# Inferring Physical Protein Contacts from Large-Scale Purification Data of Protein Complexes*[S]

**Sven-Eric Schelhorn‡§, Julián Mestre‡¶, Mario Albrecht‡‖, and Elena Zotenko‡§‖**

**Recent large-scale data sets of protein complex purifications have provided unprecedented insights into the organization of cellular protein complexes. Several computational methods have been developed to detect co-complexed proteins in these data sets. Their common aim is the identification of biologically relevant protein complexes. However, much less is known about the network of direct physical protein contacts within the detected protein complexes. Therefore, our work investigates whether direct physical contacts can be computationally derived by combining raw data of large-scale protein complex purifications. We assess four established scoring schemes and introduce a new scoring approach that is specifically devised to infer direct physical protein contacts from protein complex purifications. The physical contacts identified by the five methods are comprehensively benchmarked against different reference sets that provide evidence for true physical contacts.**

**Our results show that raw purification data can indeed be exploited to determine high-confidence physical protein contacts within protein complexes. In particular, our new method outperforms competing approaches at discovering physical contacts involving proteins that have been screened multiple times in purification experiments. It also excels in the analysis of recent protein purification screens of molecular chaperones and protein kinases. In contrast to previous findings, we observe that physical contacts inferred from purification experiments of protein complexes can be qualitatively comparable to binary protein interactions measured by experimental high-throughput assays such as yeast two-hybrid. This suggests that computationally derived physical contacts might complement binary protein interaction assays and guide large-scale interactome mapping projects by prioritizing putative physical contacts for further experimental screens.** ***Molecular & Cellular Proteomics 10: 10.1074/mcp.M110.004929, 1–15, 2011.***

From the ‡Max Planck Institute for Informatics, Campus E1.4, 66123 Saarbrücken, Germany ¶School of Information Technologies, University of Sydney, NSW 2006, Australia

[S] This article contains supplemental material.

Received September 13, 2010, and in revised form, March 15, 2011

⌘ *Author's Choice*—Final version full access.

Published, MCP Papers in Press, March 30, 2011, DOI 10.1074/mcp.M110.004929

Proteins often do not act in isolation, but cooperate in larger assemblies to fulfill their functions. The resulting protein complexes are essential in a variety of cellular processes (1). Thus, the identification and annotation of protein complexes is currently the focus of both experimental and computational analyses (2). Recent advances in experimental technologies for protein purification and identification (3), such as tandem-affinity purification techniques, enabled high-throughput purification screens for protein complexes in several model organisms (4). A typical high-throughput screen entails hundreds of purification experiments, in which a single purification assay determines *prey* proteins that associate with a given *bait* protein through multiprotein complex formation.

Because of a variety of reasons, such as experimental noise, presence of nonspecific interactors, or participation of the bait protein in multiple distinct protein complexes (5), the experimentally obtained purifications are not directly interpretable as biologically relevant protein complexes. Therefore, computational methods are applied to infer these complexes from raw purification data by scoring protein interactions within the purifications. Publication of two independent large-scale screens of protein complexes in the yeast *Saccharomyces cerevisiae* (6, 7) triggered development of several such scoring schemes (6–11) and resulted in a revised catalogue of manually curated yeast complexes (12).

Proteins within a complex are connected by protein interactions. Here, protein interactions often refer to both direct physical contacts, in which two proteins share a common binding interface, and indirect, bridging interactions, in which the proteins do not contact each other directly. Established purification scoring schemes have been shown to perform well in determining the composition of protein complexes by identifying such protein interactions in the purification data. However, these scoring schemes do not discriminate between direct physical contacts and indirect protein interactions. Consequently, less is known about which proteins in large-scale protein purifications form direct physical contacts although this information is crucial for a deeper understanding of protein complex formation and organization.

Furthermore, the difficulty of identifying physical protein contacts within protein complex purifications has hampered the comparison with results of binary protein interaction experiments such as yeast two-hybrid assays. A recent comparison found substantially more true physical contacts from binary assays than purification experiments (13). However,

this analysis did not consider that protein complex purifications contain both direct physical contacts and indirect protein interactions in contrast to binary assays. Because this results in a lower enrichment with physical contacts, a comparison of the experimental assays that concentrates only on putative physical protein contacts would provide deeper insights into the relative merits of each experimental technology.

Even though several experimental and computational methods exist that produce structural models of protein complexes at various levels of resolution (14, 15), structural data required by these approaches is not readily available for the vast majority of complexes detected by large-scale protein purifications. Thus, the main objective of this work is to assess whether and how we can make use of the available purification screens to computationally infer the network of physical contacts within the assayed protein complexes.

Our guiding principle rests upon the observation that proteins forming physical contacts within a complex exhibit stronger associations and thus are more likely to survive purification procedures than proteins that do not form such contacts. A similar observation is central to a hybrid approach developed by the Robinson group in which individual protein complexes are perturbed by experimental techniques to discover physical contacts between proteins within these complexes (16). We hypothesize that even though large-scale purification screens do not directly measure physical contacts within complexes, the resulting experimental data does contain sufficient information to reliably infer these interactions.

The three main contributions of our work are as follows. First, we propose an elegant computational method for scoring pairs of proteins based on their co-occurrence pattern within a combined set of purifications originating from multiple large-scale screens. In contrast to existing scoring schemes, which were originally developed and evaluated to detect co-complexed protein pairs regardless of their mode of interaction, our approach is tuned to detect protein pairs that form direct physical contacts and incorporates experimental replicates in a statistically correct fashion. As a consequence, our method can reliably detect true physical contacts even in the presence of many unspecific or highly transient protein interactions and especially outperforms existing scoring schemes if experimental replicates are available. These properties make our approach particularly suited for the joint analysis of physical contacts from multiple protein purification screens.

Second, we perform a comprehensive evaluation of our and four other published scoring methods on the task of detecting physical contacts. Each method scores purification data from two recent large-scale experiments in yeast (6, 7). The results of all scoring methods are benchmarked against several reference sets that represent complementary evidence for physical contacts. The reference sets are derived from experimentally determined physical interactions, three-dimensional structures of protein complexes, manually curated catalogs of protein complexes, and genetic interaction profiles. In particular, we assess the scoring methods by inferring specific physical contacts in two challenging and biologically relevant purification data sets containing repeated purifications of molecular chaperones or protein kinases.

Third, we compare top-ranking physical contacts inferred by our method to two recent high-throughput interaction data sets derived by the experimental techniques *yeast two-hybrid* (Y2H)[1] (13) and *protein fragment complementation assay* (PCA) (17) and address intrinsic differences of high-throughput approaches to mapping physical interactomes.

## EXPERIMENTAL PROCEDURES

### *Purification Data*

*Large-scale Purification Data*—We used a combined set of purifications from two large-scale screens in yeast (6, 7). Raw experimental data was obtained from the supporting web-site (http://interactome-cmp.ucsf.edu) of one of the existing scoring methods included in our evaluation, the PE method (8). Purification data from the Gavin *et al.* (6) screen was taken as it is, whereas purification data from the Krogan *et al.* (7) screen was filtered as follows. The Krogan team used two different experimental protocols, liquid chromatography MS (LCMS) and matrix-assisted laser desorption ionization (MALDI), for prey identification. For each purification, we retained preys having MALDI identification score of at least 1.25 or LCMS confidence score of at least 99%. We further filtered out preys that were identical to baits in their respective purifications from both screens. Last, we combined purifications from the two individual screens into one data set, which we denoted as Large-Scale set of purifications. Table I summarizes purification data from individual screens as well as from the Large-Scale set.

*Protein Kinase and Phosphatase Purification Data*—In addition to the Large-Scale data set, we used a specialized purification data set focusing on *kinase* and *phosphatase* interactions in yeast from a recent experimental study by Breitkreutz *et al.* (18). The bait proteins in the Breitkreutz data were screened with three different tag systems (FLAG, HA, and TAP) and the purified prey proteins include information about peptide (spectral) counts that can be used as a semi-quantitative measure of absolute protein abundance. We obtained raw purification data for all three tag systems from the supporting website (http://yeastkinome.org). Subsequently, the purifications were filtered to (1) exclude tag-specific contaminant proteins identified by control experiments in the original study and to (2) remove unreliably identified prey proteins with Mascot scores ≤35. The resulting Breitkreutz purification data set is summarized in supplemental Table S1.

### *Reference Sets*

*High-confidence Physical Contacts*—Binary gold standard (BGS) protein interactions used in a recent assessment of binary experimental methods (13) as well as the experimental Y2H data generated in

[1] Y2H, Yeast two-hybrid; BGS, Binary gold standard; ISA, improved socio-affinity; LCMS, Liquid chromatography-mass spectrometry; MALDI, Matrix-assisted laser desorption/ionization; MIPS, Munich Information Center for Protein Sequences; PCA, Protein fragment complementation assay; PDB, Protein Data Bank; PE, Purification Enrichment; SA, Socio-Affinity; SAINT, Significance Analysis of the Interactome; SGD, Saccharomyces Genome Database; 3DID, 3D interacting domains database.

the same study were obtained from the CCSB interactome database. Note that, because no true gold standard for binary protein interactions is available, we decided to use the BGS naming convention from (13) to allow better comparability of our work. Further binary interactions measured by a recent protein-fragment complementation assay (PCA) (17) were extracted from the Saccharomyces Genome Database (SGD) (19).

Binary interactions originating from experimental assays that directly measure physical protein contacts were obtained from IntAct (20) and SGD. This set of interactions was filtered to exclude physical contacts that solely rely on evidence from the Y2H and PCA binary protein interaction data sets. The filtered data set was used to (1) define two reference sets: a Chaperone reference set containing only interactions involving yeast molecular chaperones and (2) a Kinase reference set including solely interactions involving yeast kinases and phosphatases.

*Protein Complexes and Domain Interactions*—Protein complexes derived from Gene Ontology annotations as provided by SGD and manually curated protein complexes from the Munich Information Center for Protein Sequences (MIPS) (21) were imported from the websites of the respective organizations. For the validation of inferred physical contacts on the level of protein domain interactions, a mapping table from SGD was obtained to assign UniProt accession numbers to all proteins in the Large-Scale purification data. Subsequently, globular Pfam-A domain annotations for these proteins were obtained from the InterPro database (22). We restricted the used annotations to globular domains because this type of protein regions is especially well characterized and known to be involved in stable protein interactions (23). A list of Pfam-A domain interaction partners derived from structures of interacting proteins in the Protein Data Bank (PDB) (24) was obtained from the 3DID database (25). This 3DID reference set was post-processed to exclude domain interactions that solely rely on PDB intra-chain interactions.

*Genetic Interaction Profiles*—Interaction confidence scores derived from *in vivo* synthetic genetic interactions (GI) were obtained from a recent large-scale functional study in yeast (26). Of the several available data sets containing genetic interaction scores for pairs of proteins, we selected the *lenient cutoff interaction set* that offered the highest coverage of yeast proteins whereas, at the same time, including only statistically significant interactions. This data was used to reconstruct genetic interaction profiles for all proteins in the Large-Scale purification data set. Consequently, genetic interaction profile similarities for pairs of proteins were generated by computing Pearson's correlation coefficients between the genetic interaction profiles of all protein pairs. Analogous to the original study, all protein pairs with a genetic interaction profile similarity $\geq 0.2$ were used to form a functional map of the Large-Scale purification data. Protein pairs in this map are denoted as the GI reference set.

*Scoring Methods*—Let $\Phi = \{\phi_1,\ldots,\phi_N\}$ be a set of purifications where each purification $\phi_k$ is composed of a bait protein $bait_k$ and a set of prey proteins $preys_k$. We will use $n_k$ to denote the number of preys in $\phi_k$ and designate $N = \Sigma_k n_k$ as the size of the *multiset* of all $preys_k$. For a pair of proteins, $i$ and $j$, let $S_{i\to j}$ be the number of times $j$ is observed among preys in purifications performed with $i$ as bait and $preys_k$ be the number of times $i$ and $j$ are observed as preys in purifications performed with a third protein as a bait. In what follows, the experimental observations $S$ and $M$ are also denoted as *spoke observations* and *matrix observations*, respectively. Some scoring schemes combine $S$ and $M$ into one number $O_{i,j} = S_{i\to j} + S_{j\to i} + M_{i,j}$. We will denote by: $S^{\text{null}}_{i\to j}$, $M^{\text{null}}_{i,j}$, and $O^{null}_{i,j}$ random variables representing these different types of observation counts under appropriate null models. Next, we briefly introduce the scoring methods that are assessed in this work.

*Socio-affinity Scores*—The socio-affinity (SA) scoring scheme is one of the first approaches for scoring purification data and was developed by Gavin *et al.* to interpret the results of their large-scale screen (6). The score is based on three counts $S_{i\to j}$, $S_{j\to i}$, and $M_{i,j}$ is given by:

$$\text{SA}(i,j) = \log \frac{S_{i\to j}}{E[S^{\text{null}}_{i\to j}]} + \log \frac{S_{j\to i}}{E[S^{\text{null}}_{j\to i}]} + \log \frac{M_{\{i,j\}}}{E[M^{\text{null}}_{\{i,j\}}]} \qquad \text{(Eq. 1)}$$

The distributions of $S^{\text{null}}_{i\to j}$ and $M^{\text{null}}_{i,j}$ are modeled based on the assumption that purifications are drawn uniformly at random from the observed multiset of preys. This means that $\phi^{\text{null}}_k$ is formed through $n_k$ independent random selections of preys where the probability of selecting the prey protein $j$ is equal to its relative frequency $f_j = |\{\phi_k \mid j \in \text{preys}_k\}| \cdot N^{-1}$. Under this null model, the expected values of $S^{\text{null}}_{i\to j}$ and $M^{\text{null}}_{i,j}$ are given by $E[S^{\text{null}}_{i\to j}] = \Sigma_{k:i=\text{bait}_k} f_j n_k$ and $E[M^{\text{null}}_{i,j}] = \Sigma_k f_i f_j \binom{n_k}{2}$.

*Improved Socio-affinity Scores*—In this work, we propose a modification of the SA scoring scheme termed *improved socio-affinity* (ISA) score. It makes full use of repetitive purifications and concentrates on spoke observations $S$ to improve the detection of physical contacts. In particular, we adopt the null model used for SA scores and derive the ISA score as follows:

$$\text{ISA}(i,j) = -\log \Pr(S^{\text{null}}_{i\to j} \geq S_{i\to j}) - \log \Pr(S^{\text{null}}_{j\to i} \geq S_{j\to i}) \qquad \text{(Eq. 2)}$$

To compute $\Pr(S^{\text{null}}_{i\to j} \geq S_{i\to j})$ and $\Pr(S^{\text{null}}_{j\to i} \geq S_{j\to i})$, we introduce an indicator random variable $X_{j,k}$ that corresponds to the selection of protein $j$ into the set of preys of $\phi^{\text{null}}_k$, thus $\Pr(X_{j,k}) = 1 - (1 - f_j)^{n_k}$. We then note that $S^{\text{null}}_{i\to j}$ is a sum of independent binary random variables: $S^{\text{null}}_{i\to j} = \Sigma_{k:i=\text{bait}_k} X_{j,k}$. Since $\Pr(X_{j,k})$ depends on the size of $\phi_k$, it is, in general, not the same for different purifications performed with bait protein $i$. As a result, the distribution of $S^{\text{null}}_{i\to j}$ is not binomial. To alleviate this problem, we set $\Pr(X_{j,k}) = 1 - (1 - f_j)^{\hat{n}}$, where $\hat{n}$ is the average size of purifications performed with $i$, and use the binomial distribution to compute $\Pr(S^{\text{null}}_{i\to j} \geq S_{i\to j})$. To avoid a situation where a single observation with a rare prey protein receives very high scores, we adjust background prey frequencies $f_j$ by adding a constant $\epsilon$ fraction of each prey to each purification. In this work we used $\epsilon = 0.0025$.

*Purification Enrichment Scores*—The purification enrichment (PE) scoring scheme was proposed by Collins *et al.* (8) as an alternative to the original SA scores to analyze the combined set of purifications from two recent large-scale screens in yeast (8). The authors adopted a more sophisticated statistical model to score evidence for each observation $o$ separately:

$$\text{PE}(i,j) = \sum_o \log\left(\frac{\Pr(o \mid i \text{ and } j \text{ interact})}{\Pr(o \mid i \text{ and } j \text{ do not interact})}\right) \qquad \text{(Eq. 3)}$$

The detailed description of the statistical model used to derive the probabilities above is beyond the scope of this paper and the interested reader is referred to the original publication (8). We just note here that the estimation of parameters used by the model is not straightforward and requires a representative set of gold standard interactions.

*Hart scores*—Another scoring scheme was proposed by Hart *et al.* (9) with a particular emphasis on joint analysis of experimental data from several large-scale purification screens of protein complexes (9). In this approach, the scores are based on the combined number of observations, $O_{i,j}$, and are computed as:

$$\text{HART}(i,j) = -\log \Pr(O^{\text{null}}_{i,j} \geq O_{i,j}) \qquad \text{(Eq. 4)}$$

The distribution of $O_{i,j}^{\text{null}}$ is modeled based on the assumption that interactions of a protein are selected uniformly at random from the multiset of all observed interactions. More precisely, for a pair of proteins $i$ and $j$, $O_i = \Sigma_j O_{i,j}$ interactions are selected uniformly at random from the ground set of $O = \Sigma_{i,j} O_{i,j}$ interactions that contains $O_j = \Sigma_i O_{i,j}$ ''relevant'' (involving $j$) interactions and $O - O_j$ ''irrelevant'' (not involving $j$) interactions. The statistical significance of the observed $O_{i,j}$ is then assessed by determining the probability that at least $O_{i,j}$ ''relevant'' interactions are selected. Under this null model, $O_{i,j}^{\text{null}}$ has a hyper-geometric distribution and $\Pr(O_{i,j}^{\text{null}} \geq O_{i,j})$ can be efficiently computed. We note that both the SA and Hart null models are simple and parameter free, resulting in efficient computational procedures. However, whereas the SA null model preserves the structure of original purifications, the Hart null model does not.

*IDBOS Scores*—Recently, the Interaction Detection Based On Shuffling (IDBOS) scoring scheme was proposed for scoring purification data with an emphasis on the prediction of direct physical protein interactions (27). In this approach, the scores are based on the combined number of observations, $O_{i,j}$, and are computed as follows:

$$\text{IDBOS}(i,j) = \frac{O_{i,j} - E[O_{i,j}^{\text{null}}]}{S[O_{i,j}^{\text{null}}]} \qquad \text{(Eq. 5)}$$

The distribution of $O_{i,j}^{\text{null}}$ is modeled by assuming that the observed purifications are randomly permuted. The resulting null model is very similar to the one used by the SA and ISA approaches. The main difference is the accurate modeling of observed purifications—the IDBOS null model does not allow random instances where a prey appears multiple times in a single purification. However, this small gain in accuracy comes at a high computational cost. Because the resulting distribution of $O_{i,j}^{\text{null}}$ is much more complex, extensive numerical simulations are required to estimate its properties. The authors perform $10^6$ numerical randomization experiments to estimate the expected value of $O_{i,j}^{\text{null}}$, $E[O_{i,j}^{\text{null}}]$, and its standard deviation $S[O_{i,j}^{\text{null}}]$.

*SAINT Scores*—The Significance Analysis of Interactome (SAINT) scoring scheme was recently introduced to detect non-specifically interacting proteins in the Breitkreutz purification data (18). The method is depending on the use of peptide counts, an additional type of experimental data that was generated during the peptide identification phase of the Breitkreutz screen and can be used as a semi-quantitative measure of absolute protein abundance. SAINT employs a mixture of Poisson distributions to heuristically compute posterior probabilities of specific interactions between proteins based on the peptide counts. Because of the high complexity of the model, presentations of detailed theoretical underpinnings of SAINT are beyond the scope of this work. However, we note here that SAINT is a comparatively complex scoring method with many free parameters and that it requires the availability of experimental peptide count data. Both properties hinder its applicability to publicly available large-scale purification data.

*Score Implementations*—SA, ISA, and Hart scores were computed using in-house Python scripts on the Large-Scale set of purifications. Because of the computational complexity of IDBOS and PE scores, these scores were obtained from the original publications. Although PE scores were computed on the Large-Scale set of purifications by the authors, IDBOS does not support the computation of scores based on multiple sources of purification data (personal communication). Therefore, we used the IDBOS scores computed on the Gavin data because these showed the best performance among all scored data sets in the original publication (27). Because of its reliance on peptide count data, SAINT is not applicable to the Large-Scale set of purifications. SAINT scores for the Breitkreutz purification data were obtained from the original publication (18).

The purification data and the reference sets as well as all inferred physical contacts are available on request from the authors.

*Salama-Quade Rank Correlation*—The Salama-Quade rank correlation coefficient measures similarity between two different rankings of a set of elements (28). It was developed as an alternative to standard rank correlation measures, such as *Spearman's rho* and *Kendall's tau*, for situations where agreement in low ranks is more important than agreement in high ranks.

Let $\{1, \ldots, m\}$ be a set of elements, $R_1(i)$ be the rank of element $i$ under the first method, and $R_2(i)$ be the rank of element $i$ under the second method. The Salama-Quade coefficient measures the agreement between rankings $R_1$ and $R_2$ and is given by $\text{SALAMA-QUADE}(R_1,R_2) = \Sigma_{k=1}^{K} T_k/k$, where $T_k$ is the number of elements having rank less or equal to $k$ under both $R_1$ and $R_2$. For similarity values in Fig. 1*B* we used $K = 10{,}000$ and normalization $\text{SALAMA-QUADE}(R_1,R_2)/K$.

## RESULTS

*Scoring Purification Data*—A purification represents the outcome of an experiment in which a single *bait protein* is tagged and biochemically co-purified with *prey proteins* that associate with the bait by forming of one or several protein complexes. For example, in the screen by Gavin *et al.* (6), a specific purification using the $\alpha$-subunit of clathrin adaptor complex AP-2 as bait contained 22 prey proteins. Three of the co-purified preys correspond to the other subunits of this hetero-tetrameric complex; the other preys might be either unknown subunits of the AP-2 complex, subunits of other complexes in which the $\alpha$-subunit participates, or nonspecific interactors.

Even though the interpretation of a single purification is limited, combined purifications from several large-scale screens contain repeated observations of associated proteins that may indicate true protein-protein interactions. Over the years, several computational approaches were developed to integrate experimental observations across purifications in order to infer pairs of interacting proteins. Four major approaches utilizing raw experimental data are SA scores (6), PE scores (8), scores developed by Hart *et al.* (9) (Hart), and recently published IDBOS scores (27). However, none of these methods is ideally suited for identifying *direct physical contacts* between proteins within a complex through the joint analysis of purifications from several large-scale screens.

In this work we propose a novel scoring method specifically tailored for using repeated purifications. In the following, we briefly describe the main features of our new ISA scoring method.

A single purification provides two kinds of experimental evidence for protein interactions: *spoke observations* supporting interactions between the bait and each of the preys, and *matrix observations* supporting interactions between every pair of preys. In some cases, however, matrix observations are much less reliable than spoke observations. In particular, a large fraction of matrix observations from purifications containing several small complexes would support non-existing interactions between proteins in distinct complexes. The dis-

tribution of protein complex sizes in manually curated catalogues of protein complexes in yeast suggests that the majority of complexes are small; about 64% of complexes in MIPS catalog, for example, have up to four subunits. In comparison, the average number of preys in the purifications in our data set is about 10 (see Table I). It appears therefore that the majority of purifications are indeed composed of several complexes and thus provide many misleading matrix observations. Although relying on potentially misleading matrix observations does not adversely affect scoring of co-complexed protein pairs, the task of identifying direct physical contacts is more sensitive toward misleading observations. Consequently, ISA is cautious and derives interaction confidence scores solely from the more reliable spoke observations whereas completely discarding matrix observations. This is in stark contrast to the other four methods SA, PE, Hart, and IDBOS that derive their scores from a mixture of both spoke and matrix observations.

Similar to related approaches, our method uses statistical techniques to derive interaction confidence scores from spoke observations contained in the experimental purification data. Specifically, for each pair of proteins, the number of spoke observations in the experimental data is compared with the number of such observations under an appropriate null model. Our novel method ISA adopts the null model of Gavin *et al.* introduced in the context of the SA scoring method (6). This null model preserves size and content of the original purifications, but selects prey proteins for each purification uniformly at random from the multiset of preys. Even though more sophisticated null models were proposed in the context of later scoring methods such as Hart, we believe that the SA null model is ideally suited for scoring complex purification data. On one hand, it is simple enough to allow analytical derivation of statistical significance. On the other hand, it realistically models the observed data by preserving much of the structure of the original purifications such as the identity of bait proteins, purifications sizes, and frequency of prey proteins. However, one of the main problems with the SA approach is that additional observations supporting a protein interaction result in a disproportionally small increase of the SA score. This poses a problem when purifications from several independent screens are jointly analyzed. Therefore, as a major improvement over the SA method, ISA scores are derived through statistical *p-value* computations that allows attributing higher confidence to putative physical contacts with multiple supporting observations originating from experimental replicates.

TABLE I

*Summary of purification data from two independent large-scale complex purification screens in yeast, denoted here as Gavin and Krogan, as well as for the combined Large-Scale set. For each screen the number of purifications, the number of distinct bait proteins, the number of distinct prey proteins, the average number of preys per purification, and the number of distinct bait-prey and distinct bait-prey and prey-prey pairs are shown*

| | Gavin | Krogan | Large-Scale |
|---|---|---|---|
| purifications | 1912 | 3999 | 5911 |
| baits | 1754 | 2178 | 2830 |
| preys | 1813 | 3505 | 3759 |
| avg. number of preys | 10.56 | 10.31 | 10.39 |
| protein interactions (bait-prey pairs) | 18,206 | 32,525 | 47,254 |
| protein interactions (bait-prey and prey-prey pairs) | 82,202 | 182,134 | 238,154 |

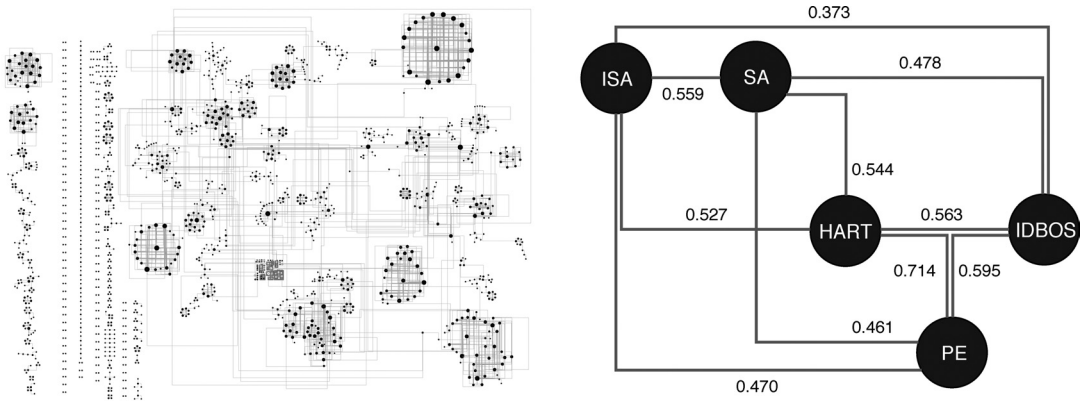
*Scoring Two Large-scale Purification Experiments in S. cerevisiae*—We used the four established scoring schemes SA, Hart, PE, and IDBOS as well as our own approach to score a combined set of purifications from two recent large-scale screens of protein complexes in *S. cerevisiae* (6, 7). In this section, we examine top-ranking inferred physical contacts between proteins by our method and relate them to results of the other four scoring methods.

Table II lists ten inferred physical contacts having the highest ISA scores. All but two interactions in the top-ten list are supported by small-scale experiments reported in the literature. Four top-ten physical contacts receive low scores under the SA method that highlights one of the main differences between the SA and ISA approaches. Consider, for example, the interaction

TABLE II

*A list of top-10 physical contacts inferred by the ISA score. For each physical contact, the number of supporting spoke observations ($S_{i\to j}$ and $S_{j\to i}$), number of supporting matrix observations ($M_{i,j}$), rank under the other scoring schemes, and number of distinct supporting SGD literature references are listed*

| i | j | $S_{i\to j}$ | $S_{j\to i}$ | $M_{i,j}$ | SA | PE | Hart | IDBOS | references |
|---|---|---|---|---|---|---|---|---|---|
| UBP2 | RUP1 | 11 | 7 | 0 | 1568 | 403 | 114 | 67 | 2 |
| RFA1 | RFA2 | 16 | 3 | 13 | 1131 | 256 | 91 | 2371 | 3 |
| HAT2 | HHT2 | 0 | 23 | 0 | 56934 | 7624 | 435 | NA | 1 |
| HIF1 | HHT2 | 0 | 22 | 0 | 53035 | 6250 | 444 | NA | 0 |
| HIF1 | HAT1 | 7 | 14 | 32 | 1857 | 70 | 5 | 321 | 2 |
| HHT2 | HAT1 | 22 | 0 | 0 | 58294 | 8010 | 622 | NA | 1 |
| SRB4 | SRB5 | 5 | 15 | 14 | 1322 | 44 | 170 | 264 | 12 |
| SPT16 | POB3 | 16 | 2 | 55 | 2705 | 791 | 4 | 2326 | 6 |
| HHT2 | PSH1 | 18 | 0 | 0 | 47045 | 6862 | 702 | NA | 0 |
| UBA2 | AOS1 | 6 | 5 | 0 | 899 | 1113 | 412 | NA | 1 |




(A) Induced network of physical contacts (B) Similarities of scoring methods

FIG. 1. **Top-ranking physical contacts inferred by the ISA method and their relation to physical contacts inferred by other methods.** *A*, A network induced by the 3000 protein interactions having the top ISA score ranks. *B*, Similarity of the inferred physical contacts generated by the five scoring methods. Nodes represent different scoring schemes. Edges are labeled with the Salama-Quade correlation coefficient, which measures agreement in the ranking of inferred protein contacts induced by the scores of the corresponding methods.

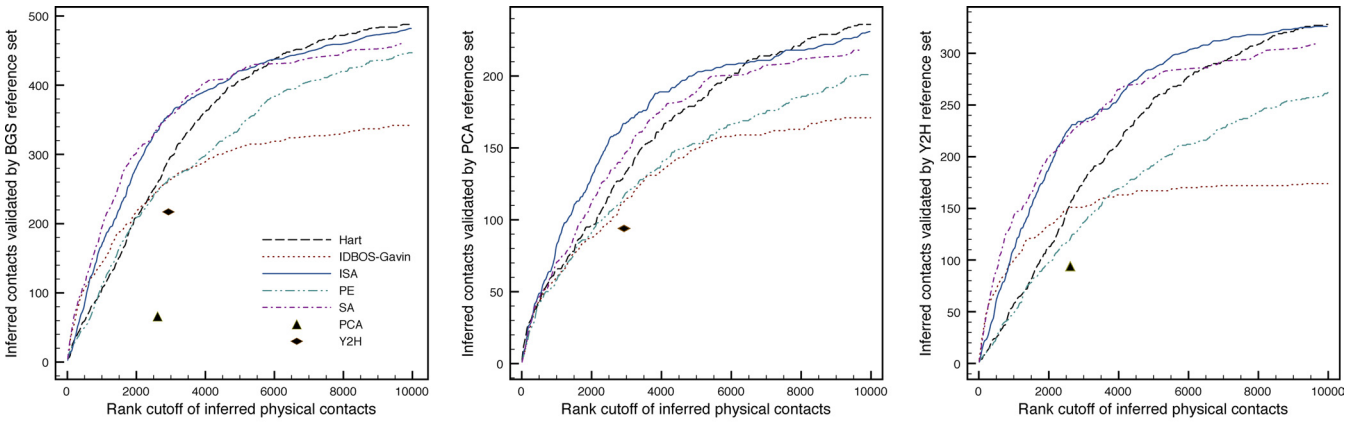
between HAT2 and HHT2 proteins, which is ranked third by the ISA method and 56,934 by the SA method. The HAT2 protein appears as prey in 23 out of 27 purifications performed with HHT2. Still, the SA method assigns low weight to repeated observations of the kind 'HAT2 purifies HHT2′, resulting in a high rank number of the corresponding physical contact.

In general, top-ranking physical contacts inferred by our method are expected to be enriched with interactions involving proteins that were purified multiple times. Indeed, if a given bait protein was purified repeatedly in multiple purifications, its interaction partners can and should be determined with increased confidence. For instance, our experimental data contains 27 purifications performed with HHT2 as bait, which support a total of 124 protein interactions. Out of these 124 interactions, 15 have sufficiently high ISA scores to be included in the top-3,000 inferred physical contacts. Fig. 1*A* depicts a network induced by the top-3,000 inferred physical contacts inferred by ISA. The network is sparse and modular, which agrees well with our intuition for the network of direct physical interactions within stable multiprotein complexes.

To assess the overall similarity among the five scoring methods we used the Salama-Quade rank correlation coefficient. The Salama-Quade coefficient belongs to a family of measures that assign greater weight to agreement in top-ranked elements and thus is more suitable for our purpose than other, more standard, rank correlation measures (see Experimental procedures). As we argue below, out of 238,154 possible physical contacts supported by raw purification data, only about 3000 can be reliably scored. Therefore, the relative ordering of inferred physical contacts beyond this cutoff is less reliable and should affect the similarity score to a lesser extent. Fig. 1*B* shows Salama-Quade rank correlation for every pair of methods. Ranking of inferred physical contacts induced by ISA scores is quite distinct from rankings induced by other scoring methods. Based on similarity scores, the methods can be grouped into two clusters, one including Hart, PE, and IDBOS methods and another containing SA and ISA methods. We hypothesize that this grouping reflects a major difference among the scoring methods, namely, treatment of matrix observations. Although the Hart, PE, and IDBOS methods treat spoke and matrix observations equally, the ISA method completely ignores matrix observations. SA implicitly downplays the contribution of matrix observations by assigning low weight to repeated matrix observations and, as a result, is grouped together with the ISA method.

*Benchmarking Against Reference Sets of Physical Contacts*—In this section, we assess the ability of the four established scoring schemes and our new approach to detect physical contacts within protein purifications. Because, to the best of our knowledge, there is no comprehensive gold standard set of protein interactions that form physical contacts within protein complexes, we approach the evaluation task from four different directions. First, we compare top-ranked inferred physical contacts between proteins to protein interactions derived by experimental techniques that directly assay protein pairs able to physically interact. However, as we argue later on, interactions detected by these techniques represent only a small fraction of physical contacts present in protein complexes. Therefore, we resort to additional, albeit less direct, procedures to assess the performance of the scoring methods by (1) relying on three-dimensional structures of protein complexes, by (2) using manually curated catalogs of protein complexes, and by (3) employing synthetic genetic interaction profiles.

*Experimentally Determined Binary Protein Interactions*—For the first evaluation, we compiled three reference sets of experimentally validated binary protein interactions. The first two data sets originate from recent high-throughput interactome screens in yeast: one employing the Y2H technique (13)




(A) BGS reference set (B) PCA reference set (C) Y2H reference set

FIG. 2. **Assessment of inferred physical protein contacts by five scoring methods against binary experimental reference sets that provide direct evidence for physical contacts.** Inferred physical contacts are ranked by scores of the corresponding scoring method. For each reference set, performance of all methods is measured by plotting the number of top-ranking inferred physical contacts of a method against the number of these contacts that are confirmed by the reference set.

and another using PCA (17). The third data set, BGS, contains manually curated yeast interactions supported by literature and is taken from an extensive validation of the Y2H method (13).

Fig. 2 shows how well the scoring methods perform in identifying true physical contacts from the reference sets. Note that although all methods are able to infer several physical contacts beyond the depicted 10,000 ranks, physical contacts at these high cutoffs have only very low confidence and are thus omitted. Notably, whereas SA and ISA methods have comparable performance across assessments, both of them outperform other approaches on all three reference sets. Moreover, the performance of all approaches starts to level off at about 3,000 to 4,000 ranks. We hypothesize that this number constitutes a reasonable limit on the number of physical contacts that can be reliably inferred given the available experimental data. This number of reliably inferable contacts also corresponds roughly to the number of direct binary interactions measured by high-throughput experimental techniques: the Y2H data set and the PCA data contain 2,930 and 2,616 interactions, respectively. As can be derived from Fig. 2*A*, Y2H and PCA data sets are less enriched in manually curated BGS interactions than an equivalent number of top-scoring interactions extracted from purification data. This suggests that physical contacts inferred by purification scoring schemes are at least qualitatively comparable and often superior to Y2H and PCA experimental data sets.

*Three-dimensional Structures of Multiprotein Complexes*—In this assessment of inferred physical contacts, we rely on experimentally determined structures of protein complexes deposited in the PDB (24). Unfortunately, only crystal structures of about 250 interactions between proteins in yeast are available (29). Therefore, the use of PDB structures for the assessment of putative physical contacts is not possible because of the low coverage of the validation set. However, physical contacts in stable multiprotein complexes are typically formed by pairs of structural protein domains (23), and members of an evolutionarily conserved domain family typically share a common set of domain binding partners. Accordingly, a set of protein interactions that correspond to physical contacts within yeast protein complexes should be enriched in domain pairs that are known to interact. Consequently, to achieve a higher coverage of true physical contacts, we perform this evaluation at the level of PDB-validated physical contacts between protein domains rather than at the level of interacting proteins.

Several resources exist that derive pairs of interacting domains from crystal structures of protein complexes in the PDB. In this work, we use the latest release of the 3DID database (30) to obtain interactions between domains that are annotated to at least one yeast gene. These interactions are denoted as 3DID reference set and compared with a set of domains induced by top-ranking inferred physical contacts. More specifically, for each method, all domain pairs were ranked according to the best-ranking inferred physical protein contact that could be formed by the domain pair. The results of this evaluation are presented in Fig. 3*A*. Again, the SA and ISA methods significantly outperform other approaches over the range of 3,000 to 4,000 inferred physical contacts that are reliably supported by experimental data. At the same time, both SA and ISA perform comparably to the Y2H binary experimental data set with about 240 true domain interactions at a rank cutoff of 4,000 physical contacts.

*Functional Similarity Derived from Genetic Interaction Profiles*—To assess the functional similarity of proteins inferred to form physical contacts, we rely on *in vivo* genetic interaction profiles measured by a recent, functionally unbiased large-scale screen (26). Although the results of this screen do not provide direct evidence for physical contacts, physically interacting proteins that carry out similar functions often exhibit

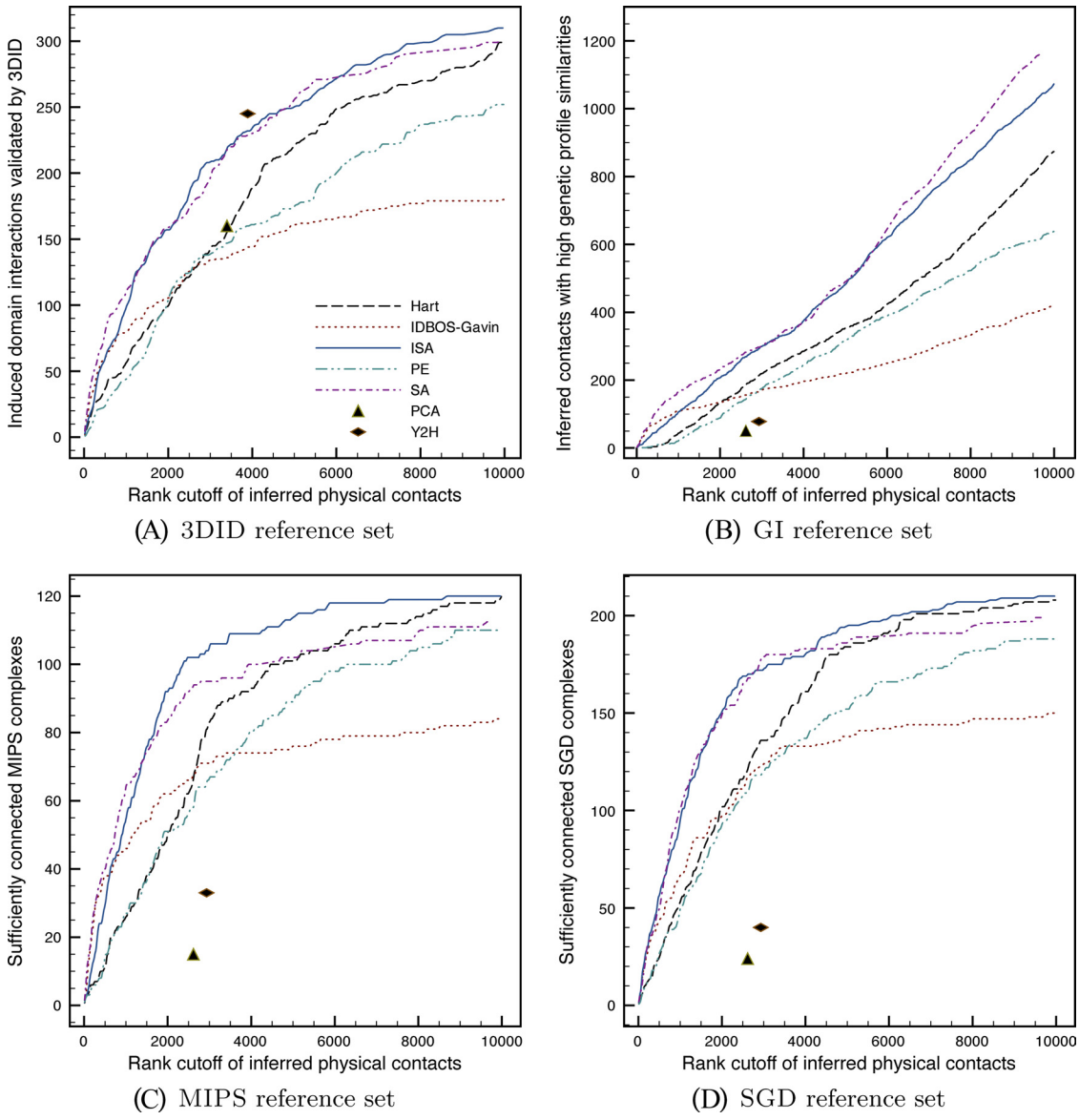
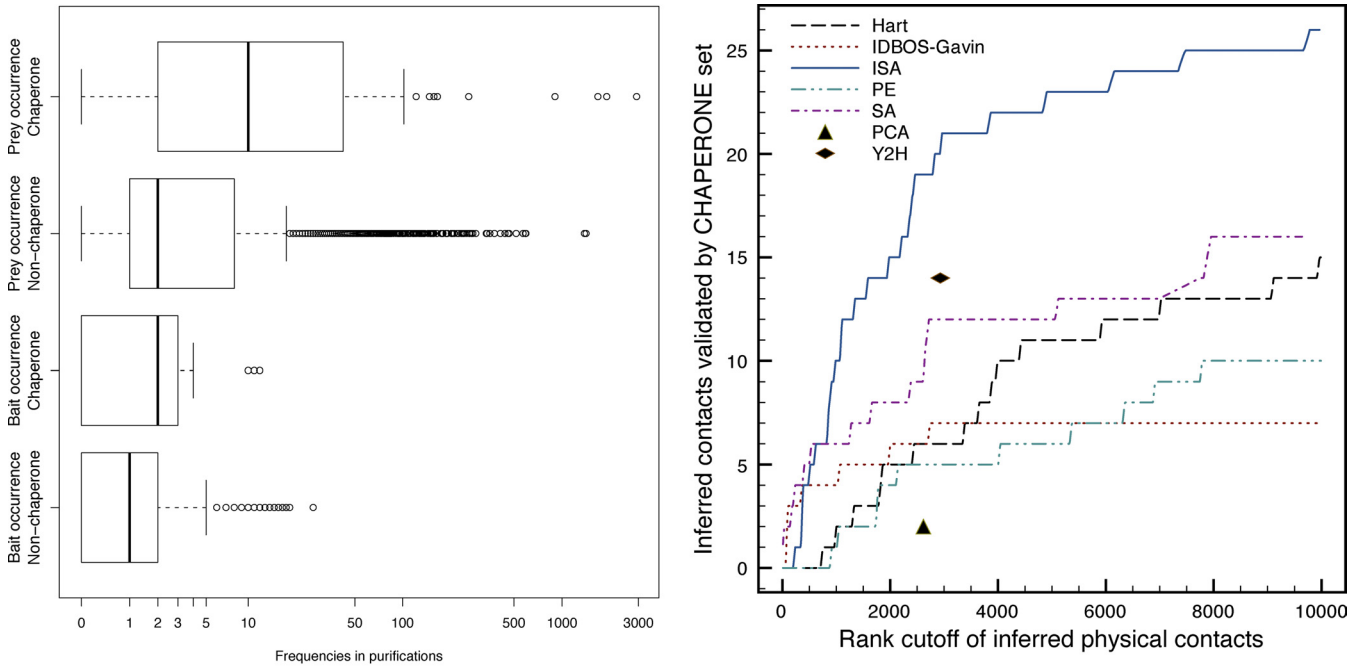



FIG. 3. **Assessment of inferred physical protein contacts by five scoring methods against reference sets that provide indirect evidence for physical contacts.** Inferred physical contacts are ranked by scores of the corresponding scoring method. *A*, Physical contacts are evaluated by their enrichment in protein domains that are known to interact in crystal structures of protein complexes. *B*, Functional similarity of proteins involved in inferred physical contacts is assessed by correlating genetic interaction profiles of these proteins. *C*, *D*, Performance is measured by plotting the number of complexes that are *sufficiently connected* by top-ranking inferred physical contacts for different rank cutoffs. We consider a complex sufficiently connected by a set of inferred physical contacts if the physical contacts reduce the number of connected components within the complex to less than 50% compared with the unconnected complex.

strongly correlated genetic interaction profiles (26). We employed the profile data of the Costanzo *et al.* study (26) to define a GI reference set containing protein pairs with high genetic interaction profile similarities. This reference set was used to assess the functional similarity of inferred physical contacts from all five methods (see Fig. 3*B*). The assessment shows that the socio-affinity based methods SA and ISA significantly outperform other scoring methods as well as the binary experimental data sets PCA and Y2H in enriching for functionally similar protein pairs.

*Manually Curated Catalogs of Multiprotein Complexes*—Several catalogues of manually curated protein assemblies in yeast are publicly available, such as MIPS and SGD complexes. Unfortunately, these high-quality data sets only provide information on the protein composition of each assembly and do not include the network of physical contacts present within the complex. Therefore, we rely on the following assumption to assess the inferred physical contacts with the MIPS and SGD data sets: physical contacts within a complex connect all its member proteins. This




(A) Protein frequency distributions (B) CHAPERONE reference set

FIG. 4. ***A*, Box plots showing the frequency distribution of chaperone-proteins and nonchaperone proteins in the Large-Scale data set, in their roles as bait or prey within purifications.** Frequencies on the abscissa are scaled logarithmically. Boxes represent 50% of the data of a given distribution, whereas bold vertical lines denote the median. *B*, Assessment of inferred and experimentally obtained physical contacts involving molecular chaperones against the Chaperone reference set of experimentally confirmed binary chaperone interactions. The BGS reference set does not contain a sufficient number of chaperone interactions to allow validation.

means that, within a given complex, every protein is connected to every other protein through a network of physical contacts. Consequently, the quality of a set of inferred physical contacts can be estimated by assessing how well these physical contacts connect manually curated complexes. Figs. 3*C* and 3*D* depict how well top-ranking inferred physical contacts from different scoring methods connect complexes in the two manually curated catalogs. It is noticeable that the results generated by purification scoring schemes seem to be significantly better suited to connect these complexes than data sets originating from Y2H and PCA techniques, with the best performing scoring methods SA and ISA connecting more than three times as many complexes than the Y2H data at a rank cutoff of approximately 3,000 physical contacts.

*Inferring Physical Protein Contacts From Repeated Purifications*—The use of experimental replicates in a statistically meaningful fashion to account for experimental errors is a current theme in interactomics research (31, 32). Although the use of orthogonal assays is already well established in experimental protocols for binary protein interactions such as yeast two-hybrid (33, 34), it is less widespread in the analysis of protein purification experiments. Our novel method ISA aims at being a generally applicable scoring scheme for inferring physical protein contacts from repeated protein complex purification experiments.

Although the ISA method performs consistently well in the assessment against experimentally determined physical contacts, its performance difference to the original SA method appears to be only minor. However, one of the main improvements of our novel method ISA is the enhanced null model that takes full advantage of additional evidence contained in repeated observations, depends on the presence of repeated purifications of the same bait protein in the experimental data. A closer analysis of the purifications in the Large-Scale data reveals that proteins were used as baits at a median number of only one time (see bait frequency distribution of nonchaperone proteins in Fig. 4*A*). Additionally, although both the Gavin and the Krogan experiments were performed on a genomic scale, only 1102 of the overall 2,830 distinct bait proteins in the Large-Scale data set were used as baits in both experiments. Therefore, the combination of the two experimental data sets resulted in relatively few repeated purifications.

In order to demonstrate the ability of ISA in utilizing repeated purifications to infer physical protein contacts with high confidence, we focused on two especially challenging purification data sets with high biological relevance that contain repeated experiments. The first analysis concentrates on inferring stable physical contacts involving *molecular chaperones* from the Large-Scale data, whereas the second analysis aims at identifying specific interactions concerning *protein kinases* and *phosphatases* from a recently published purification data set.

*Detecting Stable Interaction Partners of Molecular Chaperones*—As shown in the previous section, repeated purifications are not available for most bait proteins in the Large-Scale data. Fortunately, however, a set of 63 known molecular chaperones in yeast were screened intentionally multiple times by the Krogan group to provide experimental data for a later study (35). Molecular chaperones are a family of proteins that assist in folding of protein complexes and are therefore expected to appear in purifications with their substrate complexes. Indeed, the high abundance of this functional class of proteins results in their co-purification as preys at a median rate five times higher than non-chaperone proteins. As a consequence of selective screening in the Large-Scale purifications, chaperones were used as baits twice as often as nonchaperone proteins, resulting in a strong bias of repeatedly purified proteins in the Large-Scale experimental data toward molecular chaperones. (Compare the median number of bait and prey occurrences for chaperone and non-chaperone proteins in Fig. 4*A*).

Because of their high abundance, for some chaperones, numerous, highly transient interactions with substrates obscure more permanent interactions with co-chaperones and other regulatory proteins. This makes detection of stable interactions a difficult task for any scoring method. In fact, in order to succeed in this task, a scoring method must take full advantage of repeated purifications with chaperone bait proteins contained in the experimental data. As such, the overrepresentation of molecular chaperones in the data provides an ideal opportunity to examine the ability of scoring methods to use repeated observations in the Large-Scale purification data. To this end, we assess the performance of scoring methods in identifying stable physical contacts involving molecular chaperones from two different perspectives. First, we assess how the scoring methods perform in recovering direct physical contacts involving molecular chaperones that are confirmed by binary experimental assays. Second, we present a high-confidence network of inferred physical contacts involving molecular chaperones and investigate how well stable contacts between chaperones and their cofactors are recovered by the ISA method.

We investigated the performance of the five scoring schemes in recovering experimentally validated physical contacts involving chaperones by comparing the inferred contacts of the scoring methods to the Chaperone reference set (see Experimental procedures section for a description of the reference data). As can be seen in Fig. 4*B*, the ISA method excels in this validation and recovers 80% more physical contacts from the reference set than the SA approach at the previously determined high-confidence rank cutoff of 3,000 inferred physical contacts. Importantly, the ISA method is the only approach that demonstrates a performance higher than the best-performing binary experimental assay Y2H at the same cutoff.

Importantly, ISA does not simply promote interaction partners that have been screened repeatedly. Such an undifferentiated strategy would lead to many falsely inferred physical contacts because the highly abundant chaperones are involved in many interactions, of which only very few are likely to be reliable physical contacts. On the contrary, only 79 of the top 3,000 physical contacts inferred by the ISA method involve a molecular chaperone, much fewer than the upper limit of 7,315 spoke observations that pertain to chaperones and are present in the Large-Scale experimental data. Of these 79 physical contacts, more than 20 are validated by the reference set as shown in Fig. 4*B*. This indicates that the ability of the ISA method to make use of repeated observations in the data results in very selective promotion of true physical contacts that cannot be discovered by established scoring methods such as SA.

*Analysis of Inferred Chaperone Interactions*—To obtain a more detailed view on the relationships between molecular chaperones and their cofactors, we generated an interaction network induced by the top 3,000 physical contacts as inferred by the ISA method (see Fig. 1*A*). From this network, we extracted all physical contacts that involve at least one molecular chaperone. The resulting interaction network is displayed in Fig. 5. It contains 79 inferred physical contacts involving 31 of the 63 known yeast chaperones as well as their cofactors and putative substrates.

As can be seen in Fig. 5, physical contacts inferred by the ISA method form a sparse network. In addition, known protein assemblies, such as the RAC or Sec63 complexes, are connected by patterns of physical contacts and several fine-grained biological relationships between chaperone families are correctly recovered.

The Gimc complex, for instance, a hetero-oligomeric hexamer stabilizing nonnative proteins, consists of a dimeric core consisting of $\alpha$-subunits Gim2 and Gim5. The $\alpha$-subunits form physical contacts with each other as well as with two of four possible $\beta$-subunits (Gim1,3,4 and 6) each (36). Although the ISA score correctly identifies the physical contact between the $\alpha$-subunits by assigning it the top rank among all inferred physical contacts of that complex, it also infers contacts between the $\alpha$-subunits and each of the four possible $\beta$-subunits at a slightly lower confidence.

In addition, more intricate relationships, such as patterns of interactions between different families of chaperones, are correctly formed by the inferred physical contacts. Hsp110 homologs Sse1 and Sse2, for instance, form mutually exclusive, hetero-dimeric complexes with Hsp70 families SSA and SSB of the form SSA · SSE and SSB · SSE (37, 38). This relationship is correctly recovered by inferred physical contacts as depicted in Fig. 5, where SSA · SSE and SSB · SSE interactions are partitioned and, correctly, no physical contacts between SSA and SSB chaperones are inferred.

Importantly, the presented network of chaperone interactions is unique among scoring methods. It is inherently difficult to infer physical contacts involving chaperones from purification data. This is because of the high abundance of this

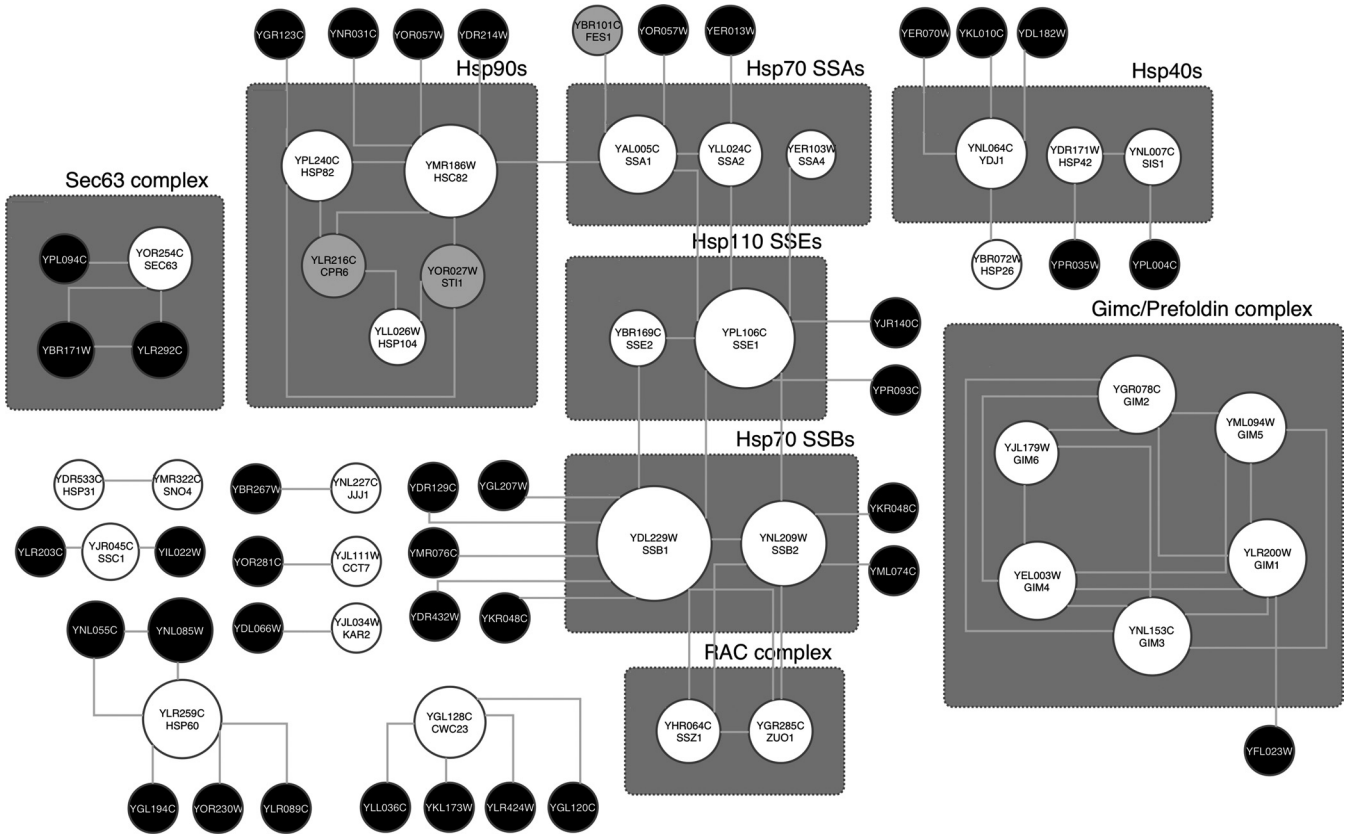
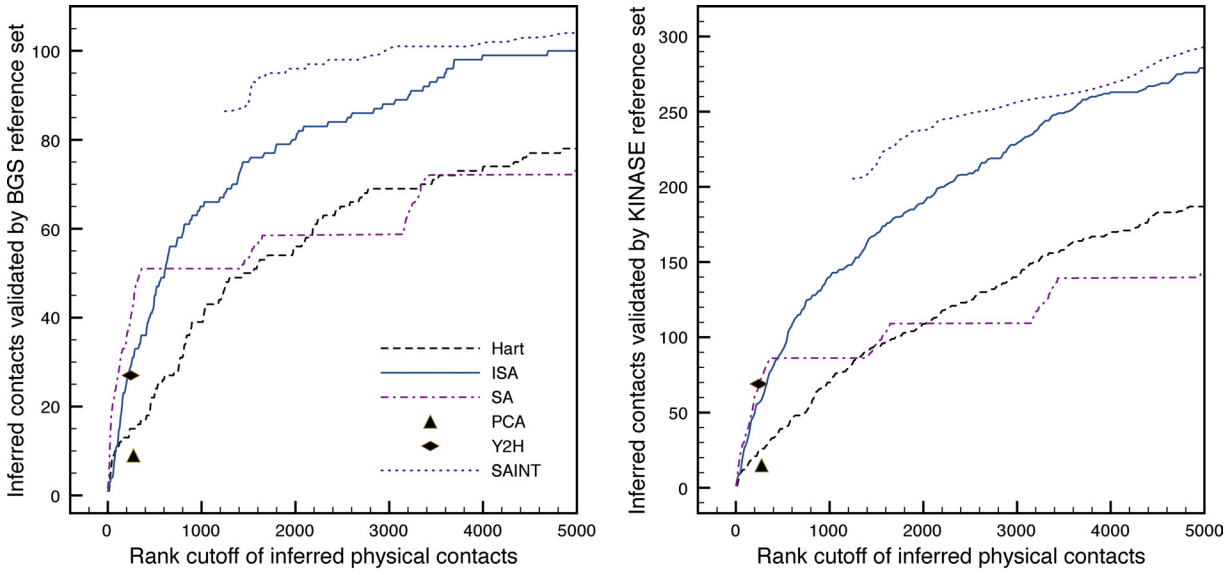



FIG. 5. **All physical contacts involving molecular yeast chaperones extracted from the overall top 3000 physical contacts as inferred by the ISA score.** Nodes and edges denote proteins present in the Large-Scale data set and their inferred physical contacts, respectively. The size of a node corresponds to its degree, that is, the number of physical contacts it is involved with. Chaperones are colored in white, whereas proteins with known chaperone-related function, such as cochaperones, are colored in gray. Black nodes denote putative substrates of chaperones. Proteins belonging to known families or assemblies are grouped in gray rectangles.

functional class of proteins that leads to a low signal-to-noise ratio for physical contacts that involve chaperones. As a result, established scoring schemes like the SA method tend to uniformly rank down chaperone interactions. Indeed, the SA method is unable to recover known biological relationships involving chaperones as shown in Fig. 5 even among its top-10,000 inferred physical contacts.

*Identifying Specific Interactions of Protein Kinases and Phosphatases*—The network of kinase and phosphatase interactions is an important component of cellular regulation and messaging. Therefore, these enzymes are highly relevant for understanding a wealth of cellular processes that are influenced by kinase signaling. Although experimentally validated kinase and phosphatase interactions have been available only sparsely in public databases, a specialized large-scale purification screen of yeast kinases and phosphatases has recently been published by Breitkreutz *et al.* (18). Protein kinases are a challenging target for scoring schemes because kinases have a propensity toward binding to a large number of other proteins and it is therefore difficult to separate specific from nonspecific interactions (18). This is illustrated by the average purification size of the Breitkreutz purification data: it is more than twice as high as the corresponding sizes in the Large-Scale data (compare Table I and supplemental Table S1). One of the main contributions of the Breitkreutz *et al.* study thus is the introduction of SAINT, a computational method that identifies nonspecific kinase interactors. SAINT and closely related methods such as COMPASS (39) primarily rely on peptide (spectral) counts, which constitute additional types of experimental data that can be interpreted as a semi-quantitative measure of protein abundance. To further increase the coverage and experimental confidence of their method, Breitkreutz *et al.* opted to perform their screen with three different tag systems, yielding multiple overlapping purifications with the same bait proteins. It is because of this intentional application of repeated purifications that we found the Breitkreutz data set especially suited for our ISA method.

We assessed the performance of the scoring methods SAINT, Hart, ISA, and SA in inferring experimentally known kinase interactions from the BGS and Kinase reference sets (see Experimental procedures section for a description of the reference data). Note that, because of the involved computations or unavailable implementations of the PE and IDBOS methods, these scores could not be evaluated here. How-




(A) BGS reference set (B) KINASE reference set

FIG. 6. **Assessment of inferred and experimentally obtained physical contacts involving protein kinases and phosphatases against the BGS and Kinase reference set of experimentally confirmed binary kinase and phosphatase interactions.** Note that the SAINT scoring scheme assigns identical score values for its top 1262 inferred interactions. Therefore, SAINT performance curve seems to start at a later point than the curves of other methods.

ever, it has been reported elsewhere that PE was not able to distinguish between true and false interactions in this setting (18). As displayed in Fig. 6, the ISA score outperforms other general-purpose purification scoring schemes on both reference sets by a large margin. Only the highly specialized SAINT scoring scheme can identify slightly more physical contacts in the data. Importantly, however, the peptide counts employed as an integral part of SAINT require additional processing during the experimental setup. Such counts are neither available for the Large-Scale purifications nor for most other publicly available purification data. In contrast, the ISA scoring scheme is generally applicable to all raw purification data without the need for additional peptide count data.

## DISCUSSION

This work is first to investigate whether direct physical protein contacts can be extracted from raw purification data contained in a combined set of large-scale protein complex purifications. We analyzed four established scoring schemes and one new approach and assessed their ability to reliably detect physical contacts within assayed complexes. Top ranking inferred physical contacts from all five methods were benchmarked against reference sets based on binary experimental protein interactions, three-dimensional structures of interacting proteins, manually curated protein complexes, and genetic interaction profiles. Inclusion of these four complementary sources of validated physical contacts allowed us to investigate aspects of the scoring schemes that were not examined before.

The results of our evaluation showed that raw purification data, if scored correctly, can indeed be exploited to infer physical contacts within protein complexes. Although established methods devised for inferring indirect protein interactions from purification data perform well in the task of identifying co-complexed protein pairs in reference protein complexes (see supplemental Fig. S1 and supplemental material), the performance of most of these methods in detecting direct physical protein contacts is considerably diminished. Only the two socio-affinity based methods SA and ISA consistently showed the best performance among all evaluated methods in inferring physical contacts.

We attribute this difference of performance between the methods to two facts. First, both SA and ISA share the concept of a simple, but elegant, null model to identify strongly associated proteins. Second, both methods have a high propensity toward using only direct, or spoke, associations between proteins as evidence for physical contacts, that is, they concentrate on interaction evidence between a bait protein and the prey proteins it purifies. This indicates that, whereas indirect, or matrix, observations are useful for detecting co-complexing protein pairs, direct observations are more informative for identifying physical contacts.

Besides intentionally concentrating on direct observations, the main innovation of the ISA method is an improved null model that allows integration of repeated purifications using the same set of bait proteins in a statistically meaningful fashion. Although improving the predictive power of our method in the presence of any repeated observations, this ability is especially relevant for inferring stable contacts

involving highly abundant proteins, such as molecular chaperones or protein kinases, whose specific interactions are especially difficult to infer in the absence of repetitions. Intuitively, the approach taken by our ISA score is similar to strategies currently discussed for interactome mapping projects where repeated experiments can allow an increased sensitivity and specificity of the resulting screens (34).

Our analysis of the Large-Scale purification data found that most proteins have been screened only once in experiments. Few proteins, such as the functional class of molecular chaperones, have been used as baits multiple times. The assessment of chaperone interactions showed that the ISA method improves upon other scoring methods when repeated observations are available. Our method recovered a range of biologically significant relationships between chaperones and their cofactors that could not be detected by the second best performing SA method. This highlights the importance of correctly using repeated observations to gain statistical confidence in the inferred physical contacts.

The use of repeated observations for gaining statistical confidence in physical contacts is not limited to the general-purpose large-scale purification experiments, but can also be applied to specialized data sets aiming at specific biological targets as demonstrated by our analysis of a recent purification study focusing on protein kinases. Because of the intentional integration of repeated purifications in that study, ISA was available to infer significantly more physical contacts from the raw purifications than any other general-purpose scoring method.

Importantly, the mechanism of improved performance of ISA based on repeated observations is not depending on single purification data sets that feature repeated purifications. Instead, ISA is able to exploit repetitions across several distinct data sets. Therefore, we expect that physical contacts inferred by our method will further improve in quality compared with established scoring schemes once additional large-scale purification data sets become available. Considering the experimental replicates within the Breitkreutz *et al.* (18) data as well as current correspondences by experimentalists about integrating multiple orthogonal assays to increase confidence in the results (31, 32), there seem to be clear indications that repeated purifications within one data set will become more widespread in future.

An adequate comparison of interactions measured by high-throughput binary experimental approaches to physical contacts deduced from purification experiments has not been possible before. This is a result of the fact that binary experimental approaches are more straightforward to interpret: each physical interaction is measured directly, whereas in purification data only a small subset of all interactions are likely to be physical contacts. As a consequence, the whole set of interactions possible in purification data has previously been interpreted as putative physical contacts. Because this resulted in a low enrichment of true physical contacts in the purification data, purification-based methodologies were determined of being less useful for measuring true physical contacts than high-throughput binary experimental techniques such as yeast two-hybrid (13).

However, a closer analysis of the performance of all five scoring schemes and a comparative assessment of the inferred physical contacts with the results of binary experimental assays such as Y2H or PCA reveals several novel findings. First, the performance of all scoring schemes in recovering physical contacts from the reference sets levels off at about 3000 top-ranking physical contacts, indicating that this number constitutes a limit on the number of interactions that can be reliably scored given current experimental data. Second, our results surprisingly suggest that, once correctly scored and ranked, physical protein contacts derived from complex purification experiments are qualitatively comparable to interactions measured by state-of-the-art Y2H and PCA techniques. In addition, the purification scoring schemes perform significantly better than the Y2H and PCA data sets in connecting manually curated protein complexes. This suggests that physical contacts derived from purification data might be more relevant for interpreting protein complexes than interactions measured by these binary experimental techniques.

Besides offering new opportunities for the interpretation of purification data and the understanding of protein complexes, there are additional application scenarios of our scoring method. Because the ISA method is optimized to make best use of repeated observations in the data, experimental research groups can repeatedly perform small-scale purifications for proteins of interest, possibly involving perturbation experiments (16) or novel purification methodologies applicable to human cells (40). These purifications can then be added to the Large-Scale experimental data. Subsequently, the ISA method can be re-applied on this newly enlarged data set. The additional information contained in the repeated purifications will then allow reliable inference of physical contacts involving the proteins of interest.

We further propose that physical contacts derived from purification data are applicable to large-scale interactome mapping projects. Several interactome screens are currently underway for model species such as *S. cerevisiae* or *D. melanogaster*. Meta-strategies for cost-effective mapping have been developed involving schemes for pooling, prioritization, and repetition of experiments to increase overall coverage and accuracy of the combined screens (33, 34). We suggest that physical contacts derived from purification data may be used complementary to binary interaction data sets by prioritizing high-confidence physical contacts for experimental validation by Y2H or PCA techniques. Such a prioritization strategy seems especially valuable when integrated in high-throughput large-scale interactome screening projects such as recently proposed by Schwartz *et al.* (34).

*Acknowledgments*—The work was conducted in the context of the DFG-funded Cluster of Excellence for Multimodal Computing and Interaction.

* Part of this study was financially supported by the German National Genome Research Network (NGFN) and by the German Research Foundation (DFG), contract number KFO 129/1-2.

§ To whom correspondence should be addressed: Max Planck Institute for Informatics, Campus E1.4, 66123 Saarbrücken, Germany. E-mail: ezotenko@mpi-inf.mpg.de; sven@mpi-inf.mpg.de.

‖ Both are joint senior authors.

REFERENCES

1. Gavin, A. C., and Superti-Furga, G. (2003) Protein complexes and proteome organization from yeast to man. *Curr. Opinion Chem. Biol.* **7,** 21–27
2. Robinson, C. V., Sali, A., and Baumeister, W. (2007) The molecular sociology of the cell. *Nature* **450,** 973–982
3. Dziembowski, A., and Séraphin, B. (2004) Recent developments in the analysis of protein complexes. *FEBS Lett*. **556,** 1–6
4. Collins, M. O., and Choudhary, J. S. (2008) Mapping multiprotein complexes by affinity purification and mass spectrometry. *Curr. Opin. Biotechnol.* **19,** 324–330
5. 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
6. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, 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. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440,** 631–636
7. 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. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature* **440,** 637–643
8. Collins, S. R., Kemmeren, P., Zhao, X. C., Greenblatt, J. F., Spencer, F., Holstege, F. C. P., 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
9. 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
10. 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
11. Friedel, C. C., Krumsiek, J., and Zimmer, R. (2009) Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. *J. Comput. Biol.* **16,** 971–987
12. Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2009) Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* **37,** 825–831
13. 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., Barabási, 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
14. 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
15. Aloy, P., Böttcher, 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
16. Hernandez, H., Dziembowski, A., Taverner, T., Seraphin, B., and Robinson, C. V. (2006) Subunit architecture of multimeric complexes isolated directly from cells. *EMBO Reports* **7,** 605–610
17. Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Serna 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
18. Breitkreutz, A., Choi, H., Sharom, J. R., Boucher, L., Neduva, V., Larsen, B., Lin, Z. Y., Breitkreutz, B. J., Stark, C., Liu, G., Ahn, J., Dewar-Darch, D., Reguly, T., Tang, X., Almeida, R., Qin, Z. S., Pawson, T., Gingras, A. C., Nesvizhskii, A. I., and Tyers, M. (2010) A global protein kinase and phosphatase interaction network in yeast. *Science* **328,** 1043–1046
19. Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., Park, J., Binkley, G., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Krieger, C. J., Livstone, M. S., Miyasato, S. R., Nash, R. S., Oughtred, R., Skrzypek, M. S., Weng, S., Wong, E. D., Zhu, K. K., Dolinski, K., Botstein, D., and Cherry, J. M. (2008) Gene ontology annotations at sgd: new data sources and annotation methods. *Nucleic Acids Res.* **36**(Database issue): D577–581
20. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuer-mann, 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. Intact-open source resource for molecular interaction data. *Nucleic Acids Res.* **35**(Database issue), D561–565
21. Guldener, U., Münsterkötter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H. W., and Stümpflen, V. (2006) Mpact: the mips protein interaction resource on yeast. *Nucleic Acids Res.* **34**(Database issue), D436–441
22. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R. R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R. D., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P. S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Orengo, C., Petryszak, R., Selengut, J. D., Sigrist, C. J., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2007) New developments in the interpro database. *Nucleic Acids Res.* **35**(Database issue), D224–228
23. Aloy, P., and Russell, R. B. (2006) Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.* **7,** 188–197
24. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The protein data bank. *Nucleic Acids Res.* **28,** 235–242
25. Stein, A., Russell, R. B., and Aloy, P. (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.* **33**(Database issue), D413–417
26. Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L. Y., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Z. Y., Lin, Liang, W., Marback, M., Paw, J., San Luis, B. J., Shuteriqi, E., Tong, A. H. Y., van Dyk, N., Wallace, I. M., Whitney, J. A., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. A., Brudno, M., Ragibizadeh, S., Papp, B., Pál, C., Roth, F. P., Giaever, G., Nislow, C. Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A. C., Morris, Q. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J., and Boone, C. (2010) The genetic landscape of a cell. *Science* **327,** 425–431
27. Yu, X., Ivanic, J., Wallqvist, A., and Reifman, J. (2009) A novel scoring approach for protein co-purification data reveals high interaction specificity. *PLoS Comput. Biol.* **5,** e1000515
28. Salama, I., and Quade, D. (1982) A nonparametric comparison of two multiple regressions by means of a weighted measure of correlation. *Commun. Statistics* **11,** 1185–1195
29. Mosca, R., Pons, C., Fernández-Recio, J., and Aloy, P. (2009) Pushing

structural information into the yeast interactome by high-throughput protein docking experiments. PLoS Comput. Biol. **5,** e1000490
30. Stein, A., Panjkovich, A., and Aloy, P. (2009) 3DID update: domain-domain and peptide-mediated interactions of known 3d structure. *Nucleic Acids Res*. **37**(Database issue), D300–304
31. Chen, Y. C., Rajagopala, S. V., Stellberger, T., and Uetz, P. (2010) Exhaustive benchmarking of the yeast two-hybrid system. *Nat. Methods* **7,** 667–668
32. Braun, P., Tasan, M., Cusick, M., Hill, D. E., and Vidal, M. (2010) Reply to "exhaustive benchmarking of the yeast two-hybrid system". *Nat. Methods* **7,** 668
33. Lappe M., and Holm, L. Unraveling protein interaction networks with near-optimal efficiency. *Nat. Biotechnol.* **22,** 98–103
34. Schwartz, A. S., Yu, J., Gardenour, K. R., Finley, R. L., and Ideker, T. (2009) Cost-effective strategies for completing the interactome. *Nat. Methods* **6,** 55–61
35. Gong, Y., Kakihara, Y., Krogan, N., Greenblatt, J., Emili, A., Zhang, Z., and Houry, W. A. (2009) An atlas of chaperone-protein interactions in saccharomyces cerevisiae: implications to protein folding pathways in the cell. *Mol. Syst. Biol.* **5,** 275
36. Leroux, M. R., Fändrich, M., Klunker, D., Siegers, K., Lupas, A. N., Brown, J. R., Schiebel, E., Dob-son, C. M., and Hartl, F. U. (1999) Mtgimc, a novel archaeal chaperone related to the eukaryotic chaperonin cofactor gimc/prefoldin. *EMBO J.* **18,** 6730–6743
37. Shaner, L., Wegele, H., Buchner, J., and Morano, K. A. (2005) The yeast hsp110 sse1 functionally interacts with the hsp70 chaperones ssa and ssb. *J. Biol. Chem.* **280,** 41262–41269
38. Yam, Y. W. A., Albanèse, V., Lin, H. T. J., and Frydman, J. (2005) Hsp110 cooperates with different cytosolic hsp70 systems in a pathway for de novo folding. *J. Biol. Chem.* **280,** 41252–41261
39. Sowa, M. E., Bennett, E. J., Gygi, S. P., and Harper, J. W. (2009) Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138,** 389–403
40. Malovannaya, A., Li, Y., Bulynko, Y., Jung, S.Y., Wang, Y., Lanz, R. B., O'Malley, B. W., and Qin, J. (2010) Streamlined analysis schema for high-throughput identification of endogenous protein complexes. *Proc. Natl. Acad. Sci. U.S.A*. **107,** 2431–2436