditions, it is usually very difficult to completely remove these during the purification or fractionation steps (see Fig. 1 for details).

A special, although not uncommon, instance where contaminants may actually appear to display exquisite specificity is in the purification of complexes that bind nucleic acids (DNA or RNA). The nuclear cap binding complex (CBC) can be taken as an example. This complex, shown to consist of a tight heterodimer, is often recovered associated with other RNA-binding protein in high throughput affinity capture ex-

periments. Of the 60 partners of the largest subunit of CBC identified by various affinity capture studies and listed in the BioGRID database (26), 49 are other proteins directly involved in RNA metabolism. This includes 29 splicing factors to which the CBC is known to associate in a pre-mRNA-dependent manner during spliceosome assembly (27–29), whereas other abundant RNA binding factors that may also interact through RNA bridges represent the majority of the remaining partners. Such carryover contamination can be avoided by using modified protocols (such as using nuclease treatment), which are known to specialists in the field (30–32) but have not yet been implemented in high throughput studies. There is the danger, however, of eliminating biologically relevant interactions as it is often difficult to discriminate between those and nonspecific binders. Other well known problems that may mar accurate complex identification include the failure to capture one or more subunits in the purified material because of their low abundance or the failure to identify a captured protein by mass spectrometry because of its small size or to poor annotation in the databases (Fig. 1, *f* and *g*).

Several of the above mentioned issues can be addressed in studies of individual complexes by performing systematic analyses of the fractionation results (Fig. 1e). For instance, processing information on the observed molecular mass of the protein bands on the gel or their apparent relative abundance, although not very accurate, can be instrumental in helping characterize subcomplexes. Such analysis can also suggest the presence of additional subunits even when no peptides have been identified, indicate possible proteolytic maturation or degradation steps, and help estimate stoichiometry relationships. Presently this information is not exploited in high throughput studies, but in principle, this could be done provided appropriate analytical methods are developed. Meanwhile to improve detection sensitivity, decrease noise due to contaminants, and afford detection of subcomplexes, the recent high throughput studies (7, 8) have systematically tagged as many proteins as possible, often representing several subunits from the same complex. In addition repeat purifications were carried out, in one instance (8) with material from over a quarter of the tagged strains purified between 3 and 14 times, and its composition was characterized with two different mass spectrometry techniques (Fig. 1). An obvious consequence of this strategy is the necessity of relying on automated analytical methods to recover biologically meaningful information.

FROM LISTS OF CO-PURIFIED PROTEINS TO COMPLEXES

The sets of protein components identified in thousands of purification runs are hence not the final product of a high throughput TAP-MS study but rather an important raw material that must be processed further to derive information on protein complexes that form in the cell. This data processing task is a key operation that involves two main steps, as detailed in Fig. 2. In the first step, the heterogeneous lists of components are converted into a network of binary protein-protein links with each link quantified by a score reflecting the confidence with which the link has been detected in the experiments. Only a fraction of these links represent direct physical interactions (Fig. 2). Next this network of weighted binary links is usually filtered to exclude low scoring links, and the resulting high confidence (HC) network is partitioned into densely connected regions with the help of computational clustering procedures (33), yielding the final complexes, each described by a list of components. The protocol outlined above involved complex computational procedures for which there was no precedent in this context, and the specific implementations of these procedures differed significantly between the two recent high throughput studies (see Fig. 2 for further details).

DIFFERENT ANALYTICAL METHODS CAN YIELD PROFOUNDLY DIFFERENT DESCRIPTIONS OF COMPLEXES

Reanalysis of the raw data made publicly available by both groups upon publication produced compelling evidence that the differences in the analytical methods used by the two groups have a profound effect on the results (20, 21). It was shown that the raw data sets from both studies were of comparable quality. The procedures for computing the network of binary links differed with measurable but limited consequences on the results. The Heidelberg study used an elegant statistical method, which has certain advantages over the particular machine learning procedure used by the Toronto consortium (see Fig. 2*c* and corresponding legend). Improved versions of this method were used in subsequent studies to derive more accurate networks using the raw data from both published studies (20, 34). However, by far the most significant difference between the Toronto and Heidelberg descriptions was attributed to the clustering methods used to partition the network (21). Clustering is an optimization problem that is usually solved with the help of heuristic algorithms whose ability to approximate the best solution (global minimum) may vary widely, and there is a wealth of clustering algorithms to choose from as discussed in a recent comparative study (33). The Toronto group used the Markov cluster algorithm (MCL) (35, 36), which is based on the simulation of stochastic flow in graphs and therefore uses information on the graph structure. This algorithm produces disjoint groups and displays good convergence and robustness (33) but can fail to separate complexes with highly interlinked shared subunits (for example RNA polymerases I, II, and III) (8, 21). The Heidelberg team applied an iterative hierarchical clustering procedure, which used the pairwise interaction score as the clustering metric and was fine tuned to yield different solutions in each run. These different solutions (complex isoforms) were then combined to produce the final set of highly overlapping protein assemblies (7). As already mentioned, these assemblies showed little overlap with the Toronto complexes.

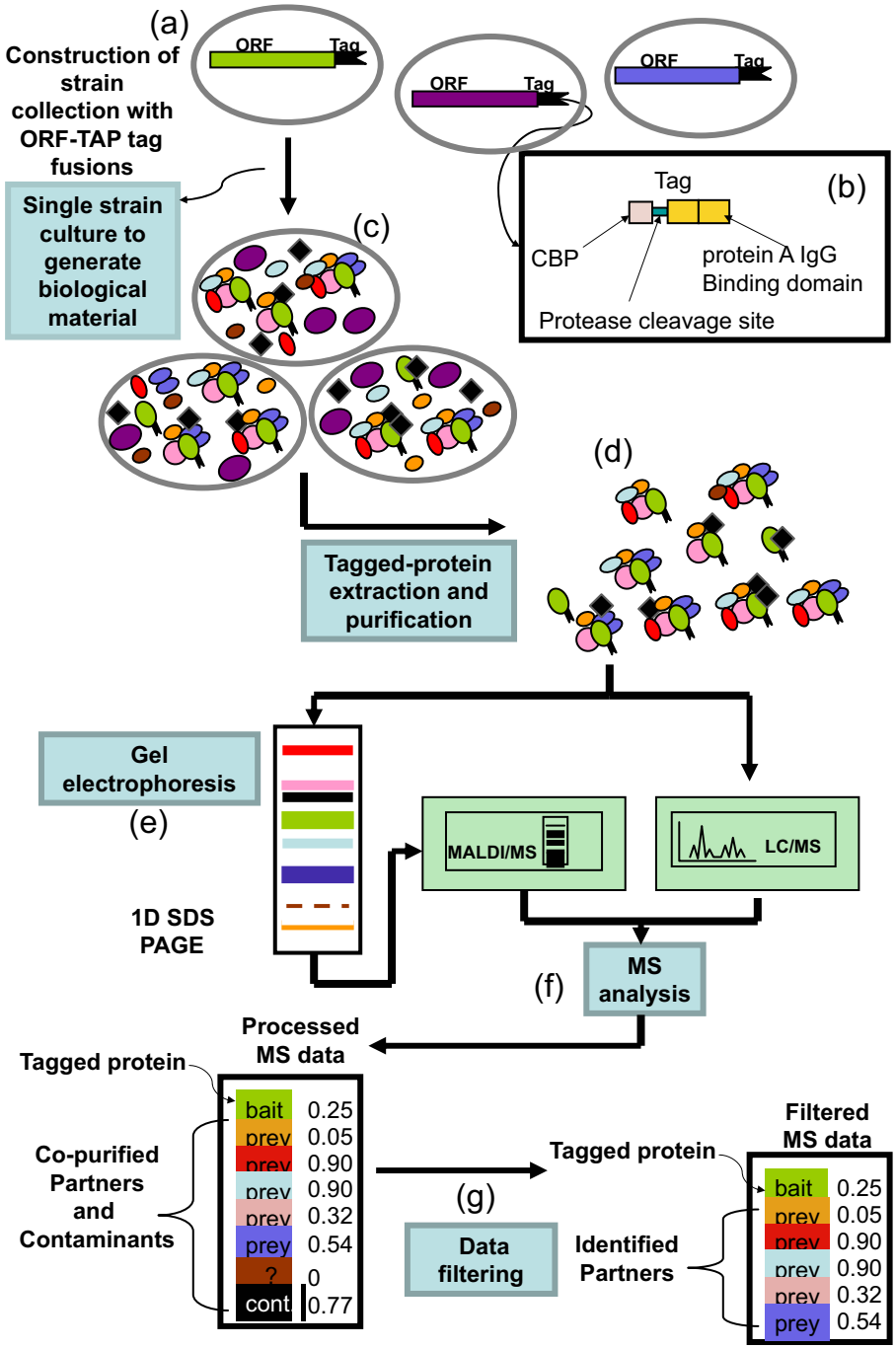


FIG. 1. **Work flow for the TAP-MS procedure for characterizing protein complexes.** Affinity purification of a tagged protein starts with the construction of cells or organisms expressing the fusion construct (ORF plus tag) schematized in *a*. The tag used in the TAP methods (22) comprises a calmodulin binding peptide (*CBP*), a protease cleavage site, and a protein A-IgG binding domain as schematized in *b*. In the TAP procedure implemented in *S. cerevisiae*, the tagged protein is produced at normal levels under the regulation of its endogenous promoter. But this may differ in other targeted biological systems (4) where regulation of the expression of the fusion protein or its location may only partly mimic the natural situation. As a result, the composition of recovered complexes might in some cases differ partly from the natural assemblies formed by untagged factors. Following expression of the tagged protein (*c*), protein material is extracted and purified (*d*). This material is schematized as different versions of the complex involving the tagged protein (*green*), each represented by its component subunits; other isolated cellular proteins (*violet*) and contaminants (*black diamonds*) are also shown. Such contaminants are often difficult to remove completely after purification as they may not interact equally well with different tagged proteins. For example, the fusion of the tag to a particular target may alter its fold, exposing hydrophobic surfaces prone to associate with chaperones and/or other abundant proteins (78). Another set of contaminants may interact with tagged proteins (or their associated partners) that bear a high electrostatic charge. Binding of contaminants may also be affected by small changes in the purification conditions. The abundance of the target complex in the extract will be determined

Interestingly, however, application of MCL followed by a postprocessing step to the published Heidelberg and Toronto networks, respectively (21) (Fig. 2), yielded in each case complexes sharing on average only 1.2 components. Moreover complexes from the two data sets now displayed a significantly better overlap than the published complexes. This overlap was further improved when exactly the same computational procedures were used to reprocess the two raw data sets, compute the networks, and derive the complexes from them (21). The resulting new sets of complexes also displayed an improved and similar level of consistency with available biological knowledge, including information on yeast complexes stored in databases, indicating that they are of comparable accuracy as will be discussed below.

These findings demonstrate that the very extensive overlap and overly modular nature of the protein complexes described in the Heidelberg study are not rooted in the properties of the "interaction" network or in potential biases introduced in building the network (*e.g.* learning from known complexes in the case of the Toronto network) but in the procedure used to partition it. This makes very good sense if one examines the high throughput TAP-MS protocol used in both studies. Indeed in this protocol, information on the many different versions of the co-purified components is combined into a single copy of each binary link and its associated confidence score (Fig. 2). The data derived in this fashion therefore represent weighted averages of the detected versions, which in turn represent temporal and spatial averages of protein associations that are stable enough and/or sufficiently abundant throughout the cell cultures of *S. cerevisiae* under the conditions in which the purifications were performed. Hence there is no way for any temporal or spatial modularity that may be occurring *in vivo* to be captured by the described high throughput TAP-MS protocols without the help of additional information, like for instance on cell state- or condition-dependent mRNA expression levels (37–39). This is a recognized limitation of these protocols as applied to date that none of the available clustering procedure, including the two discussed here, can overcome.

EVALUATING THE ERROR RATES OF INTERACTION NETWORKS AND COMPLEXES

Successfully meeting the challenges of translating the TAP-MS data into useful descriptions of the interactome also requires a means of evaluating to what extent the derived descriptions reflect biological reality by determining their error rate and coverage. Agreed upon measures for estimating the error rate of the identified binary links and complexes are still lacking, and the development of such measures is an active area of research. Complementary laboratory experiments can be carried out but only on selected subsets of the data. Preference is therefore given to quantitative criteria that measure consistency with prior knowledge. The main validation practices are summarized in Fig. 3.

Comparison against a set of highly reliable, or "gold standard," literature-curated interactions stored in databases (Fig. 4 and Table I) plays a particularly important role. Using such gold standards, error rates for the interaction data, often defined as the fraction of binary links deemed to be spurious (false positives), can be estimated. But these estimates may vary widely depending on how the gold standard itself is defined (Fig. 3). Comparing error rates for different data sets is hence meaningless unless these rates are estimated using the same gold standard and calculation method, and in the best case, only the relative magnitudes may be informative as illustrated in Fig. 3. Some of the high error rates recently estimated for the yeast interaction networks derived from the latest TAP-MS studies (40) should therefore be critically reviewed in this light. There remains

both by its expression and recovery levels. Some proteins/complexes may be lost when associated with insoluble structures (*e.g.* membranes, cell walls, chromosomes, large complexes such as the ribosomes, or the cytoskeleton), whereas complex integrity may be affected by simple changes in extraction conditions (ionic strength, pH, or the presence of divalent cations such as Mg^{2+} or Ca^{2+}). Small polypeptides may also escape detection during the fractionation step as reported for the 10th subunits of transcription factor TFIIH (79, 80). The use of heterogeneous cell populations (*e.g.* cells at different stages of their division cycle) may affect the detection of complexes present in trace amounts. The mixing of various cellular compartments that inevitably occurs during extract preparation fosters the formation of nonspecific interactions if gentle conditions are not maintained. Complexes are selectively recovered from the extracts through purification (two successive steps in the TAP procedure) where buffer and temperature conditions may affect the outcome of the experiment. Extensive washing and harsh buffer conditions favor elimination of contaminants with the risk of simultaneously losing some *bona fide* complex subunit(s). To average out the effects of such variations, the recent two high throughput studies performed multiple purifications for subsets of the cell extract preparations (7, 8). Proteins recovered following affinity purification are then identified by MS (*f*). The latter is carried out after fractionation by gel electrophoresis (*e*), using MALDI-TOF/MS, or directly on the eluate using LC-MS/MS. The gel fractionation step may decrease the sensitivity of protein detection but offers the possibility of obtaining rough estimates of the relative protein levels present in the preparation. In one of the recent analyses (8) both techniques were applied. For each sample, the MS analysis outputs a list of proteins, the tagged target (bait) and its co-purified partners (preys), each accompanied by a score representing the reliability with which the protein has been identified in the databases. Several biases may influence these analyses. Both types of analyses involve breaking the proteins in the mixture into peptides. Naturally smaller proteins are broken up into fewer peptides than larger proteins and can therefore more readily escape detection. Differences in peptide ionization may also influence detection with the detection of hydrophobic proteins being disfavored. Finally the failure to map identified peptides onto ORFs annotated in databases is not uncommon especially for small ORFs that are more often poorly annotated (81). Identical proteins present in many purifications are often classified as contaminants and automatically filtered out (*g*).

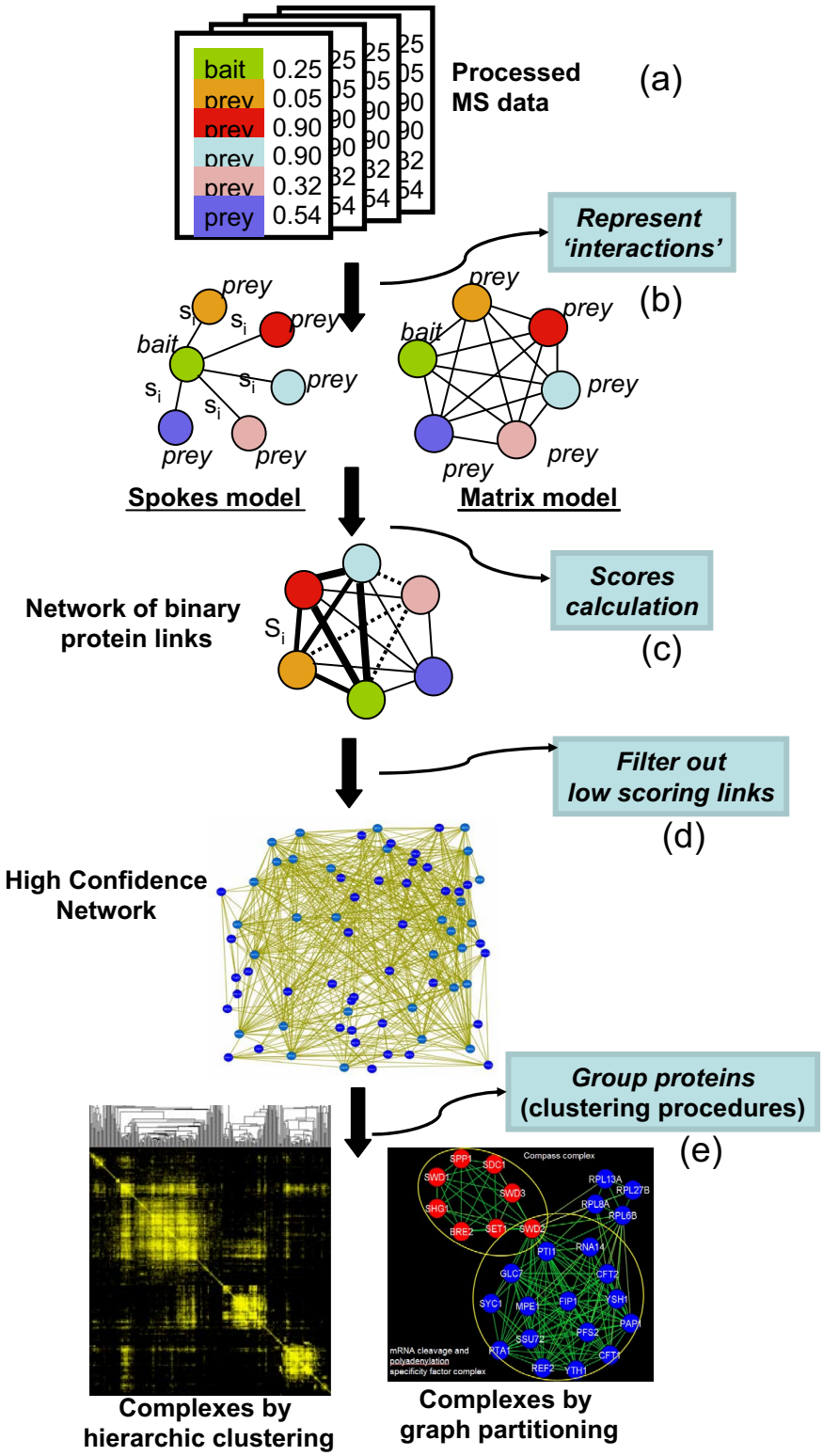


FIG. 2. **Protocol for deriving protein complexes from data produced by TAP followed by MS.** *a*, the data output by the MS analysis for each analyzed sample comprises a list of proteins, each accompanied by a score representing the reliability with which the protein has been identified in the databases. The list usually, but not always, includes the tagged protein (bait) and its co-purified partners that have been identified (preys). Each analyzed sample often represents material from repeated purifications of extracts from cells expressing the same tagged construct (Fig. 1). Only the listed proteins are used as the raw data for the subsequent computational analysis. The accompanying MS scores have so far not been exploited. *b*, next a model for representing the "interactions" is selected. In the *spokes model* only bait-prey links

the problem that the data sets most commonly used as the gold standard for yeast (see Fig. 3) tend to be outdated. It is furthermore conceivable that there is no such thing as a unique organism-specific gold standard. Indeed a recent systematic validation (41) of binary protein interactions in *S. cerevisiae* using yeast two-hybrid screens, protein complementation assays (42), and the mammalian protein-protein interaction trap technique (43) provides compelling evidence that interactions of this type tend to differ from those identified using methods for detecting complexes. Better (and usually lower) estimates of the error rate of a given type of method can therefore be obtained by comparison with gold standards that are appropriate for the same type of methods (41). Clearly therefore defining the gold standard data set remains a thorny issue.

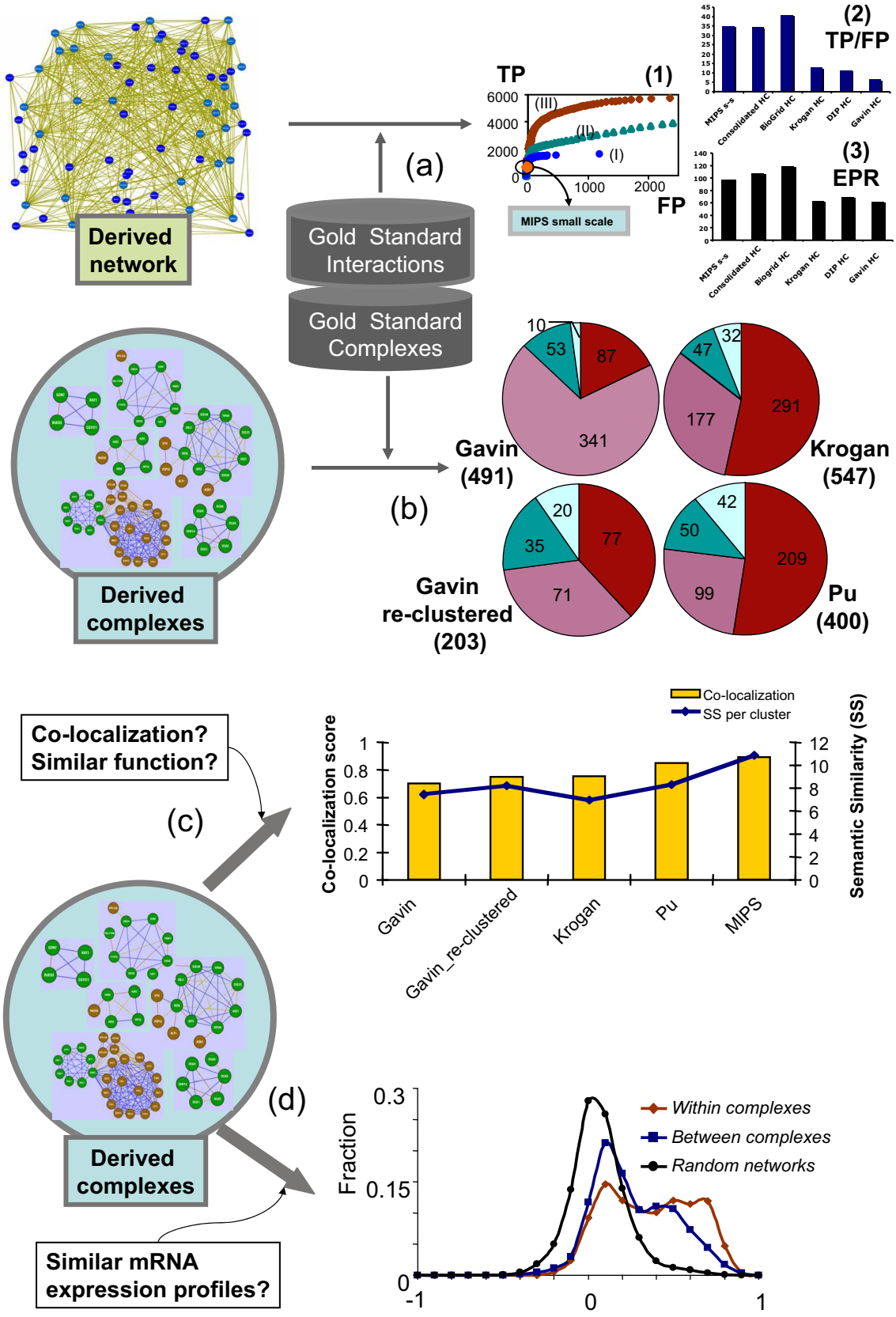
Another important use of gold standard data sets is in building the interaction networks in the first place. Portions of these data sets can be used to optimize the manner in which raw mass spectrometry data from different purifications of the same complex and different complexes are combined or more generally to consolidate different lines of evidence supporting a given binary interaction into a single confidence score (8, 20, 44). Settling on an acceptable error rate then allows defining the confidence score threshold above which interactions should be considered for further analysis.

Other reliability measures involve evaluating consistency with available information on cellular localization, the functional annotations of individual protein components stored in databases, and often also the expression profiles of the corresponding genes (45) (Fig. 3). Reliability criteria based on inferred domain-domain interactions (46) or on annotated interactions between paralogous proteins (45) have also been proposed. But all these criteria involve assumptions and approximations, and their usefulness is further limited by the quality and particular biases in the available information (*e.g.* subsets of proteins and interactions covered). Availability of high quality data sets of these different properties is thus not only important for understanding biology but is also instrumental in the discovery process.

DESCRIPTIONS OF MODEL ORGANISM INTERACTOMES ARE BECOMING INCREASINGLY ACCURATE

Despite many unresolved issue, real progress is being achieved by high throughput methods in providing accurate descriptions of the yeast interactome. Applying a new scoring scheme to the raw data sets from the two latest Heidelberg and Toronto studies, a consolidated protein interaction network was recently derived for *S. cerevisiae* (20). The estimated average error rate of the HC portion of this network, comprising 1622 proteins and 9074 interactions, is lower than for several recently derived HC networks and similar to that of the data set of binary interactions identified by small scale experiments and annotated by the Munich Information Center for Protein Sequences (MIPS) (see Fig. 3 for details). Using this HC consolidated network, a set of 400 high confidence multiprotein complexes was derived by applying the MCL procedure followed by a postprocessing step that assigned components to multiple complexes (21).

are considered (8). In the *matrix model*, both bait-prey and prey-prey links are included (7, 20). The latter model yields higher connectivity between proteins that are repeatedly co-purified and may facilitate further processing into complexes (see below). However, both models partly reflect reality because they include links representing both direct physical interactions and indirect interactions formed through third partners. Information on the number of copies of each protein in the complex (stoichiometry) is not provided. *c*, the lists of proteins output by the MS methods from different purifications using the same or different bait proteins or from different MS analyses are combined into one global network of binary links using either the spokes or the matrix model. Each link is accompanied by a score reflecting the confidence with which it has been identified in the TAP-MS analysis taken as a whole. In the Heidelberg study these so-called "confidence scores" (S_i) were derived using an unbiased statistical method (7). This method does not rely on information on known complexes to compute the scores and to some extent factors out directly links involving high abundance sticky proteins. The Toronto study used machine learning procedures (8), and a subsequent reprocessing of the data sets from both studies combined both approaches (20). With machine learning procedures, the scores are derived by "learning" from examples with the latter being taken from a set of gold standard positive (GSP) and gold standard negative (GSN) interactions (Fig. 3) with pruning of sticky proteins usually performed by the researcher. With adequate cross-validation, possible biases in the scoring scheme toward the properties of the gold standard data set are in principle eliminated. *d*, the HC portion of the network is derived by filtering out binary links whose score is below a given threshold defined by measures of the expected rate of spurious (false positive) links (20). *e*, the HC network is partitioned into densely connected regions with the help of clustering procedures (33) to yield the final complexes, each described by a list of components. To enable detailed analysis, the resulting complexes can be mapped back into the HC network from which they were derived. Results obtained by applying a hierarchical clustering procedure to the HC yeast consolidated network (20) are illustrated in the *left panel* using the TreeView software. The panel displays a zoomed-in portion of the connectivity graph representing the network after its components have been clustered as shown an the *top* of the panel. The color of individual pixels reflects the level of certainty (high, *bright yellow*; low, *black*) that the corresponding protein pair belongs to the same complex. Defining complexes usually involves selecting a threshold for this level on an *ad hoc* basis. There also is the flexibility of selecting somewhat different thresholds in specific regions of the graph. Results obtained by applying the MCL procedure (35, 36) to the same network are displayed on the *right panel*. This procedure has two main adjustable parameters, which affect the granularity of the resulting clusters (36). These parameters can be tweaked to maximize the overlap between the computed clusters and known complexes (8, 21). In the example shown *red nodes* are the proteins assigned to the complex under scrutiny. The *blue nodes* are proteins outside the complex (but in the original HC network) that form at least one link with a protein component in the complex. The picture was generated using specialized software for visualizing interactions between protein components within and between complexes (82, 83). Using this software the various networks described in this review can be visualized interactively at the Wodak laboratory website.



The resulting complexes share on average 2.5 protein components with every other complex, closely approaching the overlap observed in the curated complexes, and about 50% of these are "new" complexes not catalogued in the MIPS database. Supporting evidence for the composition of the majority of these new complexes could be obtained from the recent scientific literature and annotations in the *S. cerevisiae* database SGD (Fig. 4), suggesting that interactome descriptions derived from high throughput studies are becoming quite meaningful. Similar claims were recently made about the "second generation" consolidated data set of binary protein interactions in *S. cerevisiae*, detected by yeast two-hybrid screens, that comprises 2930 binary interactions among 2018 proteins (41). A majority of the interactions in this set were detected by three independent trials, and a subset was verified by two orthogonal experimental assays.

Clearly, however, these descriptions provide a highly abstract and incomplete representation of the biological reality. The TAP-MS studies provide only a component list for each complex but no information on stoichiometry or on the direct physical interactions formed. The yeast two-hybrid and other binary screens (42, 43) detect only binary interaction partners, which form as a result of direct physical interactions or by associations through other protein components, and neither of the above mentioned methods currently provides information on the chemically modified states (post-translational modifications) of the protein components.

FIG. 3. **Validating interactions and complexes against prior knowledge.** *a*, the derived network is validated against the GSP and GSN sets of interactions (denoted as gold standard in the text). In studies on yeast the GSP is often defined as the set of all possible binary connections between protein components annotated as belonging to the same hand-curated complex stored in the databases (84, 85). An alternative GSP can be derived from a reliable set of literature-curated binary interactions. One such set is available for *S. cerevisiae* in the MIPS database (85) (denoted as "*MIPS small scale*" or "*MIPS s-s*" on *plots 1–3*). The GSN represents the set of interactions that are unlikely to occur. But in the absence of any direct evidence that an interaction does not form in the cell, defining such a set remains speculative. Some define the GSN set as interactions between proteins belonging to distinct complexes and having non-overlapping cellular localizations (8, 21, 51). Others consider interactions between two proteins selected at random (86) or even any interactions not in the GSP (44, 87). Using these references sets, the expected error rate of a network is often computed as the ratio FP/(TP + FP) where FP (false positive) is the number of detected links that map into the GSN set, and TP (true positive) is the number of correctly recalled GSP interactions. A related quality measure is the TP/FP ratio. The GSN definition and the relative sizes of the GSP and GSN data sets may crucially influence both ratios. *Panel 1* displays the TP/FP ratios evaluated for networks built with different confidence score thresholds and plotted using a version of the receiver operating characteristic curve. Analysis of these curves is used to define the appropriate threshold value for the confidence score (20). The displayed curves are for the three recently published HC yeast networks (I, Ref. 8; II, Ref. 7; and III, Ref. 20) with the MIPS small scale data set highlighted (*orange diamond*). The HC yeast network from Ref. 7 is the one deposited in the BioGRID database. Quality measures for six yeast interaction networks, estimated using two different gold standard data sets and calculation methods, are displayed in *panels 2* and *3*. The TP/FP ratios computed using the same GSN and GSP sets as in Ref. 21 are displayed in *panel 2*. *Panel 3* displays the expression profile reliability (*EPR*) index (in %), which compares the mRNA expression profiles of interacting proteins in the data set with those in a GSP-GSN set (45). The evaluated networks are those of Krogan *et al.* (8) HC, Gavin *et al.* (7) HC, the consolidated HC (20) and three literature-curated binary interactions data sets (MIPS small scale, DIP HC ("core") (88), and BioGRID HC (59)). Comparison of *panels 2* and *3* clearly illustrates that the two evaluation methods essentially conserve the relative ranking of the different data set with the latest BioGRID HC data set approaching the reliabilities of the MIPS small scale data set and consolidated network of Collins *et al.* (20) despite very different absolute values of the quality measures. *b*, validation of newly derived complexes involves one to one comparisons between the components of these complexes with those in the gold standard set of known complexes and measuring the extent of overlap between the components (8, 21, 24). Shown here is the overlap of the components in complexes from the MIPS catalogue with complexes (clusters) derived from different high throughput analyses. These are the 491 complexes of Gavin *et al.* (7), the 547 complexes of Krogan *et al.* (8), the 203 complexes derived by reclustering the Gavin *et al.* (7) data set (21) using the MCL procedure (21) (see Fig. 2), and the 400 complexes derived by Pu *et al.* (21). Each set of derived complexes was subdivided into four categories according to the extent of overlap (*white*, >90%; *turquoise*, 50–90%; *magenta*, 5–50%; and *red*, <5%) between the components of the complex and the maximum matching MIPS complexes, and the results are displayed as a pie chart. For a more complete quantitative analysis of the overlap between the complexes in these different data sets and those derived from other reprocessed versions of the raw data from the Gavin *et al.* (7) and Krogan *et al.* (8) studies, see Ref. 21. The four sets of complexes and the corresponding networks can be interactively visualized at the Wodak laboratory website. *c*, evaluating the similarity in functional annotations between proteins within complexes. These annotations, commonly represented by the Gene Ontology (GO) terms (89), are often defined at different levels of the GO hierarchy, making comparisons cumbersome. One measure of similarity between GO annotations (90) is plotted here (*right-hand panel*) for the same four sets of derived complexes as those described in *b* as well as for the MIPS complexes used as references. The same panel also plots the extent to which components within complexes have been assigned to the same subcellular compartment. Co-localization of the interacting partners is an indication that they are likely to interact *in vivo.* But reliable data on subcellular localization are available for only a few model organisms such as *S. cerevisiae* (91, 92), and even in those, a large number of proteins tend to be found in multiple locations (18.8% in yeast), making localization data much less informative for such proteins. *d*, significant correlation between the mRNA expression profiles of two proteins can be an indication that they interact *in vivo.* Shown are the distributions (*ordinate*) of the pairwise Pearson correlation coefficient (*abscissa*) of the mRNA expression profiles of *S. cerevisiae* proteins of different categories. However, co-expressed proteins may be part of the same cellular process (pathways) and not necessarily interact physically. It must also be realized that mRNA expression levels as well as cellular localization data may depend on the cellular state and experimental conditions, which might differ from those used to characterize the interactions or complexes that need to be validated. For instance, expression data most commonly used for validation in yeast are from publicly available sets generated to investigate response to stress conditions (93), whereas the TAP studies were done on exponentially growing yeast cultured on rich medium.

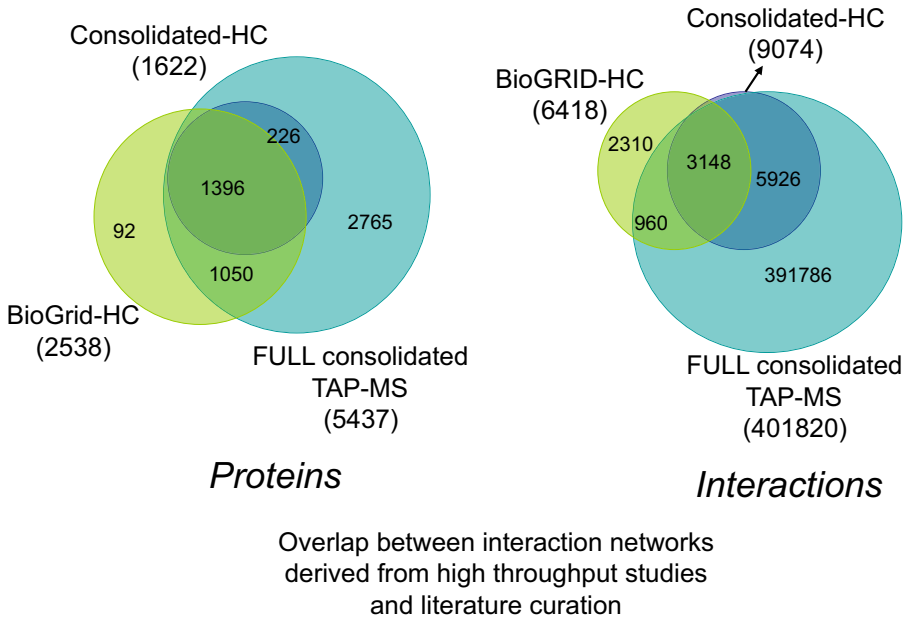


FIG. 4. **Overlap between interaction networks derived from high throughput studies and literature curation.** The Venn diagrams illustrate the overlap of the proteins (*left-hand side*) and interactions (*right-hand side*) in the HC portion of the latest literature-curated interaction data sets for yeast from the BioGRID database (26) with the HC portion of the consolidated network of Collins *et al.* (20) derived from the TAP-MS data of Refs. 8 and 7 and with the full (unthresholded) consolidated TAP-MS network of Ref. 20. The *numbers* in *parentheses* represent the protein and interaction count, respectively, for each subset category.

COVERAGE REMAINS AN ISSUE

An important goal of high throughput methods is to provide a systems level view of the interactome that is as unbiased and comprehensive as possible. This might be an elusive goal as far as interactions and complexes are concerned given their modular and dynamic nature, which is difficult to capture with current methods. Achieving good coverage of even the more stable and abundant complexes in yeast, the champion of model organisms, still remains a challenge. This is the case for the two-hybrid and related assays as well as for purification methods in both low and high throughput modes. The coverage of membrane proteins has so far been very limited (7, 8). Furthermore our current knowledge indicates that changes in cellular conditions may significantly influence the composition of some complexes. However, few if any of the purification analyses, those at high throughput in particular, have so far explored the variety of conditions encountered by living cells.

It is therefore very difficult to estimate the fraction of the protein interactions known today without at least a good guess of what the size of the complete interaction set might be. How we define interactions (binary interactions or co-complex links) will also influence the projections. Recent estimates, which suggest that 50% of all possible interactions in yeast have been identified (by all types of experimental methods combined) (40), are overly optimistic. Among other things, these estimates ignore the fact that the set of known interactions is not a random sample of the entire network (47, 48). More conservative estimates based on a careful statistical analysis, which considers the various factors affecting the process of sampling interactions, suggest that only about 15–20% of "all" binary interactions in yeast have so far been mapped (49). Similar figures were recently reported on the basis, among other things, of the observed rates at which known interactions remain undetected by yeast two-hybrid methods (41). These studies estimate the total number of binary interactions in *S. cerevisiae* to be between 18,000 and 30,000. In comparison the number of binary interactions in human was recently estimated at ~600,000 with a current coverage of less than 1% (50).

To further apprehend the issue it is useful to highlight the fact that the two recent TAP-MS studies mentioned above actually detected several hundred thousand binary links involving several thousand proteins. However, only about 2% of these links and less than 30% of the proteins end up in the HC networks from which the published interactome models were derived. The confidence scores of the remaining vast majority of the links are too low to warrant their consideration. Indiscriminately including them would result in a network containing a prohibitive proportion of false positive links possibly due to contaminants. Yet it is likely that a fraction of these low scoring links represent weaker physical interactions and/or those involving lower abundance proteins as discussed above. Failing to consider all the low scoring links thus might deprive us of important information, but distinguishing the true links (true positives) from the spurious ones is difficult. There are ways of computationally tackling the problem. They involve procedures for integrating different lines of evidence supporting a given protein-protein link (*e.g.* literature citation, similarity of mRNA expression profiles, and same subcellular localization) with the experimentally derived pairwise interaction score to yield a new consolidated confidence score. But unlike in other data integration work (34, 51, 52), scores must be computed only for the links observed in the raw TAP-MS data set. Preliminary results obtained by some of

TABLE I
Interaction databases

The availability of high quality curated information on complexes and interactions characterized in different organisms is not only important for understanding biology but also for aiding the discovery process. Several national and international efforts are devoted to producing this information as well as to standards that facilitate its exchange between different databases. The contents of the major databases are summarized are in the table. To keep up with the flood of publications dealing with the subject, database curators seek help from automatic text mining algorithms, which are rapidly gaining in accuracy (94). Nevertheless the quality of literature-curated data can be an issue as low throughput studies, sometimes based on a single experiment, can be just as, or more, error-prone than the more advanced high throughput techniques. In general, databases do not produce confidence scores for the interactions they curate, and extraction of high quality interactions from the databases remains primarily the user's responsibility. Most of the listed databases store experimentally derived protein-protein interactions obtained through literature curation. The only exception so far is the STRING database (95), which stores three types of interactions: 1) experimentally derived protein-protein interactions imported from the other databases and derived from text mining of PubMed abstracts, 2) interactions computed from genomic features, and 3) interactions transferred from model organisms based on orthology. All the listed databases support proteomics standards initiative-molecular interaction (PSI-MI) standards (see below). IntAct has the best conformity with the PSI-MI standards (96). Results deposited by high throughput TAP and yeast two-hybrid techniques include lists of the identified interactions as well as information on roles of each interactor (bait or prey). BIND (97) and DIP (88) allow retrieval of TAP-MS complexes that contain a query protein. The protein complexes in STRING, like those in SGD (98), are catalogued according to the GO (89) annotations and thus do not necessarily correspond to physical complexes. BIND and BioGRID (26) also store genetic interactions (99) (not considered in the table). Raw data (TAP purifications and peptide identification confidence scores) from high throughput studies are not available for search or download in the databases. Model organism databases such as SGD (98), Mouse Genome Database (100), WormBase (101), and FlyBase (102) usually do not independently archive protein-protein interactions. They either collaborate with major interaction databases by coordinating curation efforts (*e.g.* between SGD and BioGRID) or provide links to them (*e.g.* FlyBase and BioGRID). In addition to these major interaction databases and model organism databases, Human Protein Reference Database (103) archives 38,176 curated interactions in human, and MPACT (104) has 15,456 yeast interactions and hosts a catalogue of yeast protein complexes. BioGRID contains 38,609, 499, 22,524, 4,557, and 38,605 interactions in human, mouse, *D. melanogaster*, *C. elegans*, and *S. cerevisiae*, respectively. These figures were compiled in February 2008. PSI (105) is a community wide standard for data representation in proteomics to facilitate data comparison, exchange, and verification. PSI-MI specifies the format for exchange of molecular interactions using a controlled vocabulary. The MIMIx (minimum information required for reporting a molecular interaction experiment) (106) is a subset of the PSI-MI standard. It stipulates that a deposition must include key information that enables unambiguously defining the origin of the data, the method used to generate them, and the means to uniquely reference to other biological databases the partners of each deposited protein-protein link.

Database	Interactions	Organisms	Proteins	Complexes	Confidence
BIND[a]	82,490	1,217[b]	25,387	Yes	No
BioGRID[a]	90,577	13	322,372	No	No
IntAct	155,333	189	58,229	Pending	No
STRING	2,102,940	373	416,058	Yes	Yes
MINT[c]	103,808	30	28,186	No	Yes
DIP	56,186	161	19,490	Yes	No
MPPI[d]	>1,728	10	>900	No	No

[a] Excluding genetic interactions.
[b] Includes 232 viruses and 56 phages of various organisms.
[c] Molecular Interactions database.
[d] MIPS mammalian protein-protein interaction database.

us[2], however, show that applying this approach to the complete raw TAP-MS data sets of *S. cerevisiae* mentioned above and deriving a new HC network results in only a relatively modest gain in coverage relative to the HC interaction network of Collins *et al.* (20).

This is due in part to the nature of the overlap between the binary links in the TAP-MS network and the literature citations compiled in databases such as BioGRID (Fig. 4), which is currently considered as one of the most comprehensive literature-curated interaction compendiums (53). Indeed about 49% of the interactions in the HC BioGRID data set (59) map into TAP-MS links that already have a high confidence score. Another 36% (2310 interactions) have no match to any TAP-MS links (Fig. 4) with some of these, but not all, involving proteins different from those detected in the high throughput purification studies. This leaves only 15% of the HC BioGRID interactions (960 in all) as corresponding to poorly scoring links in the full unthresholded TAP-MS network that may need "salvaging."

Other lines of supporting evidence, such as gene expression data, offer a similar picture: protein interactions for which abundant supporting information is available are either those that are reliably identified by the TAP-MS procedures (54) or are interactions that remain undetected (and possibly not sampled) by these procedures. The high level of noise in the full TAP-MS experimental data clearly contributes to the picture as well.

Information from various two-hybrid screens has so far not been generally helpful as supporting evidence mainly because of the low coverage of the available data (5). But

[2] J. Vlasblom, S. Pu, and S. J. Wodak, unpublished data.

this is likely to change with further efforts underway to carry out more comprehensive yeast two-hybrid screens.[3] The recent achievements of large scale protein-fragment complementation assays, detecting 2770 binary links among 1124 proteins in *S. cerevisiae* (55), are also very promising in this regard. The type and affinity range of the interactions detected by these binary assays seem to be sufficiently complementary to those identified by TAP-MS experiments to provide useful supporting evidence for many of the links classified today as unreliable by the TAP-MS analyses. Lastly significant improvements in data accuracy and coverage may be expected as purification and identification techniques evolve to a point where it becomes feasible to detect transient interactions (56) or to enrich for subsets of the interactions and complexes by systematically sampling the interactome as a function of time, subcellular compartments, and cellular states.

CONCLUDING REMARKS

We focused here on the most comprehensive analyses of the interaction proteome carried out to date in a model organism. These analyses have reached the stage of becoming a powerful hypothesis-generating engine provided the biases in the data and the approximations made in deriving the final interactome descriptions are taken into account. By revealing new complexes and new memberships of previously characterized complexes and further defining the context of a given protein in the interaction network, the system level view afforded by these descriptions can yield useful insights into molecular and cellular function.

There has been great interest in unraveling the biological implications of the local and global properties of the interaction networks derived from these studies or built from combining literature-curated information obtained by various methods (38, 57–59). There is a concern, however, that such efforts may be premature (48) given the highly abstract nature of these networks, the fact that they are clearly biased toward stable interactions, and most importantly that their coverage is still limited.

Meanwhile a potentially very rewarding endeavor would be to translate the abstract description of protein links into more detailed models describing real physical interactions at near atomic or atomic scale. This will be very useful even if performed on the HC networks and complexes known today or on particular subsets of these. Such models can help reveal the physical contacts that can actually be made between proteins and suggest how these interactions might have evolved (60). They can furthermore provide atomic descriptions of the interacting interfaces that are useful for various mechanistic investigations as well as for drug design (61). Although the atomic resolution structures of large multiprotein complexes are still determined at a very slow pace, the repertoire of 3D structures of individual proteins is rapidly being filled (62) thanks in part to structural genomics initiatives (63). This repertoire can be used to model the structures of complexes with the help of computational procedures provided the 3D structure of the components is either known or can be derived from the known structure of a suitable homolog (64). Combining this approach with molecular envelope descriptions obtained from electron microscopy has enabled building partial models for dozens of yeast complexes identified by high throughput proteomics (64, 65). These models were cursory given that crucial data on stoichiometry was often unavailable and interactions between components were not optimized, but the vision of things to come is there.

With the development of promising new methods for systematically deriving quantitative information on complex stoichiometry (66–68) and for gaining information on the internal organization of complexes (69) more accurate models may in the future be built with the help of available data on known protein structures and computational procedures to optimally dock the components to each other. These so-called docking procedures are becoming increasingly powerful and accessible to non-specialists (70) as recently reviewed (71, 72). In a similar spirit, data from interaction proteomics can be used as a starting point for deriving architectural maps of very large molecular machines (73, 74), and subsequently such maps could be further refined by docking the high resolution structures of the individual components.

In a not too distance future, we might see all these approaches integrated with cryoelectron tomography (75) to yield views of complex molecular architectures in action in the cell. Such views will be a bounty for enthusiastic simulators of cellular processes (76, 77) who will finally have realistic models to feed into their computer programs.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‖The Tier 1 Canada Research Chair in Computational Biology and Bioinformatics. Supported by the Canada Institute for Health Research and the Hospital for Sick Children, Toronto, Canada. To whom correspondence should be addressed. Tel.: 416-813-8339; Fax: 416-813-8755; E-mail: shoshana@sickkids.ca.

** Supported by the McLaughlin Centre for Molecular Medicine.

§§ Supported by La Ligue contre le Cancer (Equipe Labelisée 2005), the CNRS, and the European Union FP6 project 3D-Repertoire (Grant LSHG-CT-2005-512028).

REFERENCES

1. Alberts, B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92,** 291–294
2. Formosa, T., Barry, J., Alberts, B. M., and Greenblatt, J. (1991) Using protein affinity chromatography to probe structure of protein machines. *Methods Enzymol.* **208,** 24–45
3. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert,

[3] M. Vidal, personal communication.

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, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147

4. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., 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, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415,** 180–183
5. 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. U. S. A.* **98,** 4569–4574
6. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., 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., and Rothberg, J. M. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403,** 623–627
7. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440,** 631–636
8. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrin-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440,** 637–643
9. Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433,** 531–537
10. Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., Jr., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003) A protein interaction map of Drosophila melanogaster. *Science* **302,** 1727–1736
11. Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004) A map of the interactome network of the metazoan C. elegans. *Science* **303,** 540–543
12. Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J. P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S. L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3,** 89
13. Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437,** 1173–1178
14. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122,** 957–968
15. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 4285–4288
16. Peregrin-Alvarez, J. M., Tsoka, S., and Ouzounis, C. A. (2003) The phylogenetic extent of metabolic enzymes and pathways. *Genome Res.* **13,** 422–427
17. Doolittle, R. F. (1995) The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64,** 287–314
18. Sjolander, K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* (*Oxf.*) **20,** 170–179
19. Nickell, S., Kofler, C., Leis, A. P., and Baumeister, W. (2006) A visual approach to proteomics. *Nat. Rev. Mol. Cell Biol.* **7,** 225–230
20. Collins, S. R., Kemmeren, P., Zhao, X. C., Greenblatt, J. F., Spencer, F., Holstege, F. C., Weissman, J. S., and Krogan, N. J. (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics* **6,** 439–450
21. Pu, S., Vlasblom, J., Emili, A., Greenblatt, J., and Wodak, S. J. (2007) Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics* **7,** 944–960
22. Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Seraphin, B. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* (*San Diego*) **24,** 218–229
23. Cravatt, B. F., Simon, G. M., and Yates, J. R., III (2007) The biological impact of mass-spectrometry-based proteomics. *Nature* **450,** 991–1000
24. Goll, J., and Uetz, P. (2006) The elusive yeast interactome. *Genome Biol.* **7,** 223
25. Valadkhan, S. (2007) The spliceosome: caught in a web of shifting interactions. *Curr. Opin. Struct. Biol.* **17,** 310–315
26. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34,** D535–D539
27. Colot, H. V., Stutz, F., and Rosbash, M. (1996) The yeast splicing factor Mud13p is a commitment complex component and corresponds to CBP20, the small subunit of the nuclear cap-binding complex. *Genes Dev.* **10,** 1699–1708
28. Izaurralde, E., Lewis, J., McGuigan, C., Jankowska, M., Darzynkiewicz, E., and Mattaj, I. W. (1994) A nuclear cap binding protein complex involved in pre-mRNA splicing. *Cell* **78,** 657–668
29. Lewis, J. D., Gorlich, D., and Mattaj, I. W. (1996) A yeast cap binding protein complex (yCBC) acts at an early step in pre-mRNA splicing. *Nucleic Acids Res.* **24,** 3332–3336

30. Gong, F., Fahy, D., and Smerdon, M. J. (2006) Rad4-Rad23 interaction with SWI/SNF links ATP-dependent chromatin remodeling with nucleotide excision repair. *Nat. Struct. Mol. Biol.* **13,** 902–907
31. Lai, J. S., and Herr, W. (1992) Ethidium bromide provides a simple tool for identifying genuine DNA-independent protein associations. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 6958–6962
32. Nieto, J. M., Madrid, C., Miquelay, E., Parra, J. L., Rodriguez, S., and Juarez, A. (2002) Evidence for direct protein-protein interaction between members of the enterobacterial Hha/YmoA and H-NS families of proteins. *J. Bacteriol.* **184,** 629–635
33. Brohee, S., and van Helden, J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7,** 488
34. Hart, G. T., Lee, I., and Marcotte, E. R. (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics* **8,** 236
35. Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584
36. Van Dongen, S. (2000) *A Cluster Algorithm for Graphs*, Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam
37. de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005) Dynamic complex formation during the yeast cell cycle. *Science* **307,** 724–727
38. Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P., and Vidal, M. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430,** 88–93
39. Simonis, N., van Helden, J., Cohen, G. N., and Wodak, S. J. (2004) Transcriptional regulation of protein complexes in yeast. *Genome Biol.* **5,** R33
40. Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7,** 120
41. Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J. F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabasi, A. L., Tavernier, J., Hill, D. E., and Vidal, M. (2008) High quality binary protein interaction map of the yeast interactome network. *Science* **322,** 104–110
42. Remy, I., and Michnick, S. W. (2004) Mapping biochemical networks with protein-fragment complementation assays. *Methods in Mol. Biol.* **261,** 411–426
43. Eyckerman, S., Verhee, A., der Heyden, J. V., Lemmens, I., Ostade, X. V., Vandekerckhove, J., and Tavernier, J. (2001) Design and application of a cytokine-receptor-based interaction trap. *Nat. Cell Biol.* **3,** 1114–1119
44. Kiemer, L., Costa, S., Ueffing, M., and Cesareni, G. (2007) WI-PHI: a weighted yeast interactome enriched for direct physical interactions. *Proteomics* **7,** 932–943
45. Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* **1,** 349–356
46. Deng, M., Mehta, S., Sun, F., and Chen, T. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res.* **12,** 1540–1548
47. Gentleman, R., and Huber, W. (2007) Making the most of high-throughput protein-interaction data. *Genome Biol.* **8,** 112
48. Hakes, L., Pinney, J. W., Robertson, D. L., and Lovell, S. C. (2008) Protein-protein interaction networks and biology-what's the connection? *Nat. Biotechnol.* **26,** 69–72
49. Huang, H., Jedynak, B. M., and Bader, J. S. (2007) Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput. Biol.* **3,** e214
50. Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008) Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 6959–6964
51. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302,** 449–453
52. von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33,** D433–D437
53. Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hon, G. C., Myers, C. L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., Stark, C., Ho, Y., Botstein, D., Andrews, B., Boone, C., Troyanskya, O. G., Ideker, T., Dolinski, K., Batada, N. N., and Tyers, M. (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J. Biol.* **5,** 11
54. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417,** 399–403
55. Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S. W. (2008) An in vivo map of the yeast protein interactome. *Science* **320,** 1465–1470
56. Mousson, F., Kolkman, A., Pijnappel, W. W., Timmers, H. T., and Heck, A. J. (2008) Quantitative proteomics reveals regulation of dynamic components within TATA-binding protein (TBP) transcription complexes. *Mol. Cell. Proteomics* **7,** 845–852
57. Bertin, N., Simonis, N., Dupuy, D., Cusick, M. E., Han, J. D., Fraser, H. B., Roth, F. P., and Vidal, M. (2007) Confirmation of organized modularity in the yeast interactome. *PLoS Biol.* **5,** e153
58. Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hurst, L. D., and Tyers, M. (2007) Still stratus not altocumulus: further evidence against the date/party hub distinction. *PLoS Biol.* **5,** e154
59. Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hurst, L. D., and Tyers, M. (2006) Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biol.* **4,** e317
60. Kim, P. M., Lu, L. J., Xia, Y., and Gerstein, M. B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314,** 1938–1941
61. Cochran, A. G. (2001) Protein-protein interfaces: mimics and inhibitors. *Curr. Opin. Chem. Biol.* **5,** 654–659
62. Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E., and Skolnick, J. (2006) On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 2605–2610
63. Vitkup, D., Melamud, E., Moult, J., and Sander, C. (2001) Completeness in structural genomics. *Nat. Struct. Biol.* **8,** 559–566
64. Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A. C., Bork, P., Superti-Furga, G., Serrano, L., and Russell, R. B. (2004) Structure-based assembly of protein complexes in yeast. *Science* **303,** 2026–2029
65. Aloy, P., and Russell, R. B. (2006) Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.* **7,** 188–197
66. Kitagawa, N., Mazon, H., Heck, A. J., and Wilkens, S. (2008) Stoichiometry of the peripheral stalk subunits E and G of yeast V1-ATPase determined by mass spectrometry. *J. Biol. Chem.* **283,** 3329–3337
67. Synowsky, S. A., and Heck, A. J. (2008) The yeast Ski complex is a hetero-tetramer. *Protein Sci.* **17,** 119–125
68. van den Heuvel, R. H., and Heck, A. J. (2004) Native protein mass spectrometry: from intact oligomers to functional machineries. *Curr. Opin. Chem. Biol.* **8,** 519–526
69. Hernandez, H., Dziembowski, A., Taverner, T., Seraphin, B., and Robinson, C. V. (2006) Subunit architecture of multimeric complexes isolated directly from cells. *EMBO Rep.* **7,** 605–610
70. Kamal, J. K., and Chance, M. R. (2008) Modeling of protein binary complexes using structural mass spectrometry data. *Protein Sci.* **17,** 79–94
71. Janin, J., and Wodak, S. (2007) The third CAPRI assessment meeting Toronto, Canada, April 20–21, 2007. *Structure* (*Lond.*) **15,** 755–759
72. Lensink, M. F., Mendez, R., and Wodak, S. J. (2007) Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins* **69,** 704–718
73. Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P., and Sali, A. (2007) Determining the architectures of macromolecular assemblies. *Nature* **450,** 683–694
74. Robinson, C. V., Sali, A., and Baumeister, W. (2007) The molecular sociology of the cell. *Nature* **450,** 973–982
75. Kurner, J., Frangakis, A. S., and Baumeister, W. (2005) Cryo-electron tomography reveals the cytoskeletal structure of Spiroplasma mel-

liferum. *Science* **307,** 436–438
76. Ander, M., Beltrao, P., Di Ventura, B., Ferkinghoff-Borg, J., Foglierini, M., Kaplan, A., Lemerle, C., Tomas-Oliveira, I., and Serrano, L. (2004) SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks. *Syst. Biol.* **1,** 129–138
77. Sanford, C., Yip, M. L., White, C., and Parkinson, J. (2006) Cell++—simulating biochemical pathways. *Bioinformatics* (*Oxf.*) **22,** 2918–2925
78. Mackay, J. P., Sunde, M., Lowry, J. A., Crossley, M., and Matthews, J. M. (2007) Protein interactions: is seeing believing? *Trends Biochem. Sci.* **32,** 530–531
79. Giglia-Mari, G., Coin, F., Ranish, J. A., Hoogstraten, D., Theil, A., Wijgers, N., Jaspers, N. G., Raams, A., Argentini, M., van der Spek, P. J., Botta, E., Stefanini, M., Egly, J. M., Aebersold, R., Hoeijmakers, J. H., and Vermeulen, W. (2004) A new, tenth subunit of TFIIH is responsible for the DNA repair syndrome trichothiodystrophy group A. *Nat. Genet.* **36,** 714–719
80. Ranish, J. A., Hahn, S., Lu, Y., Yi, E. C., Li, X. J., Eng, J., and Aebersold, R. (2004) Identification of TFB5, a new component of general transcription and DNA repair factor IIH. *Nat. Genet.* **36,** 707–713
81. Dziembowski, A., Ventura, A. P., Rutz, B., Caspary, F., Faux, C., Halgand, F., Laprevote, O., and Seraphin, B. (2004) Proteomic analysis identifies a new complex required for nuclear pre-mRNA retention and splicing. *EMBO J.* **23,** 4847–4856
82. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504
83. Vlasblom, J., Wu, S., Pu, S., Superina, M., Liu, G., Orsi, C., and Wodak, S. J. (2006) GenePro: a Cytoscape plug-in for advanced visualization and analysis of interaction networks. *Bioinformatics* (*Oxf.*) **22,** 2178–2179
84. Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* **26,** 73–79
85. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30,** 31–34
86. Ben-Hur, A., and Noble, W. S. (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics* (*Oxf.*) **21,** Suppl. 1, i38–i46
87. Zhang, L. V., Wong, S. L., King, O. D., and Roth, F. P. (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics* **5,** 38
88. 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
89. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29
90. Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* (*Oxf.*) **19,** 1275–1283
91. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691
92. Kumar, A., Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K. H., Miller, P., Gerstein, M., Roeder, G. S., and Snyder, M. (2002) Subcellular localization of the yeast proteome. *Genes Dev.* **16,** 707–719
93. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–4257
94. Shatkay, H., Hoglund, A., Brady, S., Blum, T., Donnes, P., and Kohlbacher, O. (2007) SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics* (*Oxf.*) **23,** 1410–1417
95. von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B., and Bork, P. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35,** D358–D362
96. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., and Hermjakob, H. (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.* **35,** D561–D565
97. Bader, G. D., Betel, D., and Hogue, C. W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31,** 248–250
98. Nash, R., Weng, S., Hitz, B., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hong, E. L., Livstone, M. S., Oughtred, R., Park, J., Skrzypek, M., Theesfeld, C. L., Binkley, G., Dong, Q., Lane, C., Miyasato, S., Sethuraman, A., Schroeder, M., Dolinski, K., Botstein, D., and Cherry, J. M. (2007) Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res.* **35,** D468–D471
99. Tong, A. H., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D. S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J. N., Lu, H., Menard, P., Munyana, C., Parsons, A. B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A. M., Shapiro, J., Sheikh, B., Suter, B., Wong, S. L., Zhang, L. V., Zhu, H., Burd, C. G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F. P., Brown, G. W., Andrews, B., Bussey, H., and Boone, C. (2004) Global mapping of the yeast genetic interaction network. *Science* **303,** 808–813
100. Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., and Richardson, J. E. (2007) The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res.* **35,** D630–D637
101. Bieri, T., Blasiar, D., Ozersky, P., Antoshechkin, I., Bastiani, C., Canaran, P., Chan, J., Chen, N., Chen, W. J., Davis, P., Fiedler, T. J., Girard, L., Han, M., Harris, T. W., Kishore, R., Lee, R., McKay, S., Muller, H. M., Nakamura, C., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E. M., Spooner, W., Tuli, M. A., Van Auken, K., Wang, D., Wang, X., Williams, G., Durbin, R., Stein, L. D., Sternberg, P. W., and Spieth, J. (2007) WormBase: new content and better access. *Nucleic Acids Res.* **35,** D506–D510
102. Crosby, M. A., Goodman, J. L., Strelets, V. B., Zhang, P., and Gelbart, W. M. (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res.* **35,** D486–D491
103. Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13,** 2363–2371
104. Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H. W., and Stumpflen, V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.* **34,** D436–D441
105. Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004) The HUPO PSI's molec-

ular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.* **22,** 177–183

106. Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stumpflen, V., Ceol, A., Chatr-aryamontri, A., Armstrong, J., Woollard, P., Salama, J. J., Moore, S., Wojcik, J., Bader, G. D., Vidal, M., Cusick, M. E., Gerstein, M., Gavin, A. C., Superti-Furga, G., Greenblatt, J., Bader, J., Uetz, P., Tyers, M., Legrain, P., Fields, S., Mulder, N., Gilson, M., Niepmann, M., Burgoon, L., De Las Rivas, J., Prieto, C., Perreau, V. M., Hogue, C., Mewes, H. W., Apweiler, R., Xenarios, I., Eisenberg, D., Cesareni, G., and Hermjakob, H. (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.* **25,** 894–898