RESEARCH ARTICLE

WI-PHI: A weighted yeast interactome enriched for direct physical interactions

Lars Kiemer[1], *Stefano Costa*[1], *Marius Ueffing*[2] *and Gianni Cesareni*[1]

[1] Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica, Rome, Italy
[2] Institute of Human Genetics, GSF – National Research Center for Environment and Health, Munich, Germany

How is the yeast proteome wired? This important question, central in yeast systems biology, remains unanswered in spite of the abundance of protein interaction data from high-throughput experiments. Unfortunately, these large-scale studies show striking discrepancies in their results and coverage such that biologists scrutinizing the "interactome" are often confounded by a mix of established physical interactions, functional associations, and experimental artifacts. This stimulated early attempts to integrate the available information and produce a list of protein interactions ranked according to an estimated functional reliability. The recent publication of the results of two large protein interaction experiments and the completion of a comprehensive literature curation effort has more than doubled the available information on the wiring of the yeast proteome. This motivates a fresh approach to the compilation of a yeast interactome based purely on evidence of physical interaction. We present a procedure exploiting both heuristic and probabilistic strategies to draft the yeast interactome taking advantage of various heterogeneous data sources: application of tandem affinity purification coupled to MS (TAP-MS), large-scale yeast two-hybrid studies, and results of small-scale experiments stored in dedicated databases. The end result is WI-PHI, a weighted network encompassing a large majority of yeast proteins.

Received: June 20, 2006
Revised: September 1, 2006
Accepted: September 14, 2006

Keywords:
Interaction network / Interactome / Protein complexes / Protein interaction / Yeast

1 Introduction

Modeling cell physiology requires a thorough and quantitative understanding of the molecular interaction mesh in a living cell. However, despite technological progress and high-throughput approaches, we are very far from a satisfactory description of the equilibrium and kinetic constants governing the interactions between proteins, proteins and metabolites, or proteins and nucleic acids. Nevertheless, a couple of recent reports on genome-wide experiments aimed at describing the complete set of interactions occurring in the yeast *Saccharomyces cerevisiae* [1, 2], combined with a comprehensive curation of the interaction data published in the scientific literature [3] have provided us with the best description ever of the protein interaction network in a living cell.

More than 50 000 interactions between yeast proteins have been described in the literature, and a large fraction of those are now deposited in protein interaction databases such as BioGRID [4], MINT [5], MPact [6], IntAct [7], and DIP [8]. It is not uncommon that a database search for ligands of any of the approximately 6000 yeast proteins returns a large number of putative interactors, ranging from a few to hundreds. However, these search outputs are often difficult to interpret because interactions of biological significance are intermixed with false positives. On top of that, direct physical interactions are mixed with indirect or

Correspondence: Professor Gianni Cesareni, Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica, Rome, Italy
E-mail: cesareni@uniroma2.it
Fax: +39 0672594599

Abbreviations: APC, anaphase promoting complex; **GO,** gene ontology; **LLS,** log-likelihood score; **MIPS,** Munich Information Center for Protein Sequences; **SA index,** socioaffinity index; **TAP,** tandem affinity purification functional associations, and interactions leading to the formation of stable complexes are mixed with transient interactions characterized by high dissociation constants. The main reasons for this heterogeneity are diversity of the experimental approaches and a high percentage of false positives in high-throughput experiments. For instance, the experimental procedure required for affinity purification allows the discovery of proteins that are stably associated with the baits, but it does not permit the detection of the binary direct interactions responsible for the precise topology of the complex. In contrast, the positive hits of a two-hybrid approach are enriched for direct interactions, and transient interactions (dissociation constants in the 10 μM range) are not uncommon. On the other hand, interactions discovered by the yeast two-hybrid approach must be confirmed by further experimental evidence because of the inherently high false-positive rate and nonphysiological experimental settings.

Despite the large number of interactions discovered in high-throughput experiments, the estimated "interactome" coverage of each experiment is still rather low. This is best illustrated by the analysis of the overlap of the results of the two recently reported high-throughput affinity purification experiments [1, 2]. Both studies used the tandem affinity purification (TAP) technology [9] to identify the proteins copurifying with most of the yeast ORFs, and applied a statistical analysis to their raw data to end up with two high confidence interactome cores of similar size (approximately 7000 interactions). However, when the two resulting datasets are compared, only approximately one-fourth of the interactions are found to be shared by both datasets (Fig. 1). Since false positives have been pruned away, the most likely conclusion is that both datasets are largely incomplete.

It has been shown, however, that by combining the evidence from different experimental approaches, it is possible to substantially increase the fraction of biologically relevant interactions [10–12]. Using either heuristic approaches or more or less sophisticated probabilistic frameworks for data integration, a number of groups have proposed trustworthy yeast protein networks. Most of these, however, combine the information on physical interaction with contextual evidence based on gene coexpression, protein colocalization, and genetic interaction [12–15]. The resulting networks provide strong evidence of functional correlation but do not necessarily reflect true, direct interactions between proteins, and are biased by a high false-positive rate when used as evidence of direct physical interactions.

We have been motivated by the recent deluge of fresh protein interaction information [1–3] to devise an approach integrating the available information to compile a list of protein interactions ranked according to a score, which weighs functional reliability and evidence of direct interaction. Although filtering, by taking into account contextual evidence and genetic interactions, has been shown to improve the functional reliability of the network [14, 16, 17], we deliberately chose to focus on direct physical interactions without contaminating our network with mere functional relationships. Contextual filters can be added to our network if desirable for any specific purpose.

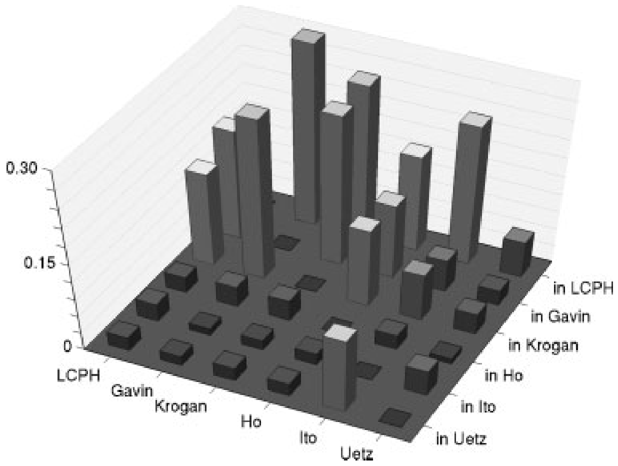

Figure 1. Comparison of different protein interaction datasets. The vertical bars represent the fraction of one dataset that is present in another one. For clarity, the diagonal is set to 0, although it should be 1. For definition of datasets see Section 2.

In the absence of accepted positive and negative standards, a purely probabilistic approach cannot be undertaken in a straightforward way. For lack of alternatives, however, we used a probabilistic approach, complemented with some heuristic intervention to adjust scores favoring interactions supported by low-throughput studies. Low-throughput studies are of higher quality because of the manual curation and labor-intensive postprocessing of the results not feasible in high-throughput experiments.

The end result is a weighted network of protein interactions with strong emphasis on direct, physical interactions. Such an interactome is a valuable contribution both for detailed functional analysis and for the field of interactome systems biology in yeast.

2 Materials and methods

2.1 Datasets

Seven datasets were assembled for analysis: five high-throughput datasets, an intermediate-scale dataset, and a low-throughput dataset. Data originating from publications reporting interactions between ten or fewer proteins were considered low-throughput, and data not conforming to this category but not being high-throughput either (arbitrarily fixed at more than 1000 interactions) were considered to be of intermediate scale. Two groups have published more than one high-throughput dataset [1, 18–20]. In this case, we considered only the most recent sets, as these are assumed to include also the data from the first smaller-scale report.

For both TAP studies [1, 2], raw data are available as Supporting Information. For yeast two-hybrid data published by Ito *et al.* [18] the full Interaction Sequence Tag (IST) set was used to obtain unfiltered data. The Uetz dataset [21] was downloaded from BioGRID [4]. The literature-curated physical interaction (LCPH) dataset was obtained by removing all the above datasets from a combination of the data downloaded from BioGRID 2.0.19, and an interrogation of MINT [5], BIND [22], IntAct [7], and MPact [6] databases. Additionally, we removed interactions supported by indirect evidence such as genetic interactions (including "synthetic lethality", "synthetic growth defect", "synthetic rescue", "phenotypic enhancement", "epistatic miniarray profile", "dosage rescue", and "phenotypic suppression") and colocalization. Interactions inferred from informatic approaches were not considered. This literature-curated dataset "LCPH" was further subdivided into low (LCPH-LS) and intermediate scale (LCPH-IS), composed of interactions discovered in experiments reporting interactions between ten or fewer proteins and experiments yielding up to 1000 interactions, respectively.

The interactomes or sets of interactions used for network analysis are: *Krogan core* [2], "filtered yeast interactome" (FYI) [13], and *Gavin core* defined as containing interactions having a socioaffinity index (SA index) higher than 5, a value proposed by the authors as a confidence threshold [1].

2.2 Generation of a benchmark set for large-scale datasets

In order to assemble a benchmark for the evaluation of high-throughput datasets, we collected the 1777 interactions supported by two independent methods in the small-scale dataset described in the previous section (LCPH-LS). This set of interactions is available as Supporting Information. An additional benchmark set based on structural information was compiled from iPfam [23]. This set alone was too small (95 interactions) to be used to benchmark the high-throughput datasets, but still it represents an important addition to the interactome (being the only dataset with irrefutable evidence for direct physical interaction).

2.3 Calculation of SA indices and dataset weights

Applying the SA index as defined by Gavin *et al.* [1] to all the datasets mentioned above, except the low-throughput dataset, allowed us to integrate the data irrespective of the experimental method.

SA indices were modified by multiplication with a weight constant depending on the accuracy of the respective set evaluated on the benchmark set mentioned above. Since the benchmark set does not describe noninteractions, evaluation was carried out by counting as noninteractors (negative set) all possible interactions between the benchmark set minus the known interactions. This allowed us to calculate a log-likelihood score (LLS) reflecting how well the individual datasets describe the benchmark set:

$$\mathrm{LLS} = \ln\left(\frac{\mathrm{TP}/\mathrm{FP}}{\mathrm{pos}/\mathrm{neg}}\right)$$

where, TP and FP are the true positive and false-positive rates of the dataset in question on covering the benchmark set, and pos and neg are the fractions of positives and negatives, respectively, in the benchmark set.

The huge differences in size between the considered datasets and the lack of a reliable benchmark of noninteracting proteins prevent a calculation of the score from the entire set. Thus, only interactions obtaining an SA score above a certain threshold were used for benchmarking. The threshold chosen was the same as that reported by Gavin *et al.* [1] as the lower boundary for trustworthy interactions. The following log likelihoods were calculated and multiplied by the SA indices to produce the final score: Gavin 3.66, Krogan 2.38, Ho 3.59, Ito 3.03, Uetz 3.95, and intermediate scale 4.64. It should be noted that two datasets are available describing the Krogan data as the prey detection step was performed with two different techniques: for interactions described by both sets, we have chosen to consider them as only one contribution and combine the scores from the two sets according to their individual performance on the benchmark set.

Following this step, the resulting scores were summed for each interaction to produce a final combined interaction score. The benchmark set, the low-throughput experiments, and the structurally based datasets were assigned high scores (20, 20, and 40, respectively) – arbitrarily chosen to make sure these high confidence interactions are included among the highest scoring – and added to the network.

2.4 Interactome validation using gene ontology (GO) and expression profiles

The GO classifications associated with each ORF were downloaded from SGD [24], and the Biological Process term was extracted. A Bayesian statistics approach adapted from Lee *et al.* [14] was used to validate the interactomes independently. This approach produces an LLS reflecting the degree of enrichment of identical GO terms between the partners in an interaction. The 50 000 best scoring interactions of WI-PHI were used to calculate prior probabilities. These 50 000 interactions cover 5951 ORFs of which more than 90% could be mapped to 569 different GO biological process terms.

The starting point of the expression analysis was the dataset provided in a large-scale study of gene expression [25]. We extracted yeast data and calculated the Pearson correlation coefficient between all pairs of genes. As the data were only used to validate the WI-PHI approach, a correlation coefficient above 0.5 was considered an indication of coexpression.

3 Results

This study consists of four main parts: (i) collecting protein–protein interaction datasets, (ii) ranking the interactions within each dataset according to a common scoring method, (iii) integrating the evidence from the different datasets by weighting the individual scores with a coefficient reflecting the performance of each dataset in reproducing the results of a high confidence interactome used as benchmark, and (iv) validation and detailed examination of the interactome. The result is WI-PHI, a weighted yeast interactome of 50 000 interactions that can be filtered according to any desired confidence threshold.

3.1 Interaction datasets

Our network assembly approach takes advantage of the published large-scale interaction datasets and a compilation of the major protein–protein interaction databases. Two datasets are the results of large-scale experimental efforts that made use of TAP technology [9] combined with affinity purification and sensitive MS analysis of the copurifying prey proteins [1, 2]. A third set originates from the MS-based identification of immunoaffinity purified, overexpressed, tagged yeast proteins [26]. A large collection of interactions became available recently, thanks to a formidable curation effort compiling more than 30 000 yeast genetic and physical interactions described in the scientific publications [3]. This dataset, downloaded from the BioGRID website, was combined with our own compilation of the protein–protein interaction databases BIND [22], MINT [5], MPact [6], IntAct [7], and DIP [8] and filtered manually to remove high-throughput experiments already covered, genetic interactions, and colocalization evidence. The resulting dataset (LCPH) is still rather heterogeneous, since it includes published data describing interactions ranging in number from one to as many as thousands in high-throughput studies. Since experiments reporting a large number of interactions are likely to be less reliable than experiments focusing on few interactions supported by different experimental evidence, we further subdivided the LCPH dataset into four groups. The first two include the results of two high-throughput yeast two hybrid (Y2H) studies, from now on dubbed Ito [18] and Uetz [21]. The remaining interactions were arbitrarily subdivided further into two groups: LCPH-LS (low-throughput scale) consisting of interactions from studies mentioning ten or fewer proteins and LCPH-IS (intermediate-scale studies) being interactions from the remaining publications. A small number of interactions between yeast proteins have also been characterized by X-ray crystallography, and these were collected in a separate dataset.

3.2 Interaction rating

As a first step, we conceived a strategy to rate the interactions within each dataset. High-throughput datasets lend themselves to scoring schemes, since statistical analysis of interactions observed at different frequencies in experimental repetitions can be readily applied. Both Gavin *et al.* [1] and Krogan *et al.* [2] subjected the raw experimental data to such, albeit different, analyses and, as a consequence, in both reports the end results are presented as ranked interaction lists. However, if we compare the scores of the interactions reported by both groups, the correlation is rather poor (correlation coefficient of 0.13). In order to have a uniform scoring scheme, we decided to apply to all datasets the SA index scoring system used by Gavin *et al.* [1] to postprocess their purification results.

The SA index measures the log-odds of the number of times two proteins are observed to interact, relative to the expected value as deduced from their frequency in the dataset. Since this scoring system takes into account the frequency of proteins within the dataset, it also penalizes interactions involving very promiscuous partners. Interestingly, Gavin *et al.* suggest that SA indices (weakly) correlate with dissociation constants and that interaction subsets having high SA indices are enriched for direct interactions. Although the SA index was specifically conceived to process the results of high-throughput pull-down experiments, we chose to apply this analysis not only to data originating from TAP experiments [1, 2] but also to data obtained by high-throughput yeast two-hybrid experiments [18, 21] as well as data from the intermediate-scale dataset. Our motivation was two-fold: (1) utilizing a single scoring system to rank interactions within distinct datasets in order to make them directly comparable, and (2) reducing the weight of highly promiscuous proteins in the final interactome. A comparison of calculated SA scores for the overlaps between the different datasets reveals a certain degree of correlation (Pearson correlation coefficient of 0.7 for Gavin *versus* Krogan and 0.4 for Uetz *versus* Ito or Ito *versus* Gavin, for example), supporting the notion that the SA score is dataset independent but comparable.

3.3 Interactome assembly

Next, we devised an integration scheme to combine the scores obtained by each interaction in the different datasets in order to assemble a high confidence yeast interactome. To take into account the different quality of the experimental datasets under consideration, we first measured the accuracy with which the individual datasets represented a benchmark of trustworthy interactions. Since no clearly established benchmark for protein interaction is available, we assembled a highly reliable set of 1777 interactions by combining all the interactions that are supported by more than two independent types of experiments in LCPH-LT (this benchmark set is available as Supporting Information). Then, we used the resulting LLSs to weigh the contribution of each supporting dataset to the final score. Finally, we incorporated the benchmark set and the low-throughput interaction set (LCPH-LT) (see Section 2 for the description of these sets) in the resulting interactome by assigning arbitrarily high scores to the corresponding interactions, thus ensuring that they would be included in any high confidence interactome. Although more than 700 000 putative interactions have been ranked by this approach, most of these are low scoring interactions, which are likely to be false positives of no biological significance. The WI-PHI interactome we consider in the following includes only the 50 000 best scoring interactions.

3.4 Interactome validation and characterization

The outcome of this study is a prioritized list of yeast protein interactions, where each interaction is assigned a score reflecting its support in different datasets and the promiscuity of each of the two proteins. However, the accumulated body of interaction data is enormous and many low-scoring interactions are dubious, due to the generally high false-positive rate in protein–protein interaction experiments [17]. Determination of the exact relationship between the accumulation of false positives and any chosen score threshold is not trivial. To gain further insights into the relation between score threshold and network characteristics, we have examined more closely two different subsets of the WI-PHI interactome.

(i) WI-PHI core: A high confidence interactome consisting only of interactions whose score is higher than 21. The score threshold chosen for this trustworthy interactome is such that even interactions from low-throughput studies can only be included if supported by another dataset.

(ii) WI-PHI extended: An extended interactome (interaction score higher or equal to 17), including in addition interactions with less experimental support. The threshold for the extended interactome is such that one third of the interactions supported by intermediate-scale experiments are included.

The WI-PHI core interactome is visualized with the VisANT software [27] in Fig. 2. The nodes annotated to a selection of relevant Munich Information Center for Protein

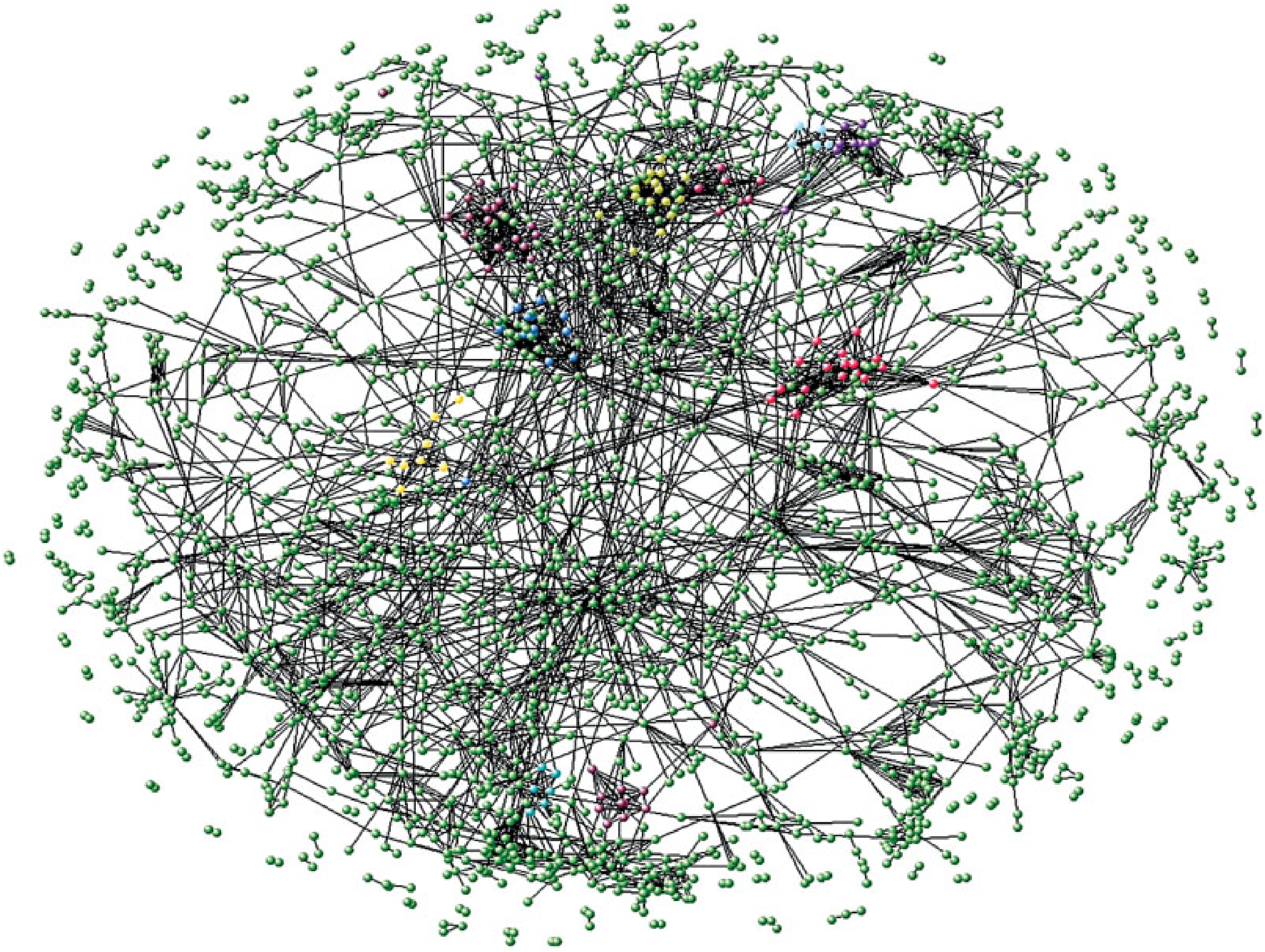

Figure 2. The WI-PHI core interactome visualized in VisANT [27]. Selected MIPS complexes are highlighted with different colors: Kornberg's mediator (SRB) complex, SAGA complex, TAFIIs, SWI/SNF transcription activator complex, RSC complex (Remodel the structure of chromatin), RNA polymerase III, 19/22S regulator, Pre-replication complex (pre-RC), Arp2p/Arp3p complex, APC.

Sequences (MIPS) complexes are highlighted in color to stress the clustering of the participant proteins in discrete, highly connected areas of the interactome. The degree distribution of the network follows a power-law, which is characteristic of scale-free biological networks (see Fig. 3). However, within the whole distribution, fewer highly connected nodes are present than expected in a purely power-law distribution. Consistent with the observation that highly connected proteins are more likely to be essential than are proteins with only a small number of links to other proteins [28], the protein products of essential genes have, in WI-PHI core, an average degree of 6.93 to be compared with 3.69 for nonessential genes. Furthermore, interactions between essential genes are three times more enriched than expected on a random base.

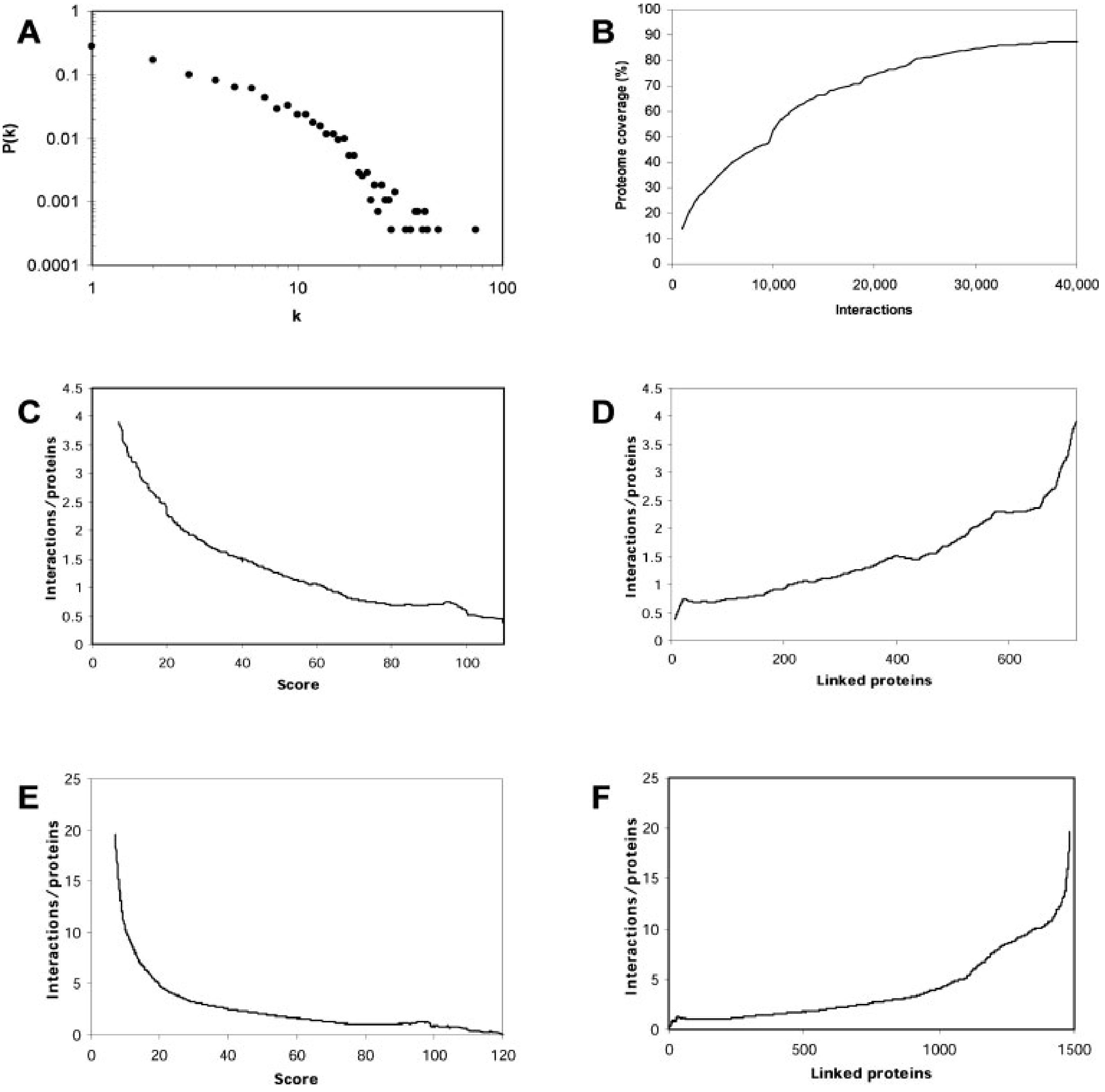

Figure 3. WI-PHI network analysis. Panel A illustrates the degree distribution which is the fraction of nodes having a given degree in a network including the 7000 highest scoring interactions. Visualized in Panel B is the percent coverage of the yeast "ORFome" as a function of interaction score threshold. It should be noted that not all the predicted ORFs of the yeast genome have been validated. Panel C displays the number of interactions per protein (only interactions between proteins within the same MIPS complex are considered) as a function of interaction score. Panel D shows the number of intracomplex interactions per protein (as in panel C) plotted against the number of proteins that have at least one link with complex comembers. (E, F): Same as C and D but for the complex cores as defined by Gavin *et al.* [1].

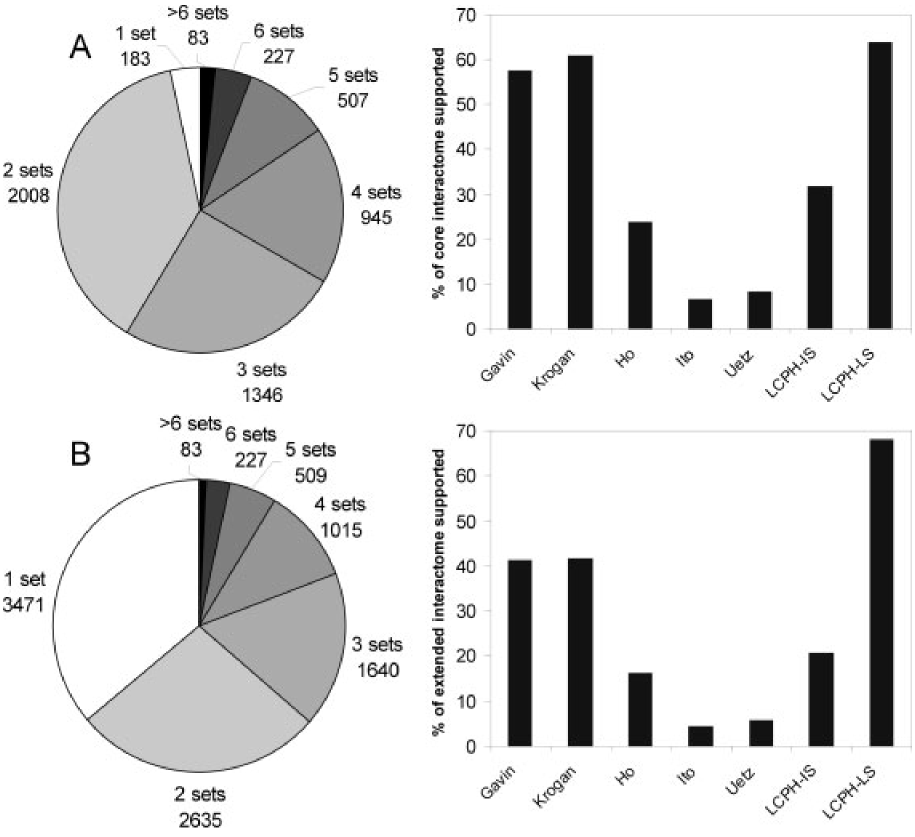

Figure 4. Dataset support for the WI-PHI core interactome (A) consisting of 5299 interactions and the WI-PHI extended interactome (B) consisting of 9580 interactions. Pie charts represent the entire set with the upper number of each slice being the number of supporting datasets, and the lower the number of interactions within that category. The bars in the right diagram depict the fraction of interactions within the interactome supported by each dataset.

Table 1. Network characterization

Interactome[a)]	Nodes	Edges	Average cluster coefficient	Average degree	Diameter
WI-PHI core	2406	5244	0.35	5.00	19
WI-PHI extended	2977	9225	0.36	6.64	18
Gavin core	1462	6942	0.53	9.95	12
Krogan core	2708	7123	0.19	5.50	12
Han *et al.* FYI	1378	2491	0.35	4.62	25

a) Datasets are defined in Section 2. Homodimers were removed prior to analysis.

As shown in Fig. 4, most of the interactions in the extended as well as core interactomes are supported by either or both of the two large affinity purification experiments. However, consistent with our effort to assemble an interactome enriched in direct binary interactions, a large part of the interactions of each of the yeast two-hybrid datasets, namely those confirmed by other supporting evidence, contribute to the two WI-PHI interactomes, although this is not reflected in Fig. 4, due to their modest size compared with the affinity purification sets. Various network properties for the two interactomes are reported in Table 1 together with the same values for other interactomes. We found that both the WI-PHI core and extended interactomes are characterized by properties common to biological networks: power-law like degree distribution, high clustering coefficient, and strong small world effect.

As our goal was to produce a yeast interactome enriched in direct physical connections between proteins, this property must be validated in detail. Hard evidence of direct interaction can be obtained only from crystallographic studies of interacting proteins. We have considered, as an independent validation, a set of 95 interactions whose direct physical contacts are supported by structural evidence (see Section 2). This benchmark set is included in WI-PHI but, for this analysis, we have subtracted its contribution from the interaction scores. As shown in Table 2, the starting datasets (considering only the “core” interactions of the TAP sets so comprising a total of 25 378 interactions) contain supporting evidence for as many as 92 of the “structurally proven direct interactions”. Ranking the interactions by our approach and compiling a network of 5299 interactions (the WI-PHI core interactome), we still retain supporting evidence for 76 structurally determined interactions, though the network is less than a fourth of the total size of the starting datasets. Even considering the limited scope of the benchmark, this is corroborating evidence for the enrichment of direct interactions in WI-PHI.

Table 2. Coverage of the structural benchmark by the different datasets

No.	Dataset	Dataset size	Structurally confirmed interactions
1	WI-PHI core	5 299	76
2	WI-PHI extended	9 580	79
3	Krogan core	7 123	61
4	Gavin core	6 942	39
5	LCPH-LS	6 514	47
6	LCPH-IS	5 958	63
7	Ito	3 275	6
8	Uetz	1 478	3
	3 + 4 + 5 + 6 + 7 + 8	25 378	92

To validate the interactome on a larger scale, we have adopted the methodology of Lee *et al*. (Science 2004), which they used to benchmark individual sets against a common standard derived from KEGG maps or GO. In this case, we have used the GO classifications available from SGD, and specifically looked at the Biological Process category. Most pairs of interacting proteins can be assumed to participate in the same biological process, although crosstalk as well as proteins with multiple functions (so-called "moon-lighting" proteins) are documented phenomena. Furthermore, due to the hierarchical nature of the GO, some ORFs may be mapped to different layers of the same general biological process producing dissimilar categories for the participating proteins, but these short-comings should be identical for the datasets. The analysis has been carried out with several score thresholds and clearly shows that the higher the WI-PHI score, the higher the likelihood of the interacting proteins sharing the same functional category in GO terms (Supporting Fig. 8). An increase in the observed frequency of shared biological process can be interpreted as an enrichment of true, relevant interactions.

Finally, we have looked at mRNA coexpression for interacting proteins in order to validate our approach. Genes with similar expression profiles can be assumed to encode proteins with close proximity in interaction space. Thus, coexpression can be viewed as corroborating evidence for interaction. From a recent compilation of microarray data in several species [25], we extracted yeast data and computed correlation coefficients between all pairs of genes for all conditions. A clear trend of higher fraction of coexpression between interacting protein pairs with higher WI-PHI score was observed (Supporting Fig. 9), further indicating that the WI-PHI ranking scheme promotes real protein–protein interactions over false positives ones.

3.5 Coverage of MIPS annotated-complexes dataset

Because of their nature, experiments based on affinity purification, which represent the vast majority of our input interactions, tend to yield highly intraconnected protein complexes. Not all of these interactions represent direct physical links between the proteins within the complex. Given that the SA index scoring system tends to favor direct interactions, and given the contribution to WI-PHI of interactions supported by methods describing direct interactions, by selecting an appropriate threshold, we expect WI-PHI to be enriched in interactions describing direct physical links. Using the MIPS set of trusted complexes [29], we monitored the connectivity of these complexes while adjusting the score threshold. By lowering the threshold for accepted scores, we initially add interactions connecting proteins to their partners in the complex, but eventually we start accumulating extra, "dispensable" links between proteins already connected within the complex. This relation is shown in Fig. 3D, which depicts the average number of interactions for proteins within MIPS complexes *versus* the number of proteins that, at any specific threshold, are connected to one of their complex comembers. The average number of links sharply increases at a threshold that connects about 700 MIPS proteins (approximately 65% of the total MIPS trusted-complex proteins). At this point, the accumulation of connections within the complexes increases at a much higher rate than the corresponding gain in new complex components. The relationship between average number of links and the score threshold is depicted in Fig. 3C. By combining the two graphs, one can conclude that, by choosing a score threshold of about 20, most proteins assigned to MIPS complexes are connected to one of their possible partners without unduly increasing the degree of any complex member. Figures 3E and F show the same analysis with a different set of complexes as defined by Gavin *et al*. [1]. The curves have similar trends supporting a threshold of 20 for minimum interaction redundancy within complexes.

Figure 5 shows the coverage of each individual trusted MIPS complex with the equivalent of the WI-PHI core interactome (corresponding to a score threshold of above 21). Although most of the complex components are connected by the interactions from our network, the members of a few complexes remained largely disconnected. Most of these annotated complexes, however, describe associations of functionally connected proteins with hardly any evidence of direct physical interactions. See, for instance, the "actin-associated proteins", the "tubulin-associated proteins", and the "other respiration chain" complexes.

The above analysis indicates that the WI-PHI core supports a substantial wiring of the MIPS-trusted complexes. However, WI-PHI core contains additional highly connected regions. After removing all the proteins annotated to yeast complexes, we identified 57 protein clusters with *k* core $\geq$3. Some of these correspond to complexes already annotated in SGD [30], others have been named according to their common functional annotation (see Supporting Fig. 7).

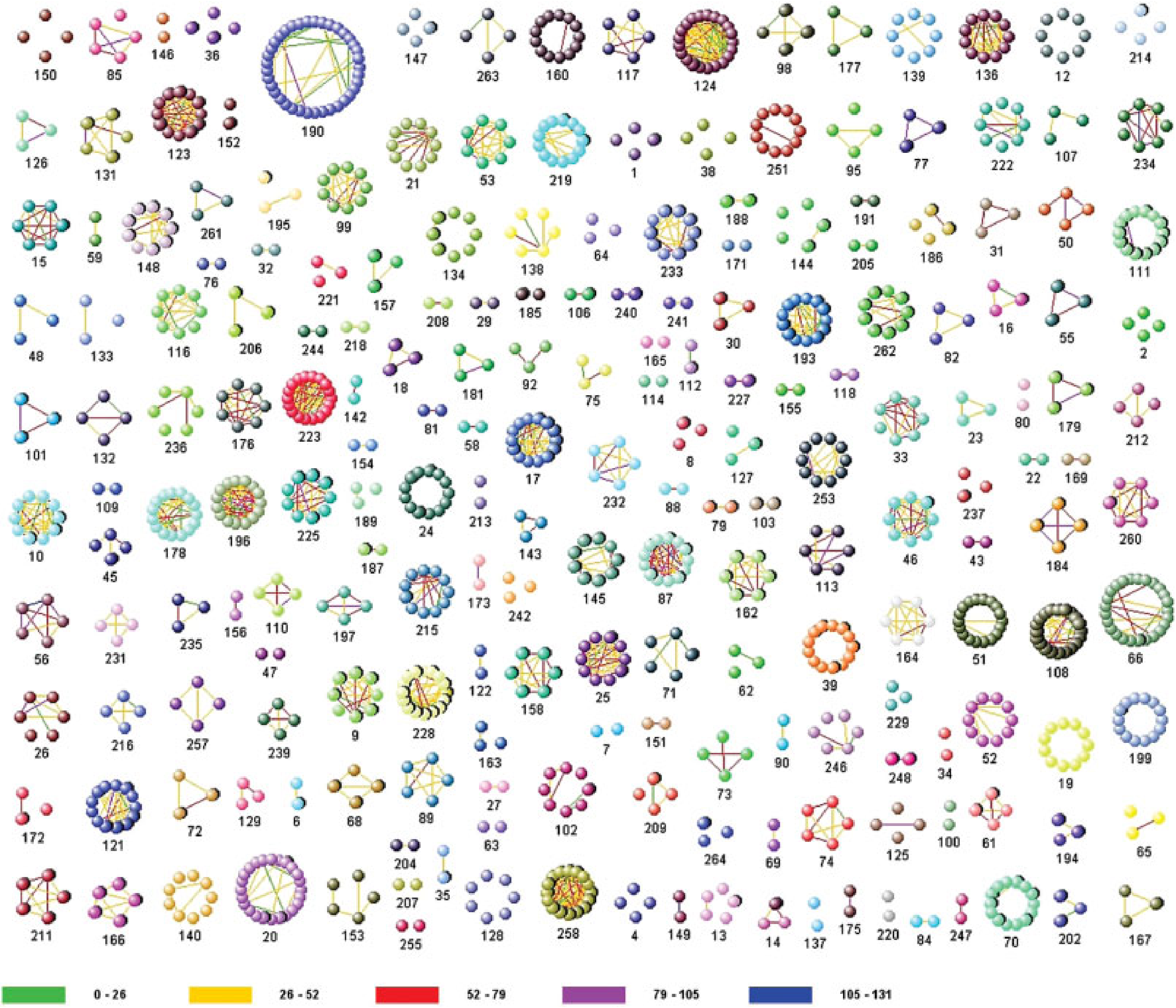

Figure 5. Connectivity of the MIPS-trusted complexes by a WI-PHI interactome with a threshold of 22. Edges are colored according to the score intervals detailed in the legend. Individual complexes are identified by numbers, which refer to a detailed description provided in the Supporting Information. Large and small ribosomal subUs are not considered in this analysis.

3.6 Wiring of the anaphase promoting complex (APC)

To further evaluate the robustness of the network on a specific biological example, and inspect the consequences of different threshold scores, we have taken a closer look at the APC. This complex serves as a ubiquitin ligase in the yeast cell and, although its structure is not known to atomic details, the basic architecture has been mapped [31]. We extracted interactions from the interactome in which at least one of the participants was a member of the yeast APC according to the MIPS-trusted complexes. From an interactome of 30 000 interactions, 11 proteins are involved in 111 interactions with themselves and 37 other proteins not annotated as APC members. From our threshold analysis, we expect most of these proteins and interactions to be false positives. Correspondingly, when the interactome is reduced in size, the number of internal interactions and the number of "attached" proteins drop dramatically (see Fig. 6). The APC complex is still coherent at even very stringent thresholds, but loses connectivity within its members perhaps reflecting more the true wiring of the complex. Even at the most stringent thresholds examined, a few additional proteins remain associated with the complex despite them not being annotated in MIPS as being members of the APC. However, a closer scrutiny reveals that two of these proteins are recognized as constituents of the APC in SGD [30], and two of the remaining ones have functional descriptions matching those of APC members (a cyclin, a ubiquitin-conjugating enzyme) indicating at least a functional relationship with the complex. This analysis of the APC complex wiring at different stringencies indicates that WI-PHI interactomes with more than 7000–9000 interactions, some of which are supported by only one experiment in a single study (score of 18-20), have an increased risk of including false-positive associations.

4 Discussion

The dynamic modeling of interactions occurring in a cell requires quantitative information about the direct physical interactions occurring between any pair of proteins in the proteome. Despite the vast amount of data published in the scientific literature, this information is not available in a structured and organized form. We were motivated by the recent publication of three large datasets to a fresh approach to the problem of integrating different types of interaction information. Our goal was to use the available information to compile a list of protein pairs ranked according to evidence of direct physical interaction and functional reliability. Different experimental approaches perform differently in this respect. A positive hit in the two hybrid method is traditionally considered strong evidence of direct physical interaction. On the other hand, the high percentage of false positives and the nonphysiological experimental settings entail caution in the interpretation of the results in the absence of further supporting evidence.

In contrast, the affinity purification method addresses the characterization of complexes that are formed in the cell and as such are likely to have physiological relevance, but the technique fails to provide evidence of direct contacts between the complex members.

To integrate the different datasets we have taken a three-step approach.

(i) First, we ranked all the interactions within the different datasets by calculating the SA indices as initially conceived by Gavin *et al.* [1] to process their own results.

(ii) Next, for each dataset, we calculated coefficients reflecting to which extent the interaction sets originating from the different experimental approaches match a set of

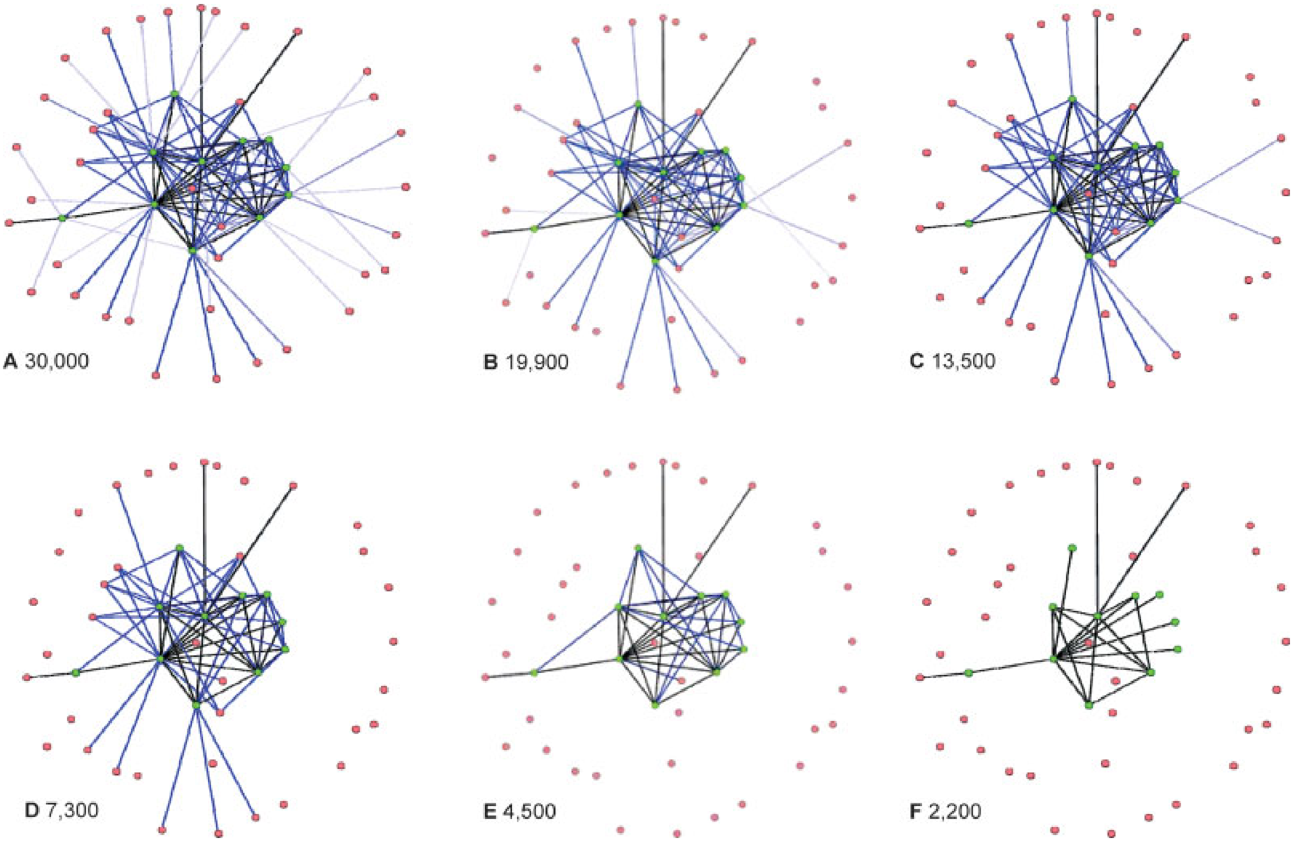

Figure 6. A closer inspection of the APC with varying interaction thresholds. Proteins reported to be part of the complex according to MIPS were extracted with all possible interactions above a score of seven, corresponding to an interactome of 30 000 interactions. Interactions are color coded according to interaction scores with darker edges corresponding to higher scores. Nodes representing proteins reported in MIPS as APC members are colored green, while the remaining nodes are first-order neighbors. For each version of the APC at different thresholds, the approximate total number of interactions in the corresponding interactome is shown. The interaction graphs were produced with Cytoscape 2.2 [39].

1777 highly trusted interactions in terms of coverage and specificity. We next used these coefficients as multipliers to combine the scores obtained by calculating the SA indices for each dataset.

(iii) Finally, we added the interactions detected in small-scale experiments with a score sufficiently high to make sure that they would be included within the best scoring 9 000 interactions even in the absence of further supporting evidence.

The end result is a ranked list of 50 000 interactions annotated with their supporting evidence, which we have chosen to name WI-PHI. This combined large network has a proteome coverage of almost 85%. However, even a smaller high confidence interactome, containing as few as 7500 interactions, connects around 50% of the yeast proteome (Fig. 3). This reflects an average of approximately 5.6 interactions per protein. The WI-PHI interaction graph properties are typical for biological networks, namely relatively high clustering coefficient, small diameter, and enrichment for interactions among products of essential genes. An interesting observation is that the WI-PHI core and extended networks have a lower average clustering coefficient than the best scoring interactions reported by Gavin *et al*. This could indicate that our integration strategy successfully exploits the complementary information present in the diverse datasets to preferentially select direct binary interactions, partially avoiding the problem of overpredicting interactions when the direct contacts between protein partners are unknown.

WI-PHI, because of its assembly strategy, has an inherently high coverage of a trusted interaction dataset. Furthermore, WI-PHI is enriched in direct interactions. There are a variety of reasons why our strategy should favor direct interactions. Firstly, the benchmark used to rate the performance of the different experiments is itself enriched in direct physical interactions. Secondly, the SA index has been shown by Gavin *et al*. to result in an increase in the proportion of direct interactions among a high scoring subset. Finally, results of small-scale experiments, which are included among the high scoring interactions in the WI-PHI network, are also enriched for direct interactions. Our validation strategies also support enrichment both in terms of shared GO classification as well as in terms of direct X-ray crystallography verified interactions. Despite this enrichment, our analysis suggests that we do not have sufficient information yet to confidently map all the direct interactions that are necessary to hold together the numerous functionally important complexes. Additional high confidence interactions require more effort in the task of mapping direct interactions and in combining this information with *in vivo* evidence.

WI-PHI represents an attempt to combine the available information and present a ranked list of interactions that can be filtered by progressively increasing the accepted score threshold. We are rather confident that the WI-PHI core network has a negligible percentage of false positives. By further lowering the accepted threshold, the risk of introducing false positives increases accordingly. The majority of false interactions between proteins in networks with a relaxed threshold are likely caused by indirect interactions which are very common in the “affinity purification” datasets. While this gives rise to a seemingly higher interconnectivity within complexes, it does not result in the inclusion of outright false interactions between completely unrelated proteins.

To further increase confidence in functional significance, we recommend applying filters based on orthogonal evidence such as coexpression data, colocalization, and other contextual evidence.

Fundamental limitations of the interactome presented here include lack of temporal and spatial information. Although integration of interaction data with spatial and temporal evidence has been limited till now by the scarcity of experimental information and by the development of suitable approaches, a number of recent reports describe experimental and computational efforts in this direction [16, 32–34]. We are confident that this strategy can be applied to other biological problems and that the interactome presented here is a good starting point for such an analysis.

Finally, we would like to point out the two most important remaining issues limiting our ability to model a dynamic interactome. The first one is the absence of information on association and dissociation kinetics. The situation is unlikely to change in the near future and requires the development of new high-throughput approaches which can address this need for kinetic information on interaction data at least in a semiquantitative way.

The second problem is the lack of information about the protein domains involved in the specific interactions. As pointed out in [35], the information represented in protein network graphs does not permit us to distinguish between a highly connected hub with the potential to interact simultaneously with many partners from one where the many partners compete for a single receptor site. Only a mapping of the regions responsible for the interactions to domains or interacting surfaces, and a transformation of the protein networks into domain networks will eventually allow confident modeling of the dynamic assembly of functional complexes. Thanks to recent technological advances and computational approaches, this goal may be within reach in the near future [36–38].

The authors wish to thank Maria Persico for fruitful discussions and Anders Fausbøll for help with figures. This work was funded by the EU-6th FW Integrated Project Interaction Proteome (LSHG-CT-2003-50520), the EU-6th FW Project ENFIN (LSHG-CT-2005-518254), and by AIRC.

5 References

[1] Gavin, A. C., Aloy, P., Grandi, P., Krause, R. *et al*., *Nature* 2006, *440*, 631–636.

[2] Krogan, N. J., Cagney, G., Yu, H., Zhong, G. *et al*., *Nature* 2006, *440*, 637–643.

[3] Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J. *et al.*, *J. Biol.* 2006, *5*, 11.

[4] Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L. *et al.*, *Nucleic Acids Res.* 2006, *34*, D535–D539.

[5] Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G. *et al.*, *FEBS Lett.* 2002, *513*, 135–140.

[6] Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P. *et al.*, *Nucleic Acids Res.* 2006, *34*, D436–D441.

[7] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S. *et al.*, *Nucleic Acids Res.* 2004, *32*, D452–D455.

[8] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P. *et al.*, *Nucleic Acids Res.* 2002, *30*, 303–305.

[9] Puig, O., Caspary, F., Rigaut, G., Rutz, B. *et al.*, *Methods* 2001, *24*, 218–229.

[10] Tong, A. H., Drees, B., Nardelli, G., Bader, G. D. *et al.*, *Science* 2002, *295*, 321–324.

[11] Vidal, M., *Cell* 2001, *104*, 333–339.

[12] Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., Botstein, D., *Proc. Natl. Acad. Sci. USA* 2003, *100*, 8348–8353.

[13] Han, J. D., Bertin, N., Hao, T., Goldberg, D. S. *et al.*, *Nature* 2004, *430*, 88–93.

[14] Lee, I., Date, S. V., Adai, A. T., Marcotte, E. M., *Science* 2004, *306*, 1555–1558.

[15] Zhang, L. V., King, O. D., Wong, S. L., Goldberg, D. S. *et al.*, *J. Biol.* 2005, *4*, 6.

[16] Hinsby, A. M., Kiemer, L., Karlberg, E. O., Lage, K. *et al.*, *Mol. Cell* 2006, *22*, 285–295.

[17] von Mering, C., Krause, R., Snel, B., Cornell, M. *et al.*, *Nature* 2002, *417*, 399–403.

[18] Ito, T., Chiba, T., Ozawa, R., Yoshida, M. *et al.*, *Proc. Natl. Acad. Sci. USA* 2001, *98*, 4569–4574.

[19] Ito, T., Tashiro, K., Muta, S., Ozawa, R. *et al.*, *Proc. Natl. Acad. Sci. USA* 2000, *97*, 1143–1147.

[20] Gavin, A. C., Bosche, M., Krause, R., Grandi, P. *et al.*, *Nature* 2002, *415*, 141–147.

[21] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A. *et al.*, *Nature* 2000, *403*, 623–627.

[22] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N. *et al.*, *Nucleic Acids Res.* 2005, *33*, D418–D424.

[23] Finn, R. D., Marshall, M., Bateman, A., *Bioinformatics* 2005, *21*, 410–412.

[24] Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A. *et al.*, *Nucleic Acids Res.* 1998, *26*, 73–79.

[25] Stuart, J. M., Segal, E., Koller, D., Kim, S. K., *Science* 2003, *302*, 249–255.

[26] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D. *et al.*, *Nature* 2002, *415*, 180–183.

[27] Hu, Z., Mellor, J., Wu, J., Yamada, T. *et al.*, *Nucleic Acids Res.* 2005, *33*, W352–W357.

[28] Jeong, H., Mason, S. P., Barabasi, A. L., Oltvai, Z. N., *Nature* 2001, *411*, 41–42.

[29] Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N. *et al.*, *Nucleic Acids Res.* 2005, *33*, D364–D368.

[30] Hirschman, J. E., Balakrishnan, R., Christie, K. R., Costanzo, M. C. *et al.*, *Nucleic Acids Res.* 2006, *34*, D442–D445.

[31] Thornton, B. R., Ng, T. M., Matyskiela, M. E., Carroll, C. W. *et al.*, *Genes Dev.* 2006, *20*, 449–460.

[32] de Lichtenberg, U., Jensen, L. J., Brunak, S., Bork, P., *Science* 2005, *307*, 724–727.

[33] Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W. *et al.*, *Nature* 2003, *425*, 737–741.

[34] Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S. *et al.*, *Nature* 2003, *425*, 686–691.

[35] Santonico, E., Castagnoli, L., Cesareni, G., *Drug Discov. Today* 2005, *10*, 1111–1117.

[36] Milstein, S., Vidal, M., *Nat. Methods* 2005, *2*, 412–414.

[37] Riley, R., Lee, C., Sabatti, C., Eisenberg, D., *Genome Biol.* 2005, *6*, R89.

[38] Landgraf, C., Panni, S., Montecchi-Palazzi, L., Castagnoli, L. *et al.*, *PLoS Biol.* 2004, *2*, E14.

[39] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S. *et al.*, *Genome Res.* 2003, *13*, 2498–2504.