nature biotechnology

Gaining confidence in high-throughput protein interaction networks

Joel S Bader[1], Amitabha Chaudhuri[2], Jonathan M Rothberg[2] & John Chant[2]

Although genome-scale technologies have benefited from statistical measures of data quality, extracting biologically relevant pathways from high-throughput proteomics data remains a challenge. Here we develop a quantitative method for evaluating proteomics data. We present a logistic regression approach that uses statistical and topological descriptors to predict the biological relevance of protein-protein interactions obtained from high-throughput screens for yeast. Other sources of information, including mRNA expression, genetic interactions and database annotations, are subsequently used to validate the model predictions without bias or cross-pollution. Novel topological statistics show hierarchical organization of the network of high-confidence interactions: protein complex interactions extend one to two links, and genetic interactions represent an even finer scale of organization. Knowledge of the maximum number of links that indicates a significant correlation between protein pairs (correlation distance) enables the integrated analysis of proteomics data with data from genetics and gene expression. The type of analysis presented will be essential for analyzing the growing amount of genomic and proteomics data in model organisms and humans.

Protein-protein interactions identified by high-throughput yeast two-hybrid screens (Y2H)[1–3] and inferred from mass spectrometry of coimmunoprecipitated protein complexes (Co-IP)[4,5] now cover three-quarters of yeast proteins. The intersection of high-throughput data sets with known interactions implies a total of 20,000–30,000 specific protein interactions in yeast, with the majority remaining to be discovered[6,7].

A challenging technical problem described in recent reviews of the Y2H[8] and Co-IP systems[9] may be the prevalence of spurious interactions in the high-throughput data. These may arise in the Y2H system from self-activators, in the Co-IP system from abundant protein contaminants and in both systems from weak, nonspecific interactions. Analysis based on concordance of interaction and expression data suggests that only 30–50% of the high-throughput interactions are biologically relevant[10].

Consequently, a crucial step in analyzing proteomics data is separating the subset of credible interactions from the background noise. We briefly review three types of methods that have been used: promiscuity and topological criteria, intersections between proteomics data sets and intersections with other types of data.

Promiscuity criteria described in the original reports[1,2,4,5] and used for later analysis[11] focused on removing proteins having many interaction partners. This method is often applied *ad hoc* because the distribution of interactions per protein shows a smooth decay, with no clear separation between 'sticky' (high connectivity) and 'nonsticky' (low connectivity) proteins. Furthermore, the scale-rich nature of biological networks suggests that highly connected proteins are a real feature of protein interaction networks[12–16]. A topological approach specific to Co-IP experiments, retaining the bait-hit (spoke) rather than bait-hit and hit-hit (matrix) interactions[7], enhances the data quality but discards 85% of the potential interactions. Topological measures of clustering have been applied to Y2H data[17].

The intersection of multiple high-throughput data sets is enriched for credible interactions[6,7,10]. A shortcoming of this method is the small number of interactions in the intersection: only 387 interactions are common to the 6,395 Y2H and 41,775 Co-IP interactions from high-throughput data. Beyond incompleteness of the screens, other factors reduce the probability that an interaction will be observed with both systems: the protein classes amenable to screening, the effective concentrations of the proteins in the engineered Y2H system relative to the *in vivo* Co-IP system, the strength of the interaction and the existence of nondirect or stabilized interactions.

Intersections with other types of data are also possible. Interacting proteins whose transcripts are coexpressed are more likely to be credible[10,18,19] and have been prioritized for experimental validation[20]. However, correlation of mRNA levels is not necessary for protein interactions, and even proteins in a permanent complex may have low transcriptional correlation owing to differences in degradation rates[19]. Even worse, as we show below for Co-IP data, correlated coexpression may be negatively correlated with predicted interaction confidence. Homology between a pair of proteins and a corresponding pair of interacting proteins has been used to enhance the confidence in high-throughput data and infer interactions across species[1,21], but the method is necessarily restricted to proteins with known homologs. Even for these proteins, a homology criterion may only identify half of the applicable true interactions as high confidence[10].

[1]Department of Biomedical Engineering, 201C Clark Hall, Johns Hopkins University, 3400 N. Charles St., Baltimore, Maryland 21218, USA. [2]CuraGen Corporation, 555 Long Wharf Drive, New Haven, Connecticut 06511, USA. Correspondence should be addressed to J.S.B. (joel.bader@jhu.edu).

Published online 14 December 2003; doi:10.1038/nbt924

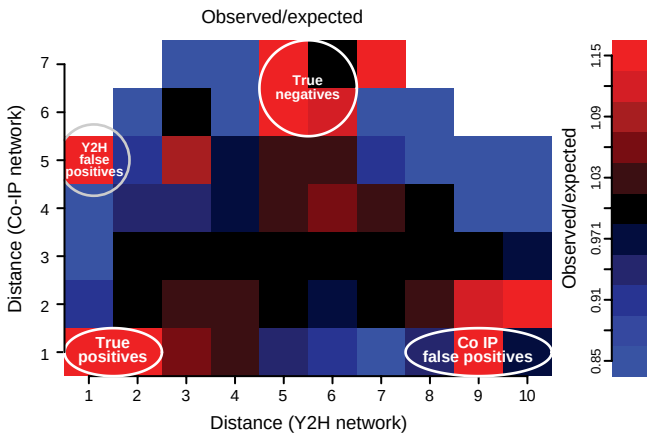

Figure 1 Network cross-comparison. Pairs of proteins have been binned according to their shortest path in networks generated from Y2H and Co-IP data. The false-color map indicates bins with more (red) or fewer (blue) interactions than expected by chance. Bins enriched for true positives, false positives and true noninteractors are indicated.

The methods described above are suitable for identifying a small, high-confidence set of interactions, but do not provide strong predictions for the majority of the high-throughput data. Here, we develop statistical models that assign a confidence score to every interaction. To avoid cross-contamination of data, we included only proteomics data in generating training sets and identifying explanatory variables; other data sources, including gene expression measurements, genetic interactions and database annotations, were used for independent validation of model predictions. Importantly, our statistical framework permits the incorporation of multiple, possibly correlated predictors.

A further challenge in analyzing the resulting network of high-confidence interactions lies in identifying the boundaries between distinct complexes: how many links can one follow from a central protein and still remain in the same complex or pathway? Although progress has been made in identifying network motifs that describe short-range clustering[22,23], it has been more challenging to identify patterns over longer length scales. We introduce a statistic that is sensitive to correlations over multiple protein-protein interaction links, predict the correlation distance of the network and validate the prediction with comparison to independent data.

Finally, we demonstrate how knowledge of the correlation distance may be used to develop algorithms that combine proteomics data with data from genetic screens, extracting known and previously unknown protein complexes with great selectivity, or with transcript profiles, providing a time-dependent view of network activity. The computational techniques developed here should have broad applicability given the increasing availability of genome-scale expression, genetic and interaction data in model organisms and humans.

RESULTS

Prediction and validation of confidence scores

We selected interactions for positive and negative training sets by comparing protein networks constructed from published Y2H and Co-IP data, using all bait-hit and hit-hit pairs when inferring Co-IP interactions (**Supplementary Table 1** online). Rather than focusing exclusively on interactions found in both data sets as positive examples, we examined a contingency table in which pairs of proteins were binned according to the distances between the paired proteins in each of the two networks (**Fig. 1**). This table shows that correlations between Y2H and Co-IP networks extend to four Y2H links and two Co-IP links and suggests that each Co-IP link is roughly equivalent to two direct Y2H interactions. As a positive training set, we selected Co-IP interactions between proteins that were one to two links apart in the Y2H network.

To select a negative training set, we reasoned that an interaction in one network between proteins that are far apart in the other network may be a technology-specific artifact. Indeed, the contingency table demonstrates an overrepresentation of Y2H interactions that are five Co-IP links apart, and Co-IP interactions that are nine Y2H links apart (**Fig. 1**). As a negative training set, we selected interactions in one training set between proteins whose distance in the other training set was larger than the median separation for random pairs of proteins, five for the Y2H network and three for the Co-IP network.

A third set of interaction distances is also enhanced: protein pairs separated by five to six links in the Y2H network and six to seven links in the Co-IP network. These pairs presumably represent proteins in unrelated pathways and could provide high-confidence examples of proteins that do not interact; we did not make use of them here.

The training sets were used to build logistic regression models that generated confidence scores for all interactions[24,25]. The complete set of explanatory variables represented data source, screening statistics, promiscuity and clustering.

For the Y2H data, the significant positive predictors were present in the Uetz I, Ito Core or literature data sources (see **Supplementary Methods** online for definitions) and the number of bridging proteins connected to both proteins in an interaction. The sole negative predictor was the number of proteins interacting with either protein in the pair.

For the Co-IP data, the positive predictors were the number of complexes containing the protein pair and the Jaccard correlation obtained as the ratio (no. of complexes containing both proteins)/(no. of complexes containing either protein). The negative predictors were the number of complexes in which the proteins were hit-hit pairs (rather than bait-hit pairs) and the number of complexes containing either protein. The Co-IP data source was not a significant predictor.

Additional measures of clustering, notably the clustering *P* value from the hypergeometric distribution[17], were significant as individual predictors but did not improve the final model for Y2H or Co-IP. Tenfold cross-validation of the training set (using 90% of the training set to predict confidence scores for the remaining 10%) indicated that the model was not over-determined. A formal description of the final model is available (**Supplementary Table 2** online). Training sets defined using other criteria yielded similar confidence scores (**Supplementary Methods** online).

To test the biological relevance of our statistical model, we verified that high-confidence Y2H and Co-IP interactions are more likely to involve proteins with similar database annotations (**Fig. 2a**). Similarly, mRNA expression correlations generally increase with Y2H interaction confidence (**Fig. 2b**).

For Co-IP interactions, however, we observed a negative correlation between interaction confidence and mRNA coexpression for low-confidence interactions; this was true for both high-throughput data sets, although slightly more pronounced for one set[5]. These results suggest that mRNA coexpression alone is not a good metric for assessing confidence in Co-IP data. Contaminant proteins or nonspecific binders with high expression levels, and thus high abundance, may be more likely to be recovered in Co-IP complexes. Indeed, higher absolute mRNA levels correspond to lower-confidence Co-IP interactions (**Fig. 2c**).

For further analysis of the high-confidence network, we retained those interactions with confidence scores greater than a threshold

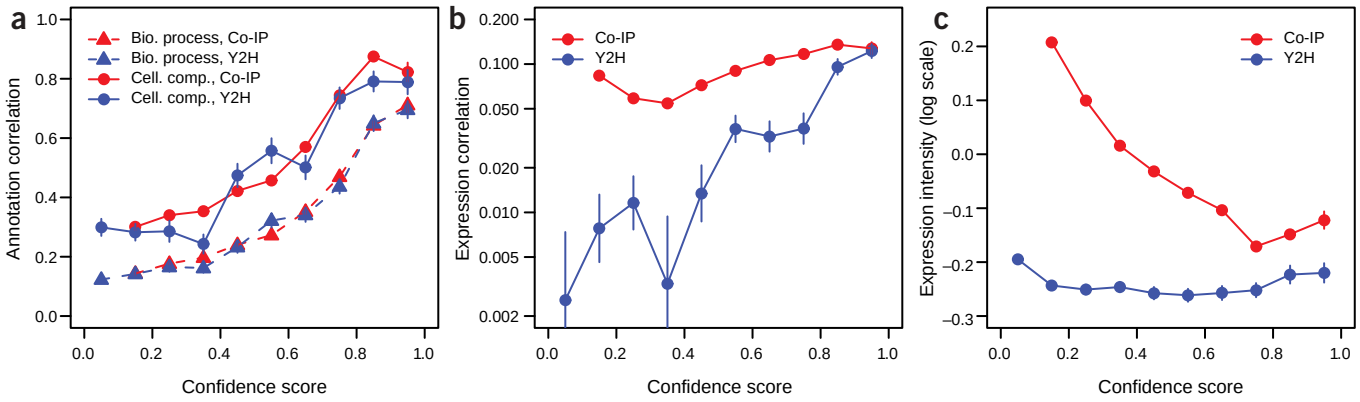

Figure 2 Confidence score validation. Protein-protein interactions inferred from Co-IP and Y2H networks were clustered according to the predicted interaction confidence. Biological measures excluded from modeling were calculated for the pairs of proteins within each bin. Lines are a guide to the eye only; vertical bars represent the s.e.m. and, when not visible, do not extend beyond the point. (**a**) Jaccard correlations were calculated from annotations for biological process and cellular component. (**b**) The mRNA expression correlation was calculated based on log-ratios from a series of 300 experiments[38]. For Co-IP experiments, both high-confidence and low-confidence pairs show enhanced coexpression. (**c**) Expression intensity for a pair of proteins was calculated as the mean of the $\log_{10}$ intensities[38]. Expression intensity is inversely correlated with interaction confidence for Co-IP interactions.

value. Although 0.5 was the threshold used to train the model, the training set was weighted to mimic a 50-50 mixture of low-confidence and high-confidence interactions. The full data set is dominated by low-confidence interactions, and a more stringent threshold is required for adequate enrichment. We estimated the enrichment of high-confidence interactions for different choices of the threshold, examined the results with a receiver operating characteristic curve and selected a threshold, 0.65, that is slightly more stringent than the 'elbow' of the curve and corresponds to 4.3-fold enrichment of biologically-relevant interactions (see **Supplementary Methods** online, confidence threshold and enrichment.)

We compared our predictions with the other published high-confidence data sets, focusing on 5,787 high-confidence interactions with confidence scores of 0.65 or better (for greater detail, see **Supplementary Table 3** online). Our high-confidence set contains 85% of the interactions in the intersection of Y2H and Co-IP data, and 15× more interactions overall. Similarly, we identified 88% of the high-confidence interactions predicted by a model based solely on local clustering, related to one proposed for Y2H interactions[17] and generalized here for Co-IP data, and 12× more interactions overall. We also identified 40–50% of the high-confidence interactions identified using mRNA coexpression and sequence homology metrics[10,20,26], and 2–6× more interactions overall. Finally, we identified 50% of the Co-IP bait-hit spokes as high-confidence interactions; our model may be effective in removing high-abundance impurity proteins from the Co-IP data (**Fig. 2c**). Furthermore, we found that the spoke model for Co-IP data is 4× more reliable than the matrix model, close to a 3× estimate reported earlier[7]. Although database annotations, expression measurements and genetic interactions are readily incorporated into the logistic regression framework and improve the model performance (**Supplementary Methods** online; see also ref. 27), we avoided using such models to maintain a wall between statistical-topological predictions and biological validation.

In summary, our confidence scores have strong concordance with previous predictions, increase the total number of high-confidence predictions by 2–10×, and, because they do not build in assumptions about coexpression or other biological criteria, can guide the incorporation of additional explanatory variables.

Prediction and validation of network clustering

We focused on the high-confidence interactions (confidence score 0.65 or higher) and, to examine clustering, initially considered the Y2H interactions and Co-IP spoke interactions (Co-IP matrix interactions were excluded because they yield trivial clusters formed by the proteins within each complex). The giant connected component formed by this subset of interactions, with 2,262 proteins and 3,854 interactions, has the properties of a small-world network: clustering at short distances and random connectivity at longer distances. Proteins in the giant component have 7.8 degrees of separation on average, compared to 5.3 degrees of separation in a comparable random network (**Fig. 3a**). The actual network may be flatter than a random network because clustering enhances local connections and suppresses far-away connections[11]. Clustering has also been observed in metabolic networks[15].

Although proteins interacting with a central protein are likely to be involved in a similar biological process, how many proteins are likely to form a single cluster? We introduce a topological statistic, the number of closed loops, that may be used to define a cluster size. The number of loops in the high-confidence network shows a significant enhancement over the randomized network (**Fig. 3b**). We developed a mathematical model for the expected number of loops in a clustered network in which connections within a cluster are enhanced over connections between clusters, derived approximate analytical expressions for the loop distribution and checked the analytical formulas by simulating clustered networks with the same parameters. For the actual network, the analytical model predicts a cluster size of 15 ± 2 proteins. The model suggests that each protein is effectively connected to 8.7 ± 0.4 other proteins in its cluster, an enhancement over the raw count of 3.4 neighbors per protein. The same model applied to a comparable randomized network does not identify any clustering. The correlation distance of the high-confidence network may be estimated as the number of links required to visit all the proteins in a cluster, $\log_{8.8} 15 \approx 1.24 \pm 0.06$ (see **Supplementary Methods** online).

Thus, even for the high-confidence network, we predicted that correlations decay over a length scale of one to two protein-protein interaction links, with the precise value depending on the significance threshold used to build the network.

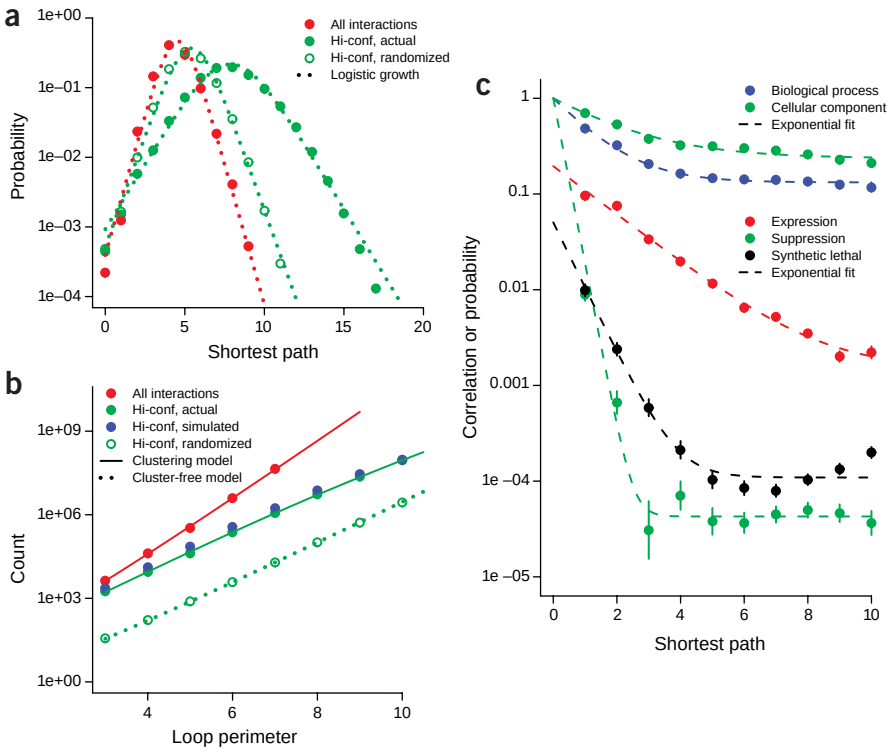

Figure 3 Distance-dependent correlations. (**a**) The distribution of shortest distances between pairs of proteins in the physical interaction network formed by high-confidence Y2H and Co-IP spoke interactions yields a mean separation of 7.8 links (filled green points). The mean separation for a comparable randomized network is 5.3 links (open green points). For comparison, the network formed by including low-confidence Y2H and Co-IP spoke interactions decreases the mean separation to 4.4 for the Y2H. The observed distributions are fit by a logistic growth model (dotted lines). (**b**) The number of loops has been calculated for each of the three networks of **a**, with the same symbols for observed data and mathematical fits. The distribution of loops for the randomized network is fit by a mathematical model that excludes clustering. The loop distribution for the actual high-confidence network requires the introduction of clusters and suggests that the correlation distance for proteins in the same cluster is one to two links. The loop distribution for the network including low-confidence interactions also shows evidence for clustering, but with much larger clusters. (**c**) A network was formed using all high-confidence Y2H and Co-IP spoke interactions, and the following quantities were calculated as a function of the shortest path connecting each pair of proteins: Jaccard correlations for biological process and cellular component annotations; mRNA coexpression correlation; and probabilities for synthetic lethal and suppression genetic interactions. Each correlation or probability was fit with exponential decay $a_1\exp(-d/a_2)$ plus a baseline, and a_2 was identified as the correlation distance in the network. Deviation from the fit for distances beyond four to five links may indicate additional structure in the network.

We tested the statistical prediction of the correlation distance by calculating distance-dependent correlations within the network. Starting from each protein in turn, we averaged database correlations, expression correlations and genetic interaction probabilities for increasingly distant shells of neighbors. Correlation distances were then obtained by fitting the distance-dependent correlations with exponential decay. Correlation distances for biological process and cellular component annotations were 1.20 ± 0.05 links and 2.0 ± 0.1 links, respectively. Correlation of mRNA expression decayed over 1.7 ± 0.1 links (**Fig. 3c**). These values are in the range of one to two links predicted by analysis of loops.

The correlation distances for synthetic lethal interactions, 0.64 ± 0.08, and suppression interactions, 0.25 ± 0.01, are much shorter (**Fig. 3c**), suggesting that genetic interactions provide specific indicators of biological function within protein complexes. Furthermore, its faster decay supports the view that suppression represents a close linkage in series, whereas synthetic lethality represents parallel or branching pathways.

We recalculated the distance-dependent correlation distances after adding the high-confidence Co-IP matrix interactions back to the network. Each correlation distance was within 1 s.d. of the Y2H + Co-IP spoke network, indicating that the cliques generated by high-confidence matrix interactions do not alter the distance-related properties of the network. Subsequent analysis was performed using the high-confidence Y2H + Co-IP matrix network.

Merging high-throughput proteomics and genetic data

Despite the statistical assurance of the biological relevance of individual interactions, analysis of the entire network remains daunting because of the high connectivity and the rapid decay of correlations. Although computationally intensive approaches have been described[28,29], the development of confidence measures and the existence of a correlation distance permit a direct approach to merging data sets.

The presence of a genetic interaction between proteins close together in the network suggests intuitively that these proteins have a high probability of being in the same complex. Indeed, the odds ratio may be calculated as the ratio of the distance-dependent genetic interaction probability to its long-distance value of 10^{-4}, approximately 100× for proteins one link apart and 10× for proteins two links apart (**Fig. 3c**). Thus, proteins separated by one or two physical interaction links and connected by a genetic interaction, together with their bridging proteins, are likely to define a single complex. We have identified every subnetwork defined by automated application of this algorithm (**Fig. 4**).

Several known complexes are immediately apparent: the actin-related protein complex, the origin recognition complex, the anaphase-promoting complex and a Skp/F-box protein degradation complex (**Supplementary Fig. 1** online). Reuse of proteins as components in multiple complexes with different functionality is illustrated by DNA replication and repair complexes.

Subnetworks built from high-confidence interactions further suggest intercomplex connections that would otherwise be obscured by low-confidence, spurious interactions (**Supplementary Fig. 1** online). These include connections between the spindle pole and myosin motor proteins, a linker protein connecting the catalytic core and the regulatory particle of the 26S proteasome, and a linker between the SAGA histone acetylation complex and a transcription mediator complex.

Merging proteomics with expression data

Static network topology is not sufficient to define function[30,31], and incorporating time-dependent expression data is important for

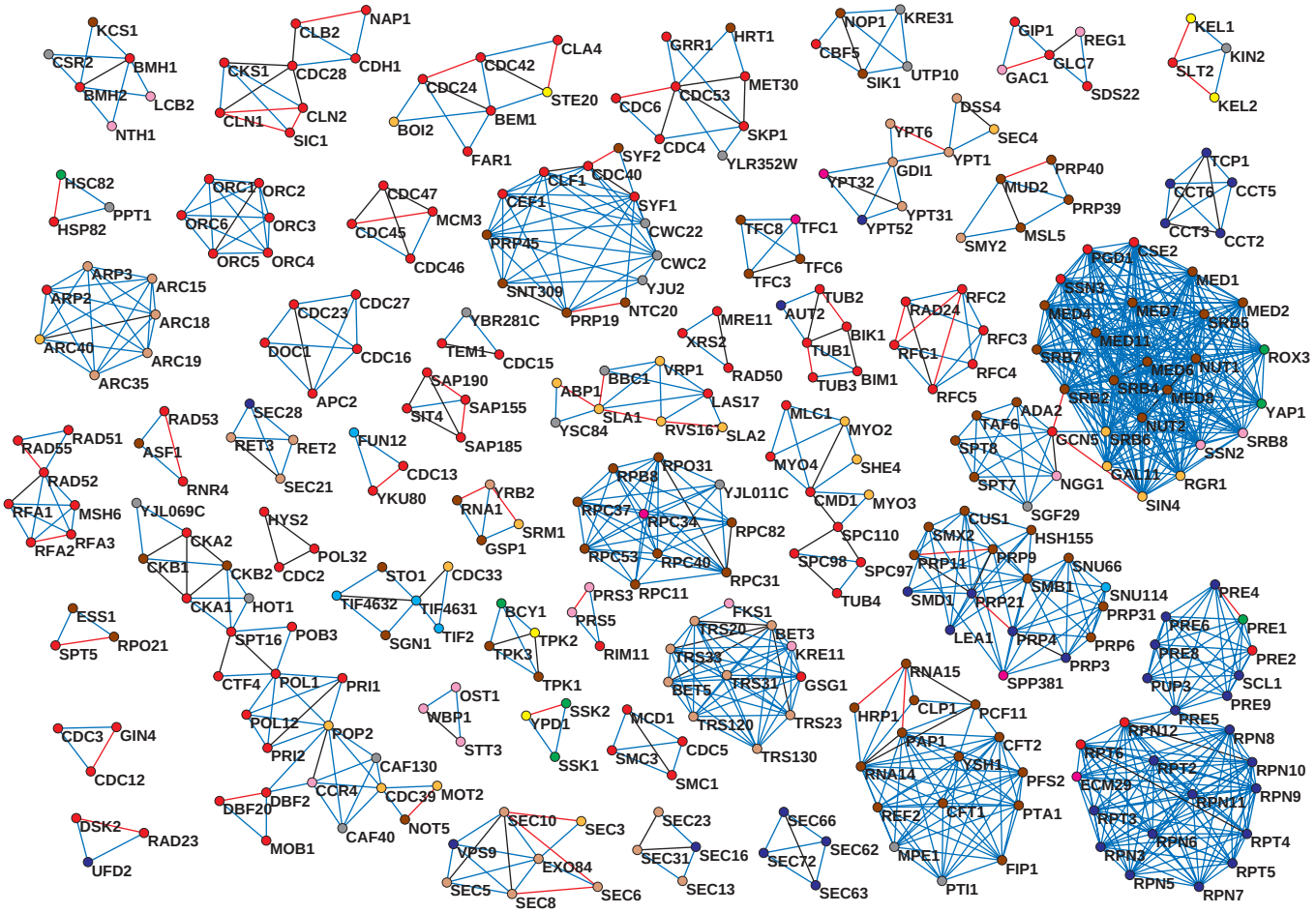

Figure 4 Joint analysis of physical and genetic interactions. Genetic interactions have been used as anchors to mine the physical interaction network. Lines indicate high-confidence physical interactions (blue), genetic interactions (red) and physical + genetic interactions (black). Protein color indicates biological process (red, cell cycle; green, cell defense; cell environment, yellow; cell fate, yellow; cell organization, magenta; metabolism, lavender; protein fate, blue; protein synthesis, cyan; transcription, brown; transport mechanisms, tan; gray, no annotation).

understanding pathway function. Inferring gene function from transcript profiles often proceeds through clustering of expression profiles to identify protein complexes and pathways. By merging transcription data with proteomics data, we resolve two problems with expression clustering: clusters of coexpressed transcripts that participate in disparate processes are split into process-specific subclusters, and proteins regulated by means other than transcription, and invisible to transcript profiling, are joined into clusters. We present the mitotic cell division cycle and the sporulation time course as examples. These have relevance as prototypical models for eukaryotic cell division, sexual reproduction, cell proliferation and cancer.

Gene expression profiles have shown that approximately 800 genes, or 10–15% of the yeast genome, are transcriptionally regulated during the cell cycle[32,33]. Within the cell cycle time course, each temporally regulated transcript has a single expression peak[34] during the cell cycle. Recent characterization of large-scale transcriptional regulatory networks in yeast provides an improved organizing framework for expression-based studies[22].

Cell division illustrates the two problems mentioned for expression clustering. First, hierarchical expression-based clustering[35] resulted in clusters with over 100 coregulated transcripts that then required expert analysis. Although subdivision of large clusters can be accomplished because much is known about yeast biology, an automated method for breaking expression clusters into biologically relevant groupings is advantageous. Second, clustering driven purely by expression ignores proteins whose transcripts are present constitutively but nevertheless have essential roles. These include cyclin-dependent kinases, key cell cycle regulators whose mRNA does not cycle and that were not identified by the mRNA profiling experiments.

We merged the static proteomics network with the expression data from the mitotic cell division cycle by selecting as anchors pairs of proteins satisfying the following criteria: (i) each protein was temporally regulated[33]; (ii) the time delay in peak transcription for the pair was no greater than 10% of the total cell cycle period; and (iii) the proteins either were connected directly or were bridged by a single protein in the high-confidence network (related methods use confidence scores to guide extraction[36]).

These anchoring pairs and all proteins directly connected to at least two anchors were extracted to yield distinct complexes containing 148 proteins and 177 physical interactions in all (**Fig. 5**; complexes listed in **Supplementary Table 4** online). Certain cell cycle proteins, including the main cyclin-dependent kinase Cdc28p, are not identified by mRNA profiling because their transcripts do not follow a cyclic expression pattern. Other proteins that are regulated but out of phase

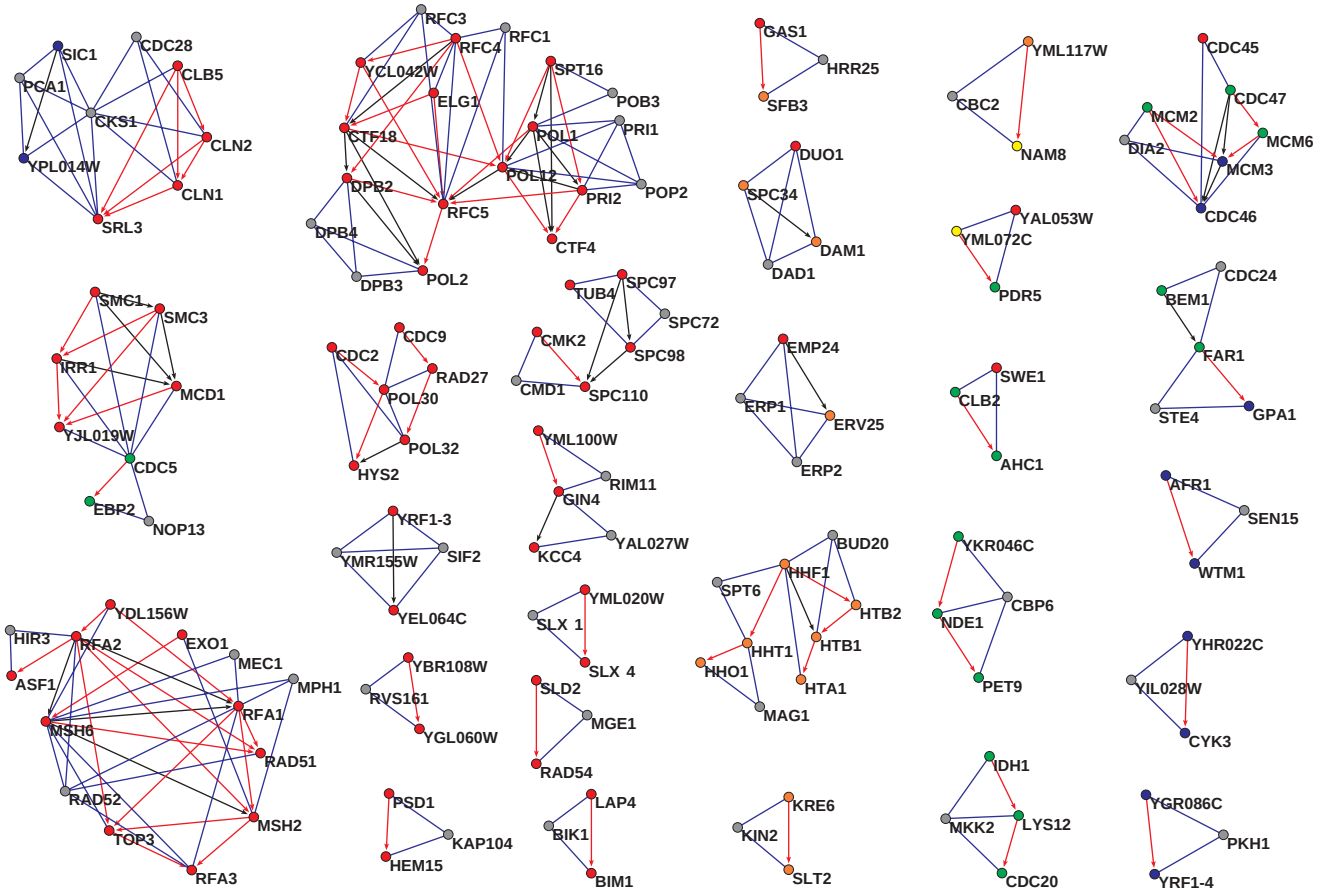

Figure 5 Joint analysis of cell division cycle coexpression with physical interactions. Subnetworks were extracted by identifying pairs of proteins one to two links apart and having peak transcription within 10 min during the yeast cell division cycle. Lines indicate physical interactions between proteins whose mRNA is transcribed close in time (black arrows), physical interactions between other pairs of proteins (blue) and noninteracting proteins transcribed close in time (red arrows). Proteins are colored according to cell cycle phase during peak expression (G1, red; S, orange; S/G2, yellow; G2/M, green; M/G1, blue; no transcriptional regulation, gray).

are nonetheless predicted to be members of the same complex: for example, cohesin (peak transcription at G1) and Cdc5p (peak transcription at G2/M). Merging expression with proteomics data clearly identifies these roles.

We performed a similar analysis for the gene expression time course of sporulation, which identified approximately 500 transcripts as induced during sporulation[37]. We defined a characteristic time for each transcript as the time to reach half-maximum log-scale expression. Pairs of induced proteins either directly connected or bridged by a single protein and with at most an hour difference in characteristic expression time were selected as anchors, and subnetworks were extracted as for the cell division cycle (**Fig. 6** and **Supplementary Table 5** online).

Here again, merging expression with proteomics data permits identification of complexes whose components may be expressed at disparate times. MAP kinase signaling components characterized as early-middle (Dig2p), middle (Ste5p) and middle-late (Rck1p) all interact with the nontranscriptionally regulated protein Fus3p, providing evidence for a single MAP kinase complex. Similarly, interactions between transcriptionally regulated Spc98p and Cdc5p with nonregulated Spc72p suggest that physical associations with the spindle pole body help regulate the activity of the protein kinase Cdc5p.

DISCUSSION

Analysis of protein physical interaction networks has often been hindered by a lack of confidence in the credibility of the individual interactions composing a network. We have demonstrated that it is possible to define a quantitative confidence measure based entirely on screening statistics and network topology. The principal assumption underlying the confidence measure is that nonspecific interactions are likely to be technology-specific. Although use of additional information—for example, expression correlation or annotation—could provide an improved measure of interaction confidence, we deliberately reserved its use for subsequent independent validation of our purely topological confidence measures. These methods will be useful for ongoing high-throughput proteomics screens[38].

The resulting high-confidence network is very complicated, essentially representing a superposition of all the pathways involving protein interactions for over one-third of the yeast proteome. The topological properties of the network indicate a hierarchical organization with a mean complex size of 15 proteins and a correlation distance of one to two links, which agrees with the correlation distance identified by database annotations and coexpression. Genetic interactions were found to represent an even finer scale of organization.

We exploited knowledge of the network correlation distance to extract biologically coherent subnetworks by merging proteomics data

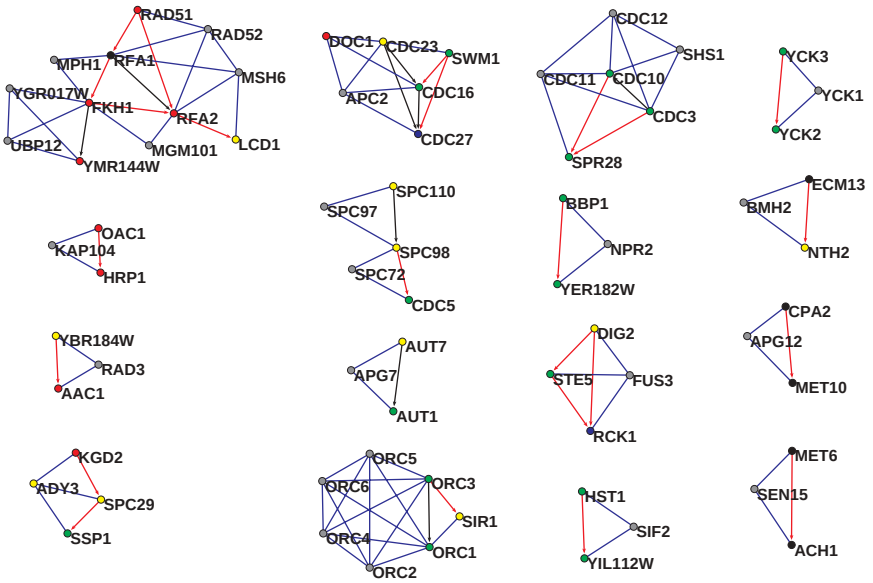

Figure 6 Joint analysis of sporulation-specific expression with physical interactions. Subnetworks were extracted by identifying pairs of proteins one to two links apart and having half-maximum transcriptional level (log-scale) within 1 h during the sporulation time course. Lines indicate physical interactions between proteins whose mRNA is transcribed close in time (black arrows), physical interactions between other pairs of proteins (blue) and noninteracting proteins transcribed close in time (red arrows). Proteins are colored according to cell cycle phase during peak expression (early I/II, red; early-middle, yellow; middle, green; middle/late and late, blue; metabolic induction, black; no transcriptional regulation, gray).

with genetic and expression data. By choosing anchoring pairs of proteins from genetic interactions and including all proteins within a short distance of both anchors, we extracted subnetworks that represent well-defined protein complexes. Interactions between complexes may suggest control points for pathways.

We then showed that combining the temporal gene expression clustering with the static proteomics network topology is a powerful, automated method for computationally extracting biologically relevant subnetworks. The method had exceptional specificity and could capture a greater number of subnetworks under less stringent criteria. We naturally resolved two problems of expression clustering: large clusters are automatically split into process-specific subclusters, and proteins whose mRNA profiles do not fit expected patterns are automatically included in the analysis.

Our analysis demonstrates the accuracy of combining qualitatively different forms of genetic, genomic and proteomics information for a systems-level understanding of biological complexes and networks.

METHODS

Data sources. Interactions were gathered from published high-throughput studies and a literature compendium[1–3,39], and all pairwise interactions were inferred for Co-IP complexes (see also http://depts.washington.edu/sfields/yplm/data/new2h.html/).

Self-validated training sets. The distances between pairs of proteins present in both networks were tabulated as $N_{obs}(D_y, D_c)$ = the number of pairs with distance D_y in the Y2H network and distance D_c in the Co-IP network. For Y2H interactions ($D_y = 1$), the positive and negative training sets were defined as $D_c = 1$ (387 pairs) and $D_c \geq 3$ (854 pairs). For Co-IP interactions ($D_c = 1$), the positive and negative training sets were defined as $D_y \leq 2$ (938 pairs) and $D_y \geq 5$ (12,471 pairs).

Explanatory variables. For Y2H interactions, we defined h_i as the number of interaction partners of protein i. Then, for an interaction between a pair of proteins arbitrarily labeled 1 and 2, we defined $h_{min} = \min(h_1,h_2)$; $h_{max} = \max(h_1,h_2)$; $h_{geom} = (h_1h_2)^{1/2}$; h_{12} = the number of proteins interacting with both protein 1 and protein 2; the Jaccard coefficient, $h_{jac} = h_{12}/(h_1 + h_2 - h_{12} - 2)$; and $h_{lod} = -\log_{10}$ (P value for shared neighbors) from the hypergeometric distribution[17]. Mathematical transforms, primarily sqrt and log, of variables were also entered, as were 1/0 indicator variables for presence/absence in each data source.

For Co-IP interactions, we used variables m_{12}, the total number of complexes including both proteins; $spoke_{12}$, the number of complexes with one protein as the bait and the other as a hit; $chord_{12}$, the number of complexes with both proteins as hits; m_1 and m_2, the total number of complexes containing m_1 and m_2, respectively; m_{jac}, the Jaccard correlation $m_{12}/(m_1 + m_2 - m_{12})$; and $-\log_{10}(P$ value for $m_{jac})$ from a hypergeometric distribution. Mathematical transforms of variables were also entered, as were 1/0 indicator variables for the data sources.

Logistic regression modeling. Logistic regression[24] as implemented by the glm function of *R* was used for statistical modeling of confidence scores. Positive and negative training set examples were weighted inversely to the number of such examples to mimic a prior distribution of 50% high-confidence, 50% low-confidence interactions. A stepwise procedure under Akaike's 'An Information Criterion' was used to identify a subset of statistically significant predictors; interaction terms involving pairs of predictors were then tested but found not to improve the model performance. Classification rates were similar for the full model and for tenfold cross-validation, indicating that the model was not over-determined.

Literature validation of predicted confidence scores. Database annotations were obtained from MIPS functional classification and subcellular localization catalogs[39], and correlations between pairs of proteins were calculated as the Jaccard correlation of shared annotation terms. Expression correlations and intensities were obtained from a compendium of 300 experimental states[40].

Empirical random networks. An empirical distribution of random networks was obtained by iterative interaction swapping using a permutation algorithm to rapidly generate trial swaps.

Mathematical random networks. Mathematical models were developed for networks in which each possible interaction has an equal probability of being present or absent, which does not match the empirical distribution of interactions per protein but has the advantage of permitting closed-form, analytical expressions for network properties in terms of simple network properties. The number of proteins d links away from a central protein in this model is $N/[1 + (N-1)J^{-d}]$, where J is the number of neighbors per protein and N is the total number of proteins in the network. A hierarchical random model was then developed to described enhanced connectivity within a protein complex. This models predicts that the number of loops of perimeter n is $(J_1 + J_2)^n/2n - J_2^n/2n + N_1(J_2^n/2n)\exp(-n^2/2N_2)$, where N_1 is the number of complexes, N_2 is the number of proteins per complex, and J_1 and J_2 are the number of between-complex and within-complex neighbors per protein, respectively.

Distance-dependent correlations. Annotation similarity, coexpression, and genetic interactions from high-throughput screens and the literature[39,41] were correlated with the shortest distance between a pair of proteins in the network. Decays were fit with the functional form $y_{fit} = a_1\exp(-d/a_2) + a_3$ by minimizing the sum of $[\log(y_{obs}) - \log(y_{fit})]^2$.

Extraction of anchored networks. Anchored networks were extracted by identifying pairs of anchor proteins (either a genetic interaction pair or transcribed with a short time-lag) connected directly or bridged by a protein in the physical interaction network. Related algorithms use the confidence of each link as a guide to network extraction[36].

More details of the methods are provided as **Supplementary Methods** online.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

J.S.B. acknowledges his colleagues at CuraGen who generated much of the data analyzed here and whose discussions have been enjoyable and productive.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 16 July; accepted 30 October 2003
Published online at http://www.nature.com/naturebiotechnology/

1. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* **403**, 623–627 (2000).
2. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
3. Tong, A.H. *et al.* A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321–324 (2002).
4. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
5. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
6. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
7. Bader, G.D. & Hogue, C.W. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.* **20**, 991–997 (2002).
8. Phizicky, E., Bastiaens, P.I., Zhu, H., Snyder, M. & Fields, S. Protein analysis on a proteomic scale. *Nature* **422**, 208–215 (2003).
9. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
10. Deane, C.M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* **1**, 349–356 (2002).
11. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).
12. Watts, D.J. & Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
13. Barabasi, A.L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
14. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
15. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabasi, A.L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
16. Wolf, Y.I., Karev, G. & Koonin, E.V. Scale-free networks in biology: new insights into the fundamentals of evolution? *Bioessays* **24**, 105–109 (2002).
17. Goldberg, D.S. & Roth, F.P. Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. USA* **100**, 4372–4376 (2003).
18. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae. Nat. Genet.* **29**, 482–486 (2001).
19. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12**, 37–46 (2002).
20. Kemmeren, P. *et al.* Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* **9**, 1133–1143 (2002).
21. Matthews, L.R. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res.* **11**, 2120–2126 (2001).
22. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science* **298**, 799–804 (2002).
23. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
24. McCullagh, P. & Nelder, J.A. *Generalized Linear Models*, edn. 2 (Chapman & Hall, London, 1983).
25. Hastie, T., Tibshirani, R. & Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2001).
26. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305 (2002).
27. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
28. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18** (Suppl. 1), S233–S240 (2002).
29. Bader, G.D. & Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
30. Guet, C.C., Elowitz, M.B., Hsing, W. & Leibler, S. Combinatorial synthesis of genetic networks. *Science* **296**, 1466–1470 (2002).
31. Bhalla, U.S., Ram, P.T. & Iyengar, R. MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. *Science* **297**, 1018–1023 (2002).
32. Cho, R.J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73 (1998).
33. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
34. Zhao, L.P., Prentice, R. & Breeden, L. Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci. USA* **98**, 5631–5636 (2001).
35. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
36. Bader, J.S. Greedily building protein networks with confidence. *Bioinformatics* **19**, 1869–1874 (2003).
37. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
38. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster. Science*; published online 6 November 2003 (doi:10.1126/science.1090289).
39. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002).
40. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
41. Tong, A.H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).