24 Bioinformatic Prediction of Yeast Gene Function

Insuk Lee, Rammohan Narayanaswamy and Edward M. Marcotte
Center for Systems and Synthetic Biology, Institute for Cellular & Molecular Biology, University of Texas at Austin, Austin, TX 78712, USA

CONTENTS

I. INTRODUCTION

The bioinformatic prediction of gene function is, although young, already an extensive field, and with the high quality of the yeast genome sequence and the already large and rapidly growing volume of yeast functional genomics data, the prediction of yeast gene function is a substantial subfield in itself. A wide variety of approaches have been developed to predict gene function, ranging from sequence analyses to assign genes into functional families (Bork and Koonin, 1998; Ponting, 2001; Bateman *et al.*, 2004), to structural analyses to assign protein folds (Honig, 1999; Schonbrun *et al.*, 2002; Godzik, 2003) and active sites (Fetrow and Skolnick, 1998; Madabushi *et al.*, 2002), to phylogenetic analyses for subdividing gene families into functional subgroups (Eisen, 1998a, b; Abhiman and Sonnhammer, 2005; Engelhardt *et al.*, 2005) or predicting interacting partners (Pazos and Valencia, 2002). As 'gene function' takes such a wide variety of forms, from the corresponding protein's biochemical activity to its physical interaction partners to membership in a given pathway, we focus here only on the latter 'network' aspects of gene function: a protein's interaction and pathway partners, and the inferences of function that derive from these.

METHODS IN MICROBIOLOGY, VOLUME 36
0580-9517 DOI:10.1016/S0580-9517(06)36024-2

One of the most effective strategies for inferring pathway-type functional information has turned out to be the general strategy of 'guilt by association' (e.g., as in (Eisen *et al.*, 1998; Marcotte *et al.*, 1999b; Aravind, 2000; Eisenberg *et al.*, 2000; Oliver, 2000; Wu *et al.*, 2002; Huynen *et al.*, 2003; Xia *et al.*, 2004; Jiang and Keating, 2005; Wolfe *et al.*, 2005), to name but a few). This chapter will discuss the inference of yeast gene function via guilt-by-association approaches, discussing a variety of relevant functional and comparative genomics approaches, and their integration to predict gene function more accurately. We focus in particular on how these approaches can be made quantitative by estimating the error rates in these data and in the predicted gene functions.

♦♦♦♦♦♦ II. PREDICTING FUNCTION THROUGH GUILT-BY-ASSOCIATION

A. A General Principle for Finding Yeast Gene Function

The general strategy of guilt by association involves implicating genes in the same biological processes. Linking an uncharacterized gene to genes known to function in ribosome biogenesis carries an implication that the uncharacterized gene functions in this general area as well. The specific linkages may imply more specific function. This strategy can be employed with many different classes of functional and comparative genomics data, some of which allow stronger inferences than others. The strength of inferences vary depending not only on the immediate links, the type of data, but also the larger data set beyond the immediate genes of interest (i.e., a data set might, for example be strong for *certain* classes of genes but weak for others), as well as the *prior* chances of such inferences being correct, an aspect that is frequently overlooked in these analyses.

In this section, we will first introduce the various classes of data useful for guilt-by-association inferences, discussing the forms of inferences that are commonly made from them. As we will see in Section III, all of these approaches can be made quantitative without explicit development of statistic models through supervised methods of benchmarking and measuring error. Many of the experimental techniques are treated in more detail in other chapters, including yeast two-hybrid assays (Chapter 6), expression analysis (Chapter 9), protein localization (Chapter 13), and synthetic genetic arrays (Chapter 16).

B. Guilt-by-Association via Functional Genomics

1. Protein interaction mapping by yeast two-hybrid and mass spectrometry

Yeast protein–protein interaction data are primarily derived from two approaches: (1) genome-wide, high-throughput yeast two-hybrid

experiments, by which over 4000 unique protein interactions were observed between yeast proteins in three large-scale experiments (Ito *et al.*, 2000; Uetz *et al.*, 2000; Ito *et al.*, 2001), and (2) affinity purification of complexes of yeast proteins, followed by identification of the proteins by mass spectrometry (Gavin *et al.*, 2002; Ho *et al.*, 2002), identifying thousands more interactions among yeast proteins.

In addition to the large-scale experimental approaches, a number of groups has collected previously measured protein–protein interactions from the biological literature (Blaschke *et al.*, 1999; Humphreys *et al.*, 2000; Proux *et al.*, 2000; Thomas *et al.*, 2000; Marcotte *et al.*, 2001). This systematic collection of known protein interaction data provides necessary checks on the quality of the large-scale interaction data; large-scale protein interaction data have varied widely in accuracy (Mrowka *et al.*, 2001; Deane *et al.*, 2002; von Mering *et al.*, 2002).

Protein interaction databases combine the interactions from large-scale screens with interactions extracted from the literature, and include the biomolecular interaction network database (BIND) (Bader *et al.*, 2003) and the general repository for interaction datasets (GRID) (Breitkreutz *et al.*, 2003) databases and the database of interacting proteins (DIP). As of this writing, the DIP (http://dip.doe-mbi.ucla.edu/; Salwinski *et al.*, 2004) currently contains >18 000 protein–protein interactions among >4900 yeast proteins. The GRID database (http://biodata.mshri.on.ca/grid) includes >20 000 yeast protein–protein interactions. The BIND database (http://www.bind.ca/) includes >71 000 yeast molecular interactions, although these include non-protein–protein interactions in the count. For example, protein–DNA interaction data are also accumulating rapidly, primarily due to the scaling of chromatin immunoprecipitation methods to genome scale using DNA microarrays (Ren *et al.*, 2000; Bulyk *et al.*, 2001; Iyer *et al.*, 2001; Mukherjee *et al.*, 2004), allowing large-scale assays of ~200 yeast transcription factor binding specificities (Lee *et al.*, 2002; Harbison *et al.*, 2004).

For the purposes of inferring function from these interaction data, it is important to consider the model under which inferences are drawn. In particular, in direct measurements of protein interactions, such as the two-hybrid and mass spectrometry data, the experiments are typically performed by measuring interactions between a 'bait' protein and whatever 'prey' proteins it may interact with. If one protein (the 'bait') is observed to interact with multiple 'prey' proteins, there is no guarantee that the 'prey' will also interact with each other, although this may be likely in the case when they are members of the same protein complex. As shown in Figure 1A, there is a distinction made (Bader and Hogue, 2002) between a 'spoke' interpretation, in which only directly observed interactions between 'bait' and 'prey' are considered, and a 'matrix' interpretation, in which 'prey' bound by the same 'bait' protein is also inferred to interact with each other. Intuitively, the spoke model may seem too restrictive at times and the matrix model too permissive. As we

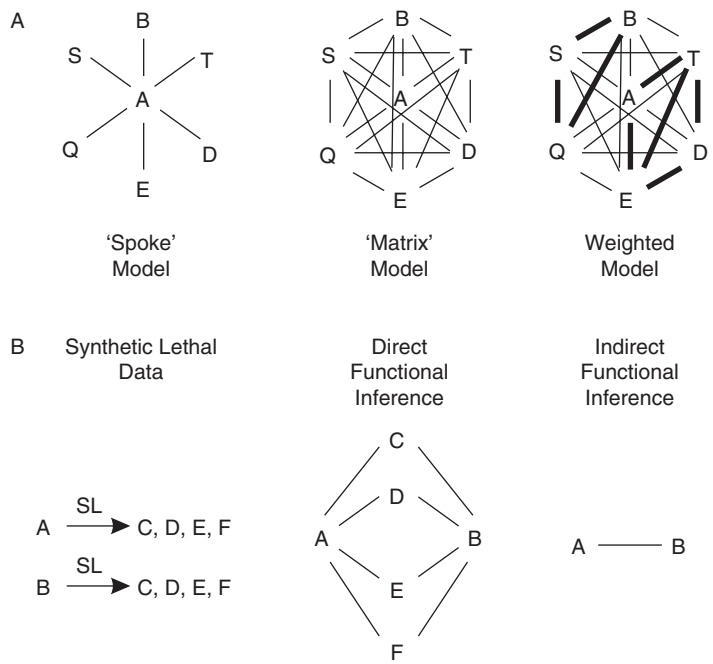

Figure 1. Alternate models for determining associations from functional genomics data. (A) Direct measurements of interactions, such as by yeast two-hybrid or mass spectrometry, can be interpreted as only providing evidence of 'bait'—'prey' associations (the 'spoke' model), as providing evidence for 'bait'—'prey' and 'prey'—'prey' associations (the 'matrix' model), or can be assigned weighted confidence scores based on interactions from the rest of the screen, as described in Section IV. (B) Genetic interaction data can provide evidence for associations between the synthetic lethal partners, or, less obviously, can provide evidence for linkages between genes synthetic lethal to the same set of other proteins.

will demonstrate in Section IV, there is an alternative model to these approaches, the weighted interaction model, which outperforms both of these strategies.

2. Genetic interactions and synthetic genetic arrays

Functional associations, far from being limited to physical interactions, can be drawn from more general associations between genes, such as those provided by genetic interactions. In yeast, the bulk of these data are from synthetic genetic array experiments, in which two mutant strains are robotically mated, sporulated, and the double mutant progeny examined for synthetic phenotypes, such as lethality (Tong *et al.*, 2001; Tong *et al.*, 2004). Unlike physical interactions, synthetic lethal relationships are not necessarily simple to interpret. They clearly represent legitimate constraints on the cell to grow properly, and it is generally perceived that the experiments have low false-positive rates (although this is hard to measure) that generally stem from technical errors, such as occasional defects in

the original yeast deletion strain collection (Giaever *et al.*, 2002), rather than biological artifacts in the screens.

Nonetheless, it has been shown (Wong *et al.*, 2004; Kelley and Ideker, 2005) that only a fraction (perhaps half) of synthetic lethal interaction partners belong to the same biological pathway. Therefore, synthetic lethal interactions give two alternate interpretations for the purposes of inferring gene function, as illustrated in Figure 1B. Given a synthetic lethal interaction between two genes, one can interpret this as partial evidence that they belong to the same pathway. However, the same inference can often be drawn in the case where two genes are not themselves synthetic lethal to each other, but have synthetic interactions *with the same set of other proteins*.

3. Co-expression and co-localization

Owing to the prevalence of publicly available large-scale mRNA expression data sets, strong functional inferences can be drawn through analyses of genes' expression patterns. These data are primarily in the form of thousands of DNA microarray experiments stored in the Stanford Microarray Database (Gollub *et al.*, 2003) and the GEO database (Barrett *et al.*, 2005). These data have proved powerful in the guilt-by-association style transfer of function, with diverse algorithms developed to mine the data, ranging from simple calculations of correlations between genes expression profiles across a bank of microarray experiments to a rich variety of clustering, classification, and deconvolution algorithms for more sophisticated grouping of genes into functional groups (e.g., as reviewed in Slonim, 2002).

Complementing the mRNA expression data is yeast protein localization data, primarily from large-scale analyses of fusion protein localizations (Habeler *et al.*, 2002; Kumar *et al.*, 2002; Huh *et al.*, 2003). These data provide an important source of functional associations that vary from extremely specific (e.g., both proteins of interest localize to the spindle pole body) to very general (e.g., both are cytosolic). These data have proved most useful for functional inference when combined with other datasets (Jansen *et al.*, 2003).

C. Guilt-by-Association via Comparative Genomics

A number of comparative genomics methods has been employed to identify yeast gene function. Here, we summarize three of these approaches, in particular:

(1) The discovery of functional associations via the observation that bacterial orthologs of the genes occur in the same operons.
(2) The discovery of functional associations based upon co-inheritance of genes across many organisms.
(3) The discovery of functional associations by observation of gene fusion events.

1. Deriving yeast gene function from bacterial genome organization

This approach relies upon the trend for bacterial genes of related function to be organized into operons. Therefore, yeast orthologs of these genes are also likely to function together. Although many operons are known for some organisms (e.g., see the RegulonDB database for known operons of *E. coli*, Salgado *et al.*, 2004), many more are uncharacterized. Two computational methods, illustrated in Figure 2A and B, have proven effective for predicting functional relationships between genes by their orthologs' tendencies to co-occur

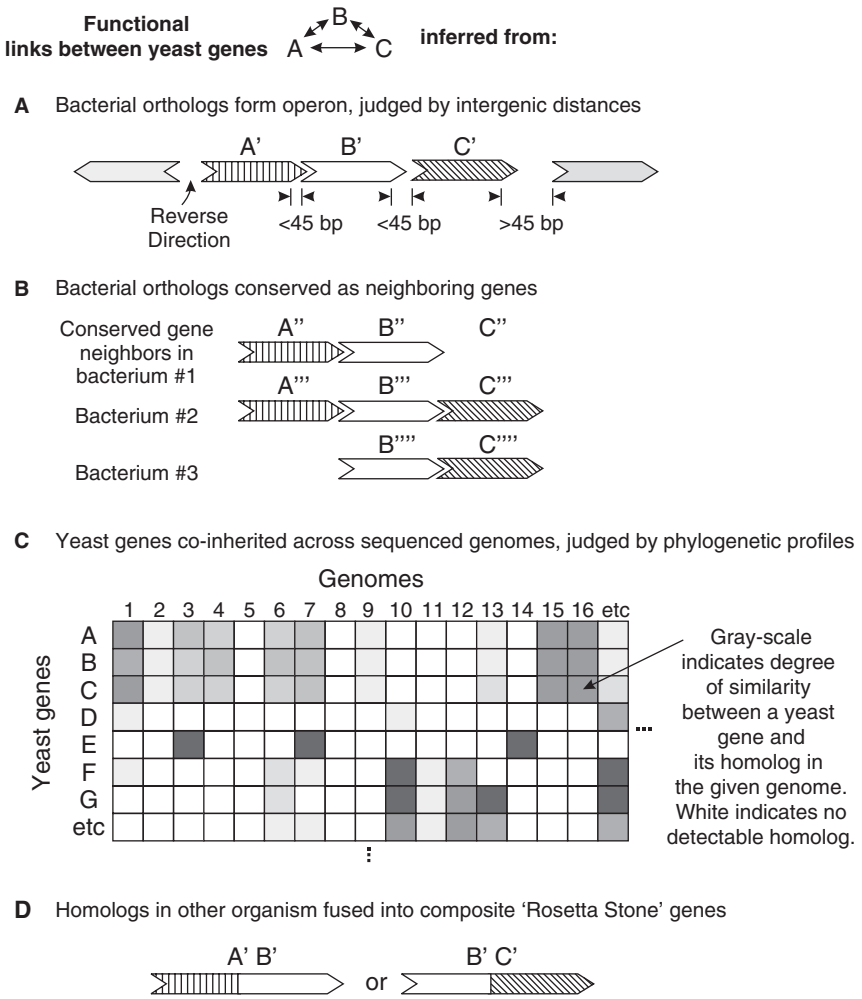

Figure 2. Functional associations derived from comparative genomics, in which the genomic organization of homologs or orthologs of the yeast genes may imply associations between the yeast genes. The genomic organization of yeast genes' bacterial orthologs into operons may imply functional associations between yeast genes, judged by (A) intergenic distances and (B) conservation of gene neighbors. Functional associations may also be inferred based upon (C) co-inheritance of genes, as measured by calculating and comparing phylogenetic profiles, or upon (D) fusions of homologous genes into composite genes.

in bacterial operons. Both of these methods ignore the identification of promoters and regulatory sequences, and instead exploit other properties to identify the operons. One approach exploits the tendency for adjacent genes in the same operon to be separated by short intergenic distances, while adjacent genes in different operons are separated by longer intergenic distances (Salgado *et al.*, 2000). The second approach, often referred to as the 'conserved gene neighbor' approach, exploits the tendency for genes in operons to be conserved as adjacent, neighboring genes in multiple bacterial genomes (Tamames *et al.*, 1997; Dandekar *et al.*, 1998; Overbeek *et al.*, 1999; Snel *et al.*, 2002; Yanai *et al.*, 2002). These approaches have proved remarkably powerful in yeast, as at least 1302 yeast proteins have bacterial orthologs with conserved gene-order information (Huynen *et al.*, 2003).

2. Phylogenetic profiles and the principle of co-inheritance

The second approach for identifying pathway and interaction partners exploits the tendency for proteins in the same pathway to be co-inherited across organisms (Pellegrini *et al.*, 1999; Huynen *et al.*, 2000), as illustrated in Figure 2C. A protein sequence is compared to sequences of all known proteins from the available fully sequenced genomes. From this analysis, a phylogenetic profile is calculated, describing which genomes contain homologs of the query gene. This phylogenetic profile can then be compared to those of all other genes in a genome to suggest functional partners of the query gene, with variations on the approach using orthologs rather than homologs (Eisen and Wu, 2002), employing probabilistic measures of profile similarity (Wu *et al.*, 2003) or measuring mutual information between the phylogenetic profiles (Huynen *et al.*, 2000; Date and Marcotte, 2003), and taking into account the phylogenetic relationships among the genomes (Vert, 2002; Sun *et al.*, 2005).

3. The Rosetta Stone/gene fusion approach

In the third approach, illustrated in Figure 2D, two genes in one organism can be inferred to be functionally linked due to the discovery of a third gene, which is the fusion of the two separate genes (Enright *et al.*, 1999; Marcotte *et al.*, 1999a). Empirically, it appears that gene fusions typically occur only between genes of related function; consequently, this method can be used to rapidly suggest functional partners for a gene. As with the previous examples, this approach is rarely used without enforcing some statistical criteria (Enright *et al.*, 1999; Yanai *et al.*, 2001; Verjovsky *et al.*, 2002; Bowers *et al.*, 2004) to rule out spurious associations that arise from the tendency of certain 'promiscuous domains' to be found in multidomain proteins with many other protein domains (Marcotte *et al.*, 1999a). This approach has also recently been combined with the

gene neighbor approach by looking for conservation of protein domains as neighbors (i.e., ignoring whether the domains belong to the same protein or different proteins, merely requiring that they be adjacent on the chromosome, Pasek *et al.*, 2005).

♦♦♦♦♦♦ III. RECOGNIZING AND ASSESSING ERROR IN FUNCTIONAL GENOMICS DATA

It is a fact that all biological data, experimental or computational, low-throughput or high-, contain errors. An error-free experiment has yet to be conducted. It is clearly critical, then, to determine the extent of errors in both experiments and predictions. Are these large or small, and can they be measured precisely? Many of the original protein interaction and genetic interaction mapping experiments were presented only as experimental results, some of these experimental *tours de force*, but rarely (at first) accompanied by estimates of the error in these techniques (e.g., Chien *et al.*, 1991; Rigaut *et al.*, 1999). Much of our appreciation of the errors in these approaches (e.g., appreciation of the classes of false positive interactions identified in typical yeast two-hybrid assays) comes from empirical observations of many individual investigators (e.g., Bartel *et al.*, 1993; Estojak *et al.*, 1995) and from computational analyses performed after the original screens (e.g., Mrowka *et al.*, 2001; Deane *et al.*, 2002; von Mering *et al.*, 2002; Patil and Nakamura, 2005). Unfortunately, data for the failure rates of these techniques are rarely published, making it difficult to systematically measure error rates from small-scale assays (Jansen and Gerstein, 2004). For several of these approaches, specialized statistics have been developed to better measure the associated errors in specific large-scale screens, such as the prediction of true and false positive yeast two-hybrid interactions based upon internal properties of a large scale screen (Bader *et al.*, 2004).

It is clear that without any appreciation of the errors involved, one can be strongly misled as to the likely functions and interactions of a gene of interest. Functional and comparative genomics have the advantage that their errors can be measured and the data interpreted accordingly. In this section, we introduce the primary approaches for measuring error in these data, in this manner estimating how correct functional inferences drawn from these data are likely to be.

A. Reference Sets for Evaluating Functional Association

In predicting gene function by guilt by association, false predictions arise when associations are identified between functionally irrelevant or uncoupled genes. These false associations are generally products of non-biological variance in the data, even with extensive filtering of the data. Therefore, one of the more successful strategies

has been to rely upon external, independent data sets to act as a form of 'gold-standard' reference for assessing the primary data quality. With high-quality reference sets, we can then evaluate different data sets using a unified quality criterion.

As with any 'supervised' approach, the quality of the reference set is critical and directly determines the quality of evaluations performed with the set. A faulty reference set can lead to faulty evaluation and, consequently, to faulty biological inferences. We consider three aspects of the reference set quality: First, the size (or coverage) of the reference set should be large enough to offer a statistically reliable evaluation of the data. Second, the set should be of high accuracy, containing a minimum number of false examples. Although the evaluation methods we will discuss are noise tolerant, a significant amount of noise can interfere with the analysis and lead to erroneous conclusions. Third, the resolution (or specificity) of information must be sufficiently high for the types of associations to be drawn. Even with all true examples in the reference set, highly generalized associations will be of less utility. For example, the fact that two proteins both localize to the cytosol only weakly indicates that they might be functionally associated, as the association is too general in nature to draw a strong association.

Diverse associations between genes, indicating different functional associations, can be used as reference sets. For example, reference sets might consist of the following:

1. protein pairs sharing functional annotation(s),
2. protein pairs sharing pathway annotation(s),
3. protein pairs found in the same complex(es), and
4. protein pairs sharing cellular localization annotation(s).

Box 1 lists a variety of yeast gene annotation sets of these types that are useful as reference sets. (Whenever possible, the precise locations of data files are given in the box, rather than the project web site gateways.) These annotations vary in both specificity and coverage. Moreover, several of these are hierarchically organized, and choosing different levels of the annotation hierarchy may generate quite different evaluations for the same data set. Generally, top-level annotations provide extensive coverage but low-specificity (resolution), while low-level annotations decrease coverage but increase specificity. Therefore, the choice of an appropriate level(s) of hierarchical annotation must be considered carefully to achieve the optimal trade-off between coverage and specificity. Once chosen, the reference set should be consistent throughout the entire data analysis.

It is striking that the current yeast annotation and reference sets are quite non-overlapping. For example, fewer than half of the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database associations are also contained in the gene ontology (GO) annotation set (Bork *et al.*, 2004). The low overlap is primarily due to different data mining methods and to inclusion bias among the annotation sets. This, however, provides the opportunity to generate

Box 1. Data sets recommended for benchmarking functional predictions.

Functional annotation

GO (Gene ontology) biological process
http://www.geneontology.org/ontology/process.ontology
GO is hierarchically organized, with the top level annotation most general and the bottom level most specific. Generally, the middle range of annotation provides a reasonable reference set. GO is notable for only annotating genes with the lowest annotation available and not listing the (implied) higher level terms, requiring the user to reconstruct these implied terms for a complete reference set.

Eukaryotic clusters of orthologous groups (KOGs)
ftp://ftp.ncbi.nih.gov/pub/COG/KOG/kog
KOGs is a comprehensive but very general annotation set, with <30 different functional terms. Due to its availability for many organisms, it is valuable for assessing poorly studied organisms.

CYGD (the comprehensive yeast genome database) functional category, hosted by MIPS (Munich Information Center for Protein Sequences)
ftp://ftpmips.gsf.de/yeast/catalogues/funcat/
CYGD is a comprehensive and detailed annotation set specific for yeast.

Pathway annotation

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway
ftp://ftp.genome.jp/pub/kegg/pathways/sce/sce_gene_map.tab
KEGG offers a three-level hierarchical annotation of pathways.

Complex annotation

CYGD (the comprehensive yeast genome database) complex, hosted by MIPS (Munich Information Center for Protein Sequences)
ftp://ftpmips.gsf.de/yeast/catalogues/complexcat/

GO (Gene ontology) cellular components
http://www.geneontology.org/ontology/component.ontology
Note: includes both complexes and subcellular locations, with similar organization as GO biological process.

Cellular localization annotation

Yeast GFP fusion localization database (hosted by UCSF)
http://yeastgfp.ucsf.edu/
The most comprehensive experimental annotation for yeast protein sub-cellular localization. It has, however, <30 different location terms, ranging widely in specificity.

GO (Gene ontology) cellular components
http://www.geneontology.org/ontology/component.ontology
Note: includes both complexes and subcellular locations, with similar organization as GO biological process.

TRIPLES (TRansposon-insertion phenotypes, localization, and expression in *Saccharomyces*)
ftp://ygac.med.yale.edu/ygac_pub_ftp/localization_pub_data_9_4_01.tab
Another localization annotation set, generated by random transposon-tagging of yeast genes.

YPL.db (Yeast Protein Localization Database)
http://genome.tugraz.at/ypl.html

more comprehensive reference sets by combination of different annotation sets. In practice, both GO and KEGG have proven to be generally reliable for benchmarking data sets.

B. Evaluating Function Predictions Using a Reference Set

Having selected a reference, the next step in functional prediction is to select a method for evaluating functional predictions against the reference set. There are multiple methods to choose from, differing primarily in the parameters for measuring data quality. Two key parameters expressing data quality are *coverage* and *accuracy*. For example, we can measure coverage of the proteome (the complete set of yeast proteins) by the proteins in the data set, or we can measure coverage of the interactome itself. As we do not currently know any complete interaction map, or even the size of a complete interaction map, the latter measure of coverage cannot be expressed as a percentage of the whole, but only as a total number of interactions. The coverage also can be represented as a fraction of the reference interactions. The accuracy can be measured by as the percentage of interaction data confirmed by the reference set (von Mering *et al.*, 2002), or alternately, as the likelihood of being true associations (Jansen *et al.*, 2003; Lee *et al.*, 2004). In the latter approach, likelihood is usually calculated as a log of the odds ratio that the data are correct to incorrect. (For example, a 2:1 odds ratio of being correct:incorrect corresponds to a 67% chance of being correct.)

This likelihood approach is often based upon Bayesian statistics, which allow relatively straightforward calculation of such likelihood ratios. To evaluate a dataset, a log likelihood ratio (LLR) can be calculated as

$$LLR = \ln\left(\frac{P(D|I)}{P(D|\sim I)}\right)$$

where $P(D|I)$ and $P(D|\sim I)$ are the probability of observing the functional genomics data (D) conditioned on the genes being associated in the reference set (I) or not being associated in the reference set ($\sim I$). By Bayes' theorem, this equation can be rewritten as:

$$LLR = \ln\left(\frac{P(I|D)/P(\sim I|D)}{P(I)/P(\sim I)}\right)$$

where $P(I|D)$ and $P(\sim I|D)$ are the frequencies of functional associations observed in the given dataset (D) between annotated genes that are associated in the reference set (I) or not associated in the reference set ($\sim I$), respectively. $P(I)$ and $P(\sim I)$ represent the prior expectations (the total frequencies of all reference set genes being associated in the reference set or not, respectively). This latter version of the equation is simpler to compute. A score of zero indicates interaction partners in the dataset being tested are no more likely than random to belong to the same pathway or to interact; higher scores indicate a more accurate dataset.

Another variation to consider for benchmarking functional inferences depends on whether we desire a single measured value for characterizing an entire data set or instead use a continuous measurement to differentiate among different quality subsets within the data. A small data set with a fairly uniform probability distribution of error can be described by a single evaluation score for the entire data set. However, if we have a parameter associated with the data set that correlates with data accuracy, it is often better to sort the data by this parameter and measure the accuracy as a function of a certain parameter range. An example of this is given in Section IV.

One formal way to evaluate data by coverage and accuracy is to plot a recall–precision curve. *Recall* (defined as the percentage of positives in the reference set correctly predicted as positives in the data set) provides a measure of coverage and *precision* (defined as the percentage of predicted positives in the data set confirmed as true positives by the reference set) provides a measure of accuracy (Figure 3). In the case of using a log-likelihood scoring scheme, the log-likelihood score would be substituted for the accuracy or precision parameter, plotted in its place on the y-axis of the plot.

Another formal assessment method is the receiver operating characteristic (ROC) curve. The ROC curve is named from its popularity in communications research, where it is used to detect the 'hit rate' of true positives at a given cost of false alarms over a noisy communications channel. The ROC approach can be applied as well to biological data, but has several shortcomings. We consider the true positive rate (defined as 100 times the number of true positives divided by the number of positives in the reference set) as the true hit rate and the false positive rate (defined as 100 times the number of false positive divided by the number of negatives in the reference set) as the false hit rate. All curves in the ROC plot (each

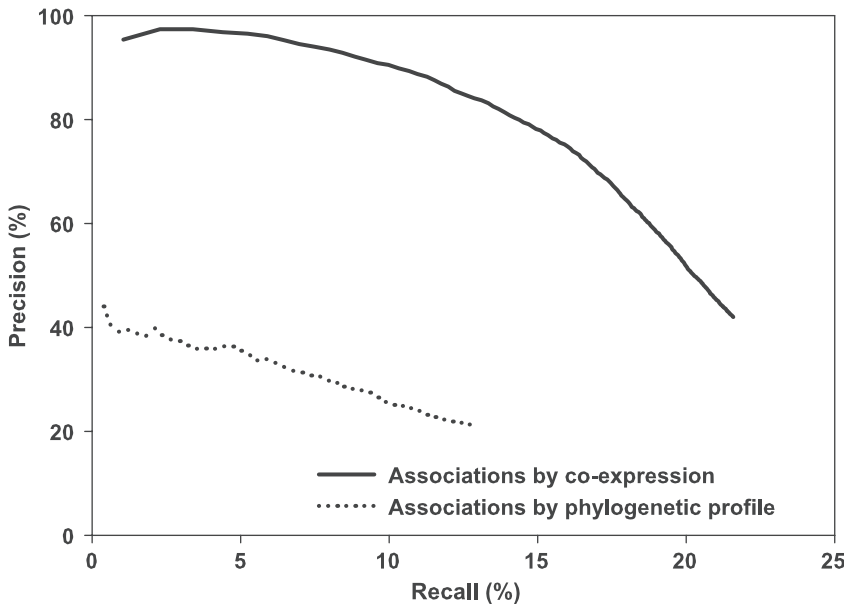

Figure 3. An example of a recall–precision curve, displaying the performance of mRNA co-expression and phylogenetic profiling at predicting functional associations. Recall (percentage of reference set positives correctly predicted as positive) and precision (percentage of predicted positives confirmed as positive by reference set) are based on a cumulative confusion matrix for the given recall level. For good predictions, we expect the evaluation curve to tend toward to the right upper corner. The given recall–precision curves indicate that these particular functional associations (Lee *et al.*, 2004) inferred from mRNA co-expression are better than those inferred from phylogenetic profiling, generally showing twice the precision at comparable levels of recall. In both Figures 3 and 4, KEGG pathways are used as the reference set (Kanehisa *et al.*, 2004).

corresponding to one data set) start at 0% in both the true- and false-positive rates and ultimately arrive at 100% in each, as illustrated in Figure 4. For a data set of randomly chosen associations, we would expect to observe the same rates of true and false positives, giving rise to a diagonal line. For a real data set, however, we hope to observe a higher rate of gain of true positives. The area under each ROC curve is therefore proportional to the quality of the data sets. The ROC analysis has a notable shortcoming – to be strictly comparable, each data set must be evaluated over the same reference set. So, for comparisons of multiple data sets, we have to define a subset of the original reference set common to all data sets in the plot, which can substantially diminish the size of the reference set used and which may even introduce a bias in the evaluation favoring specific data sets – it is possible that the *common subset* of reference associations has a significantly higher or lower proportion of matches for specific data sets than do the remaining set of associations.

A final benchmarking strategy that has proven useful is the modeling of a data set as a mixture of true and false positives

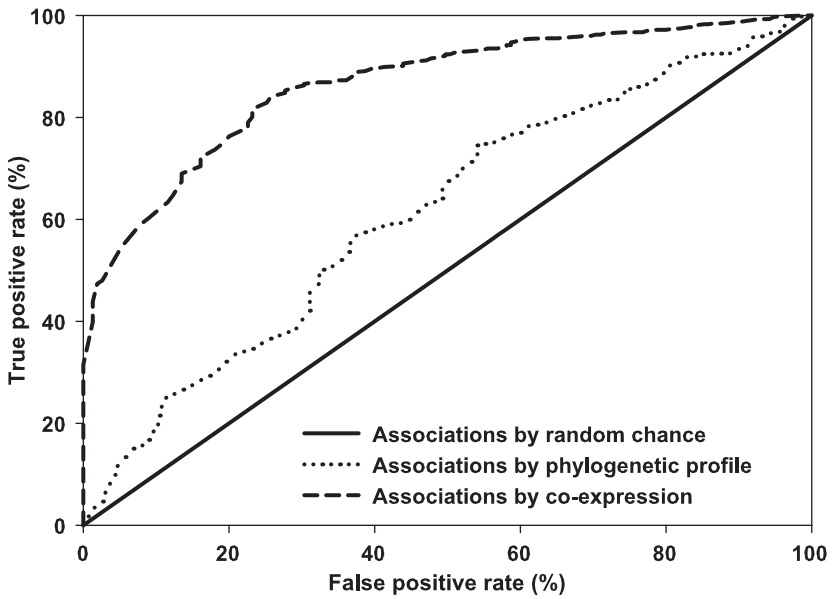

Figure 4. An example of a ROC curve, plotting the performance of mRNA co-expression and phylogenetic profiling at predicting functional associations. True positive rate (or true hit rate = percentage of reference set positives correctly predicted as positive) and false positive rate (or false hit rate = percentage of reference set negatives incorrectly predicted as positives) are plotted by ranking all predictions by their associated scores, then calculating the hit rates as a function of increasing rank. Random guesses generally show equal rates of true and false hits, generating a diagonal line in the ROC curve. Any prediction that is better than random guessing shows a curve above the diagonal. Overall performance can be estimated by the area under the ROC curve. For these particular data (the same data as in Figure 3), co-expression-based functional inferences out-perform phylogenetic profiling-based inferences.

(Mrowka *et al.*, 2001; Deane *et al.*, 2002). Therefore, the behavior of the data set on some external benchmark (such as testing the extent of correlation of the gene pairs' mRNA expression patterns) can be mathematically fit as a linear combination of the behaviors of two reference sets (a positive and a negative sets), with the fit percentage of false positives providing the false positive rate of the data. An example of this strategy is calculating error rates for yeast two-hybrid data based upon the co-expression of the genes across a bank of DNA microarray experiments. Random gene pairs show one distribution of co-expression, true positives show another, skewed distribution, and the two-hybrid set being tested shows a mixture of the two distributions (Deane *et al.*, 2002; Kemmeren *et al.*, 2002).

Figure 5 illustrates the benchmarking of a number of the functional genomics data sets introduced earlier, using the recall–precision analysis described above. The relative scores indicate the utility of these data sets for inferring functional associations between genes linked by the particular methods. The reference set

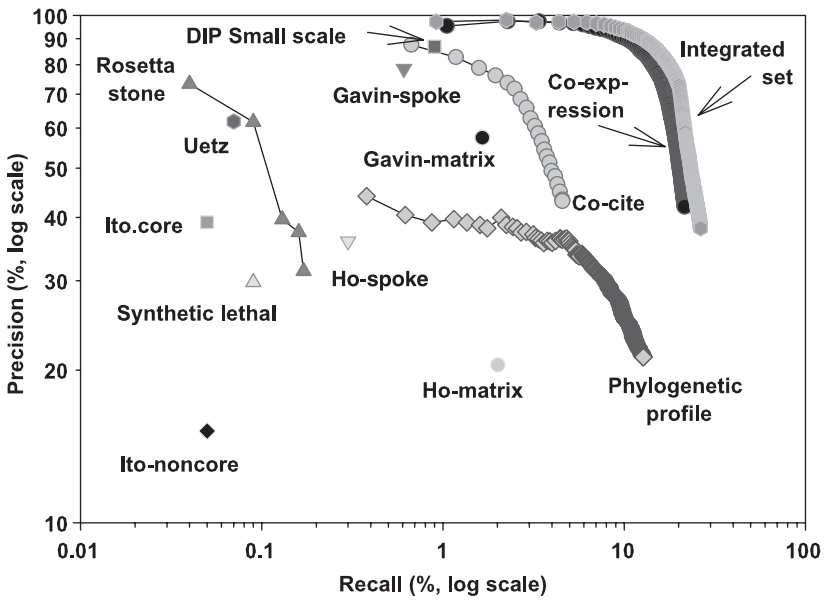

Figure 5. Benchmarking functional interactions from various experimental and computational approaches against the KEGG pathway reference set. The data plotted (analyzed as described in Lee *et al.*, 2004) include a collection of small scale experimental interaction data from the DIP (Small scale; Xenarios *et al.*, 2002), yeast proteins co-cited in Medline abstracts (co-cite), yeast genes whose mRNAs are co-expressed across 491 microarrays from the Stanford Microarray Database (co-expression; Gollub *et al.*, 2003), co-inherited yeast genes (phylogenetic profile; Date and Marcotte, 2003), yeast genes linked by gene fusion analysis (Rosetta stone), two spoke models and two matrix models of mass spectrometry data (Gavin-spoke, Gavin-matrix from Gavin *et al.*, 2002; and Ho-spoke, Ho-matrix from Ho *et al.* 2002), three yeast two-hybrid data sets (Uetz *et al.*, 2000; Ito-core, Ito-noncore from Ito *et al.*, 2001), and synthetic lethal interactions (synthetic lethal, Tong *et al.*, 2004). Data sets with associated continuous scores were evaluated in binned groups of 1000 gene pairs each. The reference set consists of ~49 000 positive examples, made by pairing genes belonging to the same KEGG pathway (Kanehisa *et al.*, 2004), and ~629 000 negative examples, made by pairing genes that do not belong to the same KEGG pathways. The integrated data set was generated by the weighted sum log-likelihood method of (Lee *et al.*, 2004).

was generated by pairs of genes belonging to the same KEGG pathways (~49 000 positive examples). Each data set is thus evaluated for its utility in inferring two genes belong to the same pathway. Better inferences tend toward the upper right corner of the plot. As can be seen, there is a wide spectrum of accuracy and coverage across the different data sets. For example, two proteins found to interact in the 'Uetz' yeast two-hybrid assay (Uetz *et al.*, 2000) are ~60% likely to belong to the same KEGG pathway, giving an indication of their general functional relatedness, while two genes that are synthetic lethal to each other in the Tong *et al*. (2004) data set are ~30% likely to belong to the same KEGG pathway. The best functional inference is accomplished by integrating the individual data sets.

♦♦♦♦♦♦ IV. A QUANTITATIVE ERROR MODEL FOR YEAST TWO-HYBRID, MASS SPECTROMETRY, AND OTHER INTERACTIONS

Most interaction data sets, such as yeast two hybrid (Uetz *et al.*, 2000; Ito *et al.*, 2001) and protein complex affinity purification followed by mass spectrometry (Gavin *et al.*, 2002; Ho *et al.*, 2002), are accompanied by the simple observation of 'interacting' or 'not interacting'. However, for the purposes of predicting gene function, we would prefer to have some more fine-grained measure of confidence in the interactions and in the value of the transferred function (as in the 'weighted' interaction model of Figure 1A). One particularly simple and general theoretical framework for error is based upon the hypergeometric distribution. Variants of this approach appear to work well for many linkage and interaction types (Verjovsky *et al.*, 2002; Samanta and Liang, 2003; Schlitt *et al.*, 2003). In this section, we present an approach for calculating the hypergeometric error model and demonstrate its effectiveness.

The essential notion is that when data is assembled from many interaction experiments, such as deriving from a large-scale protein interaction assay, proteins take on varied numbers of interaction partners, with some proteins interacting promiscuously with many partners. This may be due to errors in the interaction screen or may be a legitimate reflection of the multi-functionality for the given proteins. In either case, assignment of specific function by guilt-by-association is non-trivial, with our intuition that we probably want to assign less confident scores for these particular interactions. Following this line of reason, we can apply a statistical re-interpretation to the interaction data to assign different confidence scores for different interactions depending on their tendency to participate in many or few interactions. With this approach, even the matrix model of mass spectrometry data becomes highly informative, with subsets showing extensive coverage with reliable quality.

Rather than use these coarse-grained models, we suggest a simple hypergeometric error model for calculating the probability of two proteins interacting by random chance given their behavior in the large-scale screen, assigning a probability (p-value) to the pair as:

$$p(\#\text{interactions} \geq k|n,m,N) = \sum_{i=k}^{\min(n,m)} p(i|n,m,N)$$

where :

$$p(i|n,m,N) = \frac{\binom{n}{i}\binom{N-n}{m-i}}{\binom{N}{m}} = \frac{n!(N-n)!m!(N-m)!}{(n-i)!i!(m-i)!(N-n-m+i)!N!},$$

and where k is the number of experiments in which an interaction is observed between proteins A and B (e.g., the number of yeast

two-hybrid interactions or co-purifications involving both A and B), n the total number of experiments in which protein A is observed, m the total number of experiments in which protein B is observed, and N the total number of experiments with ≥ 1 interaction measured (e.g., the overall number of yeast two-hybrid interactions observed in a large-scale two-hybrid experiment, or the overall number of pull-down experiments with at least one interaction partner observed in a large-scale affinity purification mass spectrometry experiment). The resulting probability score indicates how likely the specific interaction between proteins A and B was to be observed by random chance given the other interactions observed in the screen, in effect taking into account how promiscuous A and B are in their interactions.

Figure 6 demonstrates the dramatic improvement in the confidence in yeast protein–protein interactions under this hypergeometric error model, using as examples a genome scale yeast two-hybrid experiment data set (Ito *et al.*, 2001) and interacting protein data from affinity purifications followed by mass spectrometry (Gavin *et al.*, 2002). The cumulative prediction accuracy with the hypergeometric error model shows equivalent power to the whole data set under the matrix model. However, the model achieves a much higher resolution of information, separating out the high- and low-quality interactions in the set, leading to more powerful data integration and interpretation. As discussed in the earlier sections, the 'spoke' model provides more accurate but less extensive interactions, while the 'matrix' model increases coverage at the expense of accuracy.

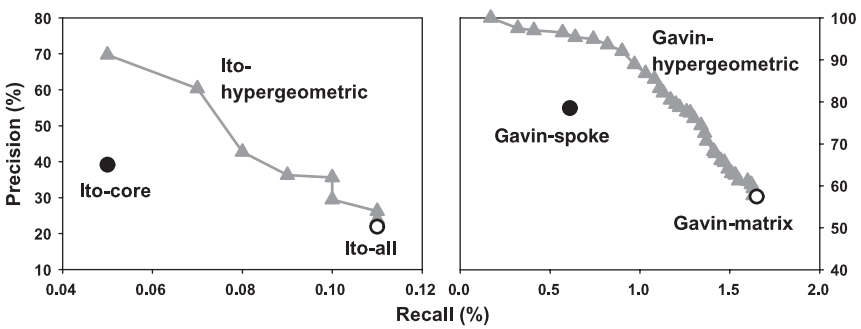

Figure 6. Re-interpreting interaction data sets using the hypergeometric error model. Large-scale yeast two-hybrid interaction sets and mass spectrometry interaction sets generally provide binary 'scores' (observed or not-observed). The hypergeometric error model will generate continuous accuracy scores for these data. These scores are illustrated for yeast two hybrid data (Ito *et al.*, 2001) and mass spectrometry interaction data (Gavin *et al.*, 2002). Recall–precision curves for the error models are plotted using bins of 500 gene pairs. For the two-hybrid data, the core data set ('Ito-core') is more accurate than the entire data set ('Ito-all'). The hypergeometric model shows substantially higher accuracies for the same level of coverage. For the mass spectrometry data, the 'spoke' model ('Gavin-spoke') is more accurate, but has lower coverage, than the 'matrix' model ('Gavin-matrix'). The hypergeometric model is considerably more accurate than the spoke model at the same level of coverage, and eventually converges on the matrix model.

However, the hypergeometric error model can differentiate the reliable interactions from the others in the original experimental data, resulting in generally more extensive coverage at accuracy equal to the spoke model. Note that this approach is quite generic, and can be applied to many different forms of linkages, interactions, and predicted linkages.

♦♦♦♦♦♦ V. STRONGER INFERENCES VIA DATA INTEGRATION

Data integration for predicting gene function has recently drawn considerable attention for a couple of major reasons. First, no single experimental or computational approach has proven able to analyze the yeast gene set to completion, due either to biases in the system or to increasing technical difficulties in taken screens to 100% of the genes/proteins. Second, different approaches show bias in the sets of genes and classes of functional inference that they are more useful for. Integrating inferences from multiple data types is therefore mandated. Agreement between inferences drawn from distinct data sets also increases our confidence in their correctness. In general, we expect better functional inferences when integrating multiple data sets than from even the best single data set. There are many different approaches for data integration, ranging from quick-and-dirty methods, such as the voting method, to more statistically empowered approaches, such as Bayesian networks.

A. Simple Voting Methods

There are many variations to the voting approach. If we have only binary (interact or not interact) inferences from the given data set, an intersection or a union method can be adopted (e.g., as in Marcotte *et al.*, 1999b; Mellor *et al.*, 2002). While the former takes a subset of the data with support from multiple data sets after applying suitable thresholds for each analysis, the latter takes a combined set with support from at least one data set. The 'intersection' approach works best with redundant data sets and gives the better integration of the two while the 'union' approach performs well when relatively independent data sets are involved and tends to give improved integration without significant loss in accuracy. If the data sets have associated confidence scores, the method could be extended by taking the maximum or the averaged scores of multiple lines of evidence to generate an integrated data set.

B. Bayesian Statistical Methods

The foremost goal of a formal statistical approach to data integration entails the normalization of scores from different data sets with

different features. Normalizing all the data sets for integration can be performed using the reference sets discussed earlier. The final integrated data will be biased and characterized by the types of information implemented in it. For example, if we want to learn protein–protein physical interactions using an integrated data set, our evaluation of the pre-integration data sets should be based on a reference set of protein–protein physical interaction.

One of the common approaches to data integration is using a Bayesian network (e.g., as in Jansen *et al.*, 2003; Troyanskaya *et al.*, 2003). In this approach, a graph is created with each node (vertex) representing a specific condition (a combination of data set and parameter) and probability of true instance. If two different conditions are not completely independent (i.e., if A happens, B tends to happen as well), they are connected with some probability. In this case, the construction of the graph is not entirely straightforward. Finding the correct degree of dependence between conditions is critical for choosing an appropriate model for integration.

There are two distinct styles of Bayesian network approach. In the fully connected Bayesian network, we assume that all conditions are connected, and find a model for the best integration by evaluating all possible combinations of conditions. For a simple model, this approach is feasible, but for a complex one, the problem of exponentially increasing combinations precludes this approach. Biological data, being highly complex, often face this problem. In an attempt to simplify the data integration, a naïve Bayesian network approach has been adopted in some instances. This simpler approach, however, assumes that the different data sets are completely independent. Interestingly, such simple naïve Bayesian network approaches have performed well in functional genomics data integration (Jansen *et al.*, 2003; Troyanskaya *et al.*, 2003; Patil and Nakamura, 2005), which may indicate that existing genomics data are relatively sparse, with only minimal overlap between different data sets.

Into the future, however, we expect genomics data will continue to increase and the problem of measuring correlation between different data sets will have to be addressed. Therefore, it is imperative to identify an efficient way to handle information redundancy for integration of large data sets. For example, one simple but incomplete approach to this is the weighted sum method (Lee *et al.*, 2004), which has the merit of simplicity but at the cost of only treating a subset of correlations in the data sets. An example of integration by this approach is shown in Figure 5.

C. Other Methods

A number of more complex probabilistic and kernel-based supervised approaches have been proposed (e.g., see Yamanishi *et al.*,

2004 for kernel canonical correlation analysis, Jiang and Keating (2005) for a combined approach using probabilistic assignment of function combined with decision trees, and Tanay *et al.* (2004) for graph partitioning based approaches) that reliably assign biological function from combinations of high-throughput data.

Note that we have taken a fairly narrow view of data integration for function prediction here, focusing primarily on the guilt-by-association method as determined from network models of gene associations. However, other methods worth noting include models of data integration that are completely independent of network structure, such as those based on learning association rules by training on reference sets (Clare and King, 2003), as well as approaches for predicting genetic interactions (Wong *et al.*, 2004) and diverse approaches for learning specific ontology descriptions or gene annotations for uncharacterized genes, such as by the effective propagation of information through gene networks (see, e.g., Vazquez *et al.*, 2003; Deng *et al.*, 2004a, b; Karaoz *et al.*, 2004; Nabieva *et al.*, 2005) or through algorithmic classifiers (Pavlidis *et al.* 2002).

♦♦♦♦♦♦ VI. METHODS AND PROTOCOLS FOR EMPLOYING PRE-CALCULATED FUNCTIONAL PREDICTIONS

In this section, we highlight several of the main functional- and comparative-genomics based integrated models of yeast gene function. Each is available either as binary functional linkages between yeast genes or in the form of an internet server for interactively searching for functional linkages (Box 2).

A. AVID (Jiang and Keating, 2005)

URL: http://web.mit.edu/biology/keating/AVID/

(1) The user can enter the gene name or the GO ID in the query box.
(2) AVID is a computational function prediction framework that integrates data from high-throughput experiments, such as yeast two-hybrid, mass spectrometry, DNA microarrays, protein localization data, and protein sequence similarity.
(3) AVID initially assumes that the query gene interacts with all other genes, then successively filters low-confidence partners from the list based on varied criteria, resulting in a list of functional interactions with candidate genes at the level of three GO categories – molecular function, biological process, and cellular component.

Box 2. Predictions of yeast gene function via integrating multiple datasets.

AVID (Jiang and Keating, 2005)
http://web.mit.edu/biology/keating/AVID/

FinalNet (Lee *et al.*, 2004)
Online supplemental data with paper, also available for download from SGD database: ftp://genome-ftp.stanford.edu/pub/yeast/data_download/systematic_results/published_computational_predictions/lee_pmid_15567862/

LIANG (Samanta and Liang, 2003)
http://www.systemix.org/PP/partners/index.php

MAGIC (Troyanskaya *et al.*, 2003)
http://genome-www.stanford.edu/magic/

PIT (Jansen *et al.*, 2003)
http://networks.gersteinlab.org/intint/index-2.html

PLEX (Date and Marcotte, 2005)
http://bioinformatics.icmb.utexas.edu/plex/

Prediction of CCPs (Zhang *et al.*, 2004)
Online supplemental data with paper

PREDICTOME (Mellor *et al.*, 2002)
http://predictome.bu.edu/index.php

PROLINKS (Bowers *et al.*, 2004)
http://dip.doe-mbi.ucla.edu/pronav

STRING (von Mering *et al.*, 2005)
http://string.embl.de/

B. FinalNet (Lee *et al.*, 2004)

Available as supplemental data from the on-line publication, or from the following URL:
ftp://genome-ftp.stanford.edu/pub/yeast/data_download/systematic_results/published_computational_predictions/lee_pmid_15567862/

(1) This dataset awaits an interactive web feature. It can be readily accessed, however, through the above *Saccharomyces* genome database (SGD) FTP site.
(2) The user can download the yeast_FinalNet.txt.gz and unzip the file as a text (.txt) file.
(3) This file has comprehensive listings of the interacting partners of all yeast genes computed by an integrated Bayesian formalism. The probabilistic functional prediction is calculated by computing

a LLS for each functional genomics dataset to arrive at an integrated LLS score. The datasets include mRNA co-expression, Rosetta Stone gene fusions, phylogenetic profiles, literature mining (co-citation), and genetic and protein interaction experiments. This initial integrated network is further rescored by identifying linkages that are dependent on the network context. A final integrated network is arrived at, on which is applied a threshold based upon 'gold-standard' small-scale protein interaction assays. The resulting 'ConfidentNet' constitutes 34 000 linkages between 4681 genes. Further hierarchical clustering yields a network with 627 modules of functionally linked genes spanning 3285 genes, giving an estimate of protein systems in a yeast cell.

C. LIANG (Samanta and Liang, 2003)

URL: http://www.systemix.org/PP/partners/index.php

(1) The user can enter the name of the query gene in the query box. The server will return candidate interacting partners.
(2) This function prediction works on the hypothesis that two proteins sharing common interacting partners more than just by random chance are likely to interact with each other with higher probability. The known interacting proteins from DIP (Salwinski *et al.*, 2004)) are taken to illustrate this hypothesis. Further analyses of the interacting partners of these proteins are done to generate the complete dataset.

D. MAGIC (Troyanskaya et *al.*, 2003)

URL: http://genome-www.stanford.edu/magic/

Currently, the Multisource Association of Genes by Integration of Clusters (MAGIC) dataset does not have an interactive web interface. However the user can access the final output at the above internet site (also available from the FTP download site of the SGD database). The predicted interactions for a given gene can be seen as different GO category IDs at different stringencies or cut-off scores (cut-off 0.9 and cut-off 0.75 available at the FTP site)

(1) MAGIC defines gene pairs having 'functional' relationships (i.e., involved in the same biological processes, defining processes as in the GO database).
(2) The Bayesian network that is implemented in MAGIC draws from known protein–protein interactions in the GRID database and protein–DNA interactions in the promoter database of *Saccharomyces cerevisiae* (Zhu and Zhang, 1999). Expression data is also incorporated. A score for each data source is calculated that represents the strength of each method's confidence in the

existence of a relationship between a gene pair. A combined probability score is then calculated.

E. PIT (Jansen *et al.*, 2003)

URL: http://networks.gersteinlab.org/intint/index-2.html

(1) A web interface exists to search these physical interaction predictions, in which a user can paste or type in the systematic name of the query gene and be given predicted interaction partners.
(2) Briefly, a probabilistic combination of multiple datasets is used to predict physical protein–protein interactions. The protein complexes represented in the MIPS database are used as the positive 'gold-standard dataset' whereas a negative gold-standard dataset consisting of protein pairs in separate sub-cellular compartments are used to train the Bayesian network. Two separate probabilistic interactomes (PIs) are constructed, one called the PIP (PI predicted) uses genomic context, mRNA co-expression, GO process, MIPS function and gene essentiality as datasets and the other called PIE (PI experimental) that uses data from co-immunoprecipitation/mass spectrometry and yeast two-hybrid experiments. These two interactomes along with the gold-standards are further integrated to a total PI (PIT) that represents a comprehensive view of known and predicted protein complexes in yeast.
(3) The user can choose a likelihood ratio cut-off (Lcut) as well as choose the PI version to be browsed for a particular analysis. An Lcut of 600 defines a threshold where a given protein pair is predicted to exist in the same complex with a better than 50% chance and therefore serves as a useful default.
(4) The output lists the proteins interacting with the query in an easily interpretable tabular form.

F. PLEX (Date and Marcotte, 2005)

URL: http://bioinformatics.icmb.utexas.edu/plex/

(1) PLEX returns functional associations predicted from phylogenetic profiles, operon neighbors, and gene fusions. Click on 'Submit a new job' and paste a protein sequence into the Protein Link Explorer (PLEX) query box or its Genbank GI number from the NCBI non-redundant database.
(2) Once the sequence is obtained, PLEX proceeds to compare the query sequence against ~350 000 proteins from 89 fully sequenced genomes using BLAST (Altschul *et al.*, 1997) sequence alignments. A phylogenetic profile is constructed from BLAST scores of the top-matching homologs in each genome. The

profile is displayed as a series of colored boxes, with blue indicating absence of the query in a given genome as opposed to red, which indicated varying degrees of confidence with which the query is present in different genomes.
(3) After a profile is created, the user can compare it to the profiles of all known proteins in any of the 89 genomes by entering a mutual information score cut-off for the comparison of phylogenetic profiles.
(4) PLEX searches are iterative: the first search associates the query gene (which may be from any source) with genes from the database genomes; successive iterations return pre-calculated associations (including operon neighbors and gene fusions) among database genes.

G. Predictome (Mellor *et al.*, 2002)

URL: http://predictome.bu.edu/index.php

(1) Predictome contains pre-calculated associations among genes from yeast and other organisms. The associations are derived from a variety of functional and comparative genomics approaches. It can be searched via a web interface, in which the user can paste the gene name (derived from SGD) into the query box to get the proteins predicted to be associated with the query protein.
(2) The output yields putative protein links with the query protein by integrating both experimental (yeast two-hybrid, co-immunoprecipitation and co-expression) and computational (phylogenetic profiles, gene fusion, and gene neighbor) datasets.
(3) The user can further view the interaction results in a network format using the VISANT applet.

H. PROLINKS (Bowers *et al.*, 2004)

URL: http://128.97.39.94/cgi-bin/functionator/pronav

(1) Enter the GI number of the query protein or the gene name and the genome to query.
(2) The user can then choose the sequence ID corresponding to the query gene name.
(3) The annotation page that is displayed lists general information such as sequence, chromosome, and existing annotation. Additional tabs in the page allow the user to display the homologs, the functional linkages 'PROLINKS' or to a graph depicting the protein interaction network involving the query protein.
(4) PROLINKS arrives at the output by computing the phylogenetic profile, Rosetta Stone, gene-neighbor and gene-cluster scores of a query protein. The user can further set a minimum confidence threshold to test for the strength of interactions.

I. STRING (von Mering *et al.*, 2005)

URL: http://string.embl.de/

(1) Enter the SWISS-PROT identifier of the query protein or its amino acid sequence in the query box.
(2) A search can be performed in the 'COG' mode which links with the COG database to classify the query protein to a particular COG category in an all-or-none fashion. A 'protein' mode search can also be done that relies on a matrix of pre-computed all-against-all protein similarity scores consisting of ~750 000 proteins.
(3) Once the query protein is associated with a COG category or with partner proteins as the case may be, search tool for the retrieval of interacting genes/proteins (STRING) computes a 'combined' score from seven individual sub-scores using a Bayesian prediction scheme. These include gene neighbors, gene fusions, phylogenetic profiles, mRNA co-expression, large-scale experiments (e.g., yeast two-hybrid), database imports that constitute previous knowledge, and literature mining (co-citation). This 'joint' probability score is often of higher confidence as compared to individual sub-scores.
(4) A final result is displayed as a network depiction with the individual sub-scores and the combined scores for every functional prediction. These scores serve to give a quick insight into the possible nature of interactions of a query protein, especially for proteins that are un-annotated.

♦♦♦♦♦♦ VII. AN EXAMPLE APPLICATION TO THE PARTIALLY CHARACTERIZED GENE *PRP43*

As an example, we will now take *PRP43* (*YGL120C*), a yeast RNA helicase, through several of the function prediction databases discussed above. *PRP43* is an essential yeast DEAH box protein, one of the family of proteins thought to possess RNA helicase activity and that function extensively in the coordination and catalysis of the pre-mRNA splicing reaction. *PRP43* itself, although not shown to exhibit an RNA helicase activity, was implicated to function in the disassembly of the U2/U5.U6 spliceosomal complex post-catalysis and subsequent release of the lariat RNA (Staley and Guthrie, 1998; Martin *et al.*, 2002) .

Figure 7 shows the functional associations of *PRP43* predicted from the different methods introduced in Section VI. The genes linked to *PRP43* are labeled according to their broad functional categories (based upon their GO annotations and literature) – 'mRNA processing', including mRNA transactions, such as pre-mRNA splicing; 'rRNA processing and biogenesis', constituting the

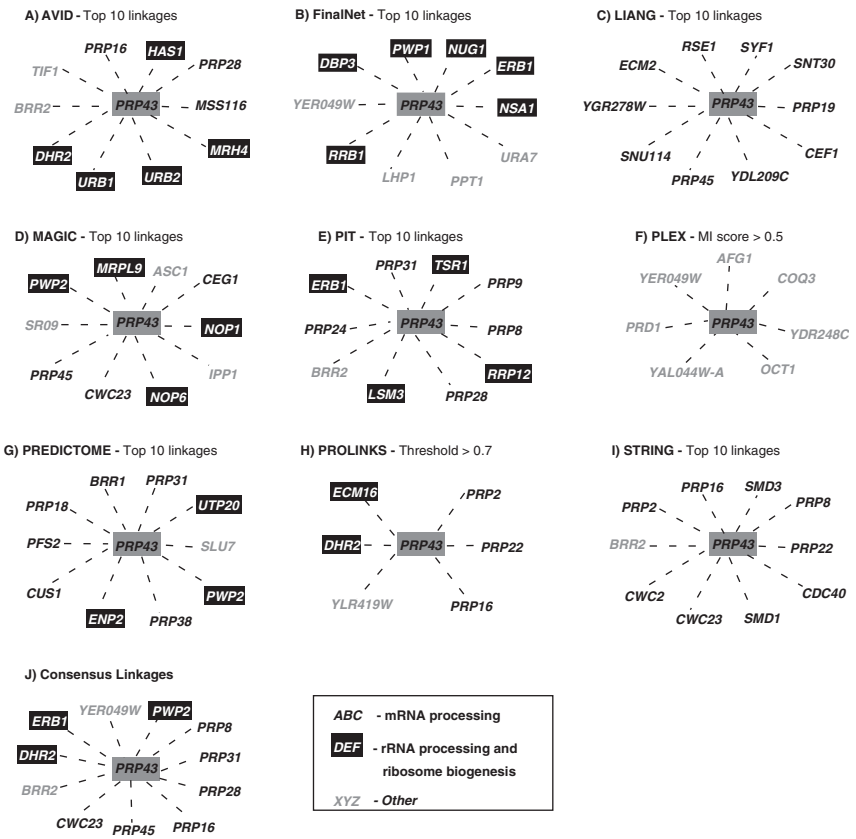

Figure 7. Functional associations of *PRP43* predicted by seven algorithms. Default parameters were used for querying the gene in each of seven functional prediction internet servers, and in each case, only the top 10 functional associations are shown (or fewer, if less than 10 are returned passing the default confidence threshold). Genes involved in mRNA processing are labeled in black text with white background, gene involved in rRNA processing and ribosome biogenesis are labeled in white fonts on black background, and gene of other functions are labeled in gray text on a white background. (A) Depicts the top 10 linkages as predicted by AVID, (B) depicts the top 10 linkages as predicted by FinalNet, (C) depicts the top 10 linkages as predicted by Liang *et al.*, (D) depicts the top 10 linkages as predicted by MAGIC, (E) depicts the top 10 linkages as predicted by PIT , (F) depicts the top 10 linkages as predicted by PLEX, (G) depicts the top 10 linkages predicted by PREDICTOME, (H) depicts all the linkages predicted by PROLINKS at a threshold greater than 0.7, (I) depicts the top 10 linkages predicted by STRING, and (J) depicts the consensus linkages that are predicted by more than one server, considering only their top 10 linkages.

synthesis, processing, and assembly of rRNAs into ribosomes; and 'other', referring to functions besides these two categories.

Although it may appear at first glance in Figure 7 that the algorithms are returning different predictions, most of the apparent disagreement is simply a function of including only the top 10 associations from each – as the methods include different data types and have different scoring functions, they tend to exhibit trivial

differences in the ranking of associated genes, and we have omitted the full set of predictions from many methods for reasons of space, concentrating only on the top 10 predictions per method. Nonetheless, although the specific predictions vary, certain linkages are identified by multiple algorithms, such as the linkages of *PRP43* to *ERB1*, *YER049W*, *CWC23*, *PWP2*, *PRP2*, *PRP28*, *DHR2*, *PRP8*, *BRR2*, *PRP31*, and *PRP16* (see Figure 7J). An examination of the broad functional categories of these predicted genes shows that the majority lie in mRNA processing such as pre-mRNA splicing, the known function of *PRP43* (Martin *et al.*, 2002). However, a number of associations are inferred with genes implicated in rRNA processing and ribosome biogenesis. This observation turns out to be in agreement with very recent data (Lebaron *et al.*, 2005; Combs *et al.*, 2006; Leeds *et al.*, 2006) that indicate that *PRP43* serves an essential role in the biogenesis of both ribosome subunits while having a non-essential role in pre-mRNA processing. Thus, the two primary functions of *PRP43* are correctly inferred by the algorithms. Although further experiments would be needed to verify the exact candidates involved in this process with *PRP43* and their manner of involvement, it is clear that integrated function prediction databases can be immensely valuable at generating new and testable hypotheses.

Acknowledgment

This work was supported by grants from the N.S.F. (IIS-0325116, EIA-0219061, 0241180), N.I.H. (GM06779-01), Welch (F1515) and a Packard Fellowship (E.M.M.).

References

Abhiman, S. and Sonnhammer, E. L. (2005). Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins* **60**, 758–768.

Altschul, S. F. *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.

Aravind, L. (2000). Guilt by association: contextual information in genome analysis. *Genome Res.* **10**, 1074–1077.

Bader, G. D., Betel, D. and Hogue, C. W. (2003). BIND: the biomolecular interaction network database. *Nucleic Acids Res.* **31**, 248–250.

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

Bader, J. S., Chaudhuri, A., Rothberg, J. M. and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* **22**, 78–85.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Nagu, W. C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W. and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Res.* **33**, D562–D566.

Bartel, P., Chien, C. T., Sternglanz, R. and Fields, S. (1993). Elimination of false positives that arise in using the two-hybrid system. *Biotechniques* **14**, 920–924.
Bateman, A. *et al.* (2004). The Pfam protein families database. *Nucleic Acids Res* **32**, D138–D141.
Blaschke, C., Andrade, M. A., Ouzounis, C. and Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. Int. Conf Intell. Syst. Mol. Biol.* 60–67.
Bork, P. *et al.* (2004). Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14**, 292–299.
Bork, P. and Koonin, E. V. (1998). Predicting functions from protein sequences – where are the bottlenecks?. *Nat. Genet.* **18**, 313–318.
Bowers, P. M. *et al.* (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* **5**, R35.
Breitkreutz, B. J., Stark, C. and Tyers, M. (2003). The GRID: the general repository for interaction datasets. *Genome Biol.* **4**, R23.
Bulyk, M. L., Huang, X., Choo, Y. and Church, G. M. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. USA* **98**, 7158–7163.
Chien, C. T., Bartel, P. L., Sternglanz, R. and Fields, S. (1991). The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. USA* **88**, 9578–9582.
Clare, A. and King, R. D. (2003). Predicting gene function in *Saccharomyces cerevisiae*. *Bioinformatics* **19**(Suppl 2), II42–II49.
Combs, D. J., Nagel, R. J., Ares, M. J. and Stevens, S. W. (2006). Prp43p is a DEAH-box spliceosome disassembly factor essential for ribosome biogenesis. *Mol. Cell. Biol.* **26**, 513–522.
Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328.
Date, S. V. and Marcotte, E. M. (2003). Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **21**, 1055–1062.
Date, S. V. and Marcotte, E. M. (2005). Protein function prediction using the Protein Link EXplorer (PLEX). *Bioinformatics* **21**, 2558–2559.
Deane, C. M., Salwinski, L., Xenarios, I. and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high-throughput observations. *Mol. Cell. Proteomics* **1**, 349–356.
Deng, M., Chen, T. and Sun, F. (2004a). An integrated probabilistic model for functional prediction of proteins. *J. Comput. Biol.* **11**, 463–475.
Deng, M., Tu, Z., Sun, F. and Chen, T. (2004b). Mapping gene ontology to proteins based on protein–protein interaction data. *Bioinformatics* **20**, 895–902.
Eisen, J. A. (1998a). A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res.* **26**, 4291–4300.
Eisen, J. A. (1998b). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8**, 163–167.
Eisen, J. A. and Wu, M. (2002). Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor. Popul. Biol.* **61**, 481–487.
Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868.

Eisenberg, D., Marcotte, E. M., Xenarios, I. and Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature* **405**, 823–826.
Engelhardt, B. E., Jordan, M. I., Muratore, K. E. and Brenner, S. E. (2005). Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput. Biol.* **1**, e45.
Enright, A. J., Iliopoulos, I., Kyrpides, N. C. and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90.
Estojak, J., Brent, R. and Golemis, E. A. (1995). Correlation of two-hybrid affinity data with *in vitro* measurements. *Mol. Cell. Biol.* **15**, 5820–5829.
Fetrow, J. S. and Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* **281**, 949–968.
Gavin, A. C. *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147.
Giaever, G. *et al.* (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.
Godzik, A. (2003). Fold recognition methods. *Methods Biochem. Anal.* **44**, 525–546.
Gollub, J. *et al.* (2003). The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* **31**, 94–96.
Habeler, G. *et al.* (2002). YPL.db: the yeast protein localization database. *Nucleic Acids Res.* **30**, 80–83.
Harbison, C. T. *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104.
Ho, Y. *et al.* (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183.
Honig, B. (1999). Protein folding: from the levinthal paradox to structure prediction. *J. Mol. Biol.* **293**, 283–293.
Huh, W. K. *et al.* (2003). Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691.
Humphreys, K., Demetriou, G. and Gaizauskas, R. (2000). Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pac. Symp. Biocomput.* 505–516.
Huynen, M., Snel, B., Lathe, W., 3rd and Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10**, 1204–1210.
Huynen, M. A., Snel, B., von Mering, C. and Bork, P. (2003). Function prediction and protein networks. *Curr. Opin. Cell Biol.* **15**, 191–198.
Ito, T. *et al.* (2000). Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. USA* **97**, 1143–1147.
Ito, T. *et al.* (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574.
Iyer, V. R. *et al.* (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538.
Jansen, R. *et al.* (2003). A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302**, 449–453.
Jansen, R. and Gerstein, M. (2004). Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.* **7**, 535–545.

Jiang, T. and Keating, A. E. (2005). AVID: an integrative framework for discovering functional relationships among proteins. *BMC Bioinform.* **6**, 136.
Kanehisa, M. *et al.* (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**(Database issue), D277–D280.
Karaoz, U. *et al.* (2004). Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. USA* **101**, 2888–2893.
Kelley, R. and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566.
Kemmeren, P. *et al.* (2002). Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell.* **9**, 1133–1143.
Kumar, A. *et al.* (2002). The TRIPLES database: a community resource for yeast molecular biology. *Nucleic Acids Res.* **30**, 73–75.
Lebaron, S., Froment, C., Fromont-Racine, M., Rain, J. C., Monsarrat, B., Caizergues-Ferrer, M. and Henry, Y. (2005). The splicing ATPase prp43p is a component of multiple preribosomal particles. *Mol. Cell. Biol.* **25**, 9269–9282.
Lee, I., Date, S. V., Adai, A. T. and Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558.
Lee, T. I. *et al.* (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804.
Leeds, N. B. *et al.* (2006). The splicing factor Prp43p, a DEAH box ATPase, functions in ribosome biogenesis. *Mol. Cell. Biol.* **26**, 513–522.
Madabushi, S. *et al.* (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
Marcotte, E. M. *et al.* (1999a). Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753.
Marcotte, E. M. *et al.* (1999b). A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86.
Marcotte, E. M., Xenarios, I. and Eisenberg, D. (2001). Mining literature for protein–protein interactions. *Bioinformatics* **17**, 359–363.
Martin, A., Schneider, S. and Schwer, B. (2002). Prp43 is an essential RNA-dependent ATPase required for release of lariat-intron from the spliceosome. *J. Biol. Chem.* **277**, 17743–17750.
Mellor, J. C. *et al.* (2002). Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.* **30**, 306–309.
Mrowka, R., Patzak, A. and Herzel, H. (2001). Is there a bias in proteome research?. *Genome Res.* **11**, 1971–1973.
Mukherjee, S. *et al.* (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36**, 1331–1339.
Nabieva, E. *et al.* (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* **21**(Suppl 1), i302–i310.
Oliver, S. (2000). Guilt-by-association goes global. *Nature* **403**, 601–603.
Overbeek, R. *et al.* (1999). The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896–2901.
Pasek, S. *et al.* (2005). Identification of genomic features using microsynthesis of domains: domain teams. *Genome Res.* **15**, 867–874.

Patil, A. and Nakamura, H. (2005). Filtering high-throughput protein–protein interaction data using a combination of genomic features. *BMC Bioinform.* **6**, 100.
Pavlidis, P., Weston, J., Cai, J. and Noble, W. S. (2002). Learning gene functional classifications from multiple data types. *J. Comput. Biol.* **9**, 401–411.
Pazos, F. and Valencia, A. (2002). *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219–227.
Pellegrini, M. *et al.* (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288.
Ponting, C. P. (2001). Issues in predicting protein function from sequence. *Brief. Bioinform.* **2**, 19–29.
Proux, D., Rechenmann, F. and Julliard, L. (2000). A pragmatic information extraction strategy for gathering data on genetic interactions. *Proc. Int. Conf. Intell Syst. Mol. Biol.* **8**, 279–285.
Ren, B. *et al.* (2000). Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309.
Rigaut, G. *et al.* (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032.
Salgado, H. *et al.* (2004). RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* **32**, D303–D306.
Salgado, H., Moreno-Hagelsieb, G., Smith, T. F. and Collado-Vides, J. (2000). Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA* **97**, 6652–6657.
Salwinski, L. *et al.* (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32**(Database issue), D449–D451.
Samanta, M. P. and Liang, S. (2003). Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci. USA* **100**, 12579–12583.
Schlitt, T. *et al.* (2003). From gene networks to gene function. *Genome Res.* **13**, 2568–2576.
Schonbrun, J., Wedemeyer, W. J. and Baker, D. (2002). Protein structure prediction in 2002. *Curr. Opin. Struct. Biol.* **12**, 348–354.
Slonim, D. K. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.* **32**(Suppl), 502–508.
Snel, B., Bork, P. and Huynen, M. A. (2002). The identification of functional modules from the genomic association of genes. *Proc. Natl. Acad. Sci. USA* **99**, 5890–5895.
Staley, J. P. and Guthrie, C. (1998). Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**, 315–326.
Sun, J. *et al.* (2005). Refined phylogenetic profiles method for predicting protein–protein interactions. *Bioinformatics* **21**, 3409–3415.
Tamames, J., Casari, G., Ouzounis, C. and Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66–73.
Tanay, A., Sharan, R., Kupiec, M. and Shamir, R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. USA* **101**, 2981–2986.

Thomas, J., *et al*. (2000). Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.* 541–552.
Tong, A. H. *et al.* (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368.
Tong, A. H. *et al.* (2004). Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813.
Troyanskaya, O. G. *et al.* (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. USA* **100**, 8348–8353.
Uetz, P. *et al.* (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627.
Vazquez, A., Flammini, A., Maritan, A. and Vespignani, A. (2003). Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol.* **21**, 697–700.
Verjovsky Marcotte, C. J. and Marcotte, E. M. (2002). Finding functionally linked proteins from gene fusions with confidence. *Appl. Bioinform.* **2**, 93–100.
Vert, J. P. (2002). A tree kernel to analyse phylogenetic profiles. *Bioinformatics* **18**(Suppl 1), S276–S284.
von Mering, C. *et al.* (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403.
von Mering, C. *et al.* (2005). STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**(Database issue), D433–D437.
Wolfe, C. J., Kohane, I. S. and Butte, A. J. (2005). Systematic survey reveals general applicability of ''guilt-by-association'' within gene coexpression networks. *BMC Bioinform.* **6**, 227.
Wong, S. L. *et al.* (2004). Combining biological networks to predict genetic interactions. *Proc. Natl. Acad. Sci. USA* **101**, 15682–15687.
Wu, J., Kasif, S. and DeLisi, C. (2003). Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* **19**, 1524–1530.
Wu, L. F. *et al.* (2002). Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* **31**, 255–265.
Xenarios, I. *et al.* (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305.
Xia, Y. *et al.* (2004). Analyzing cellular biochemistry in terms of molecular networks. *Ann. Rev. Biochem.* **73**, 1051–1087.
Yamanishi, Y., Vert, J. P. and Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics* **20**(Suppl 1), I363–I370.
Yanai, I., Derti, A. and DeLisi, C. (2001). Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl. Acad. Sci. USA* **98**, 7940–7945.
Yanai, I., Mellor, J. C. and DeLisi, C. (2002). Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet.* **18**, 176–179.
Zhang, L. V., Wong, S. L., King, O. D. and Roth, F. P. (2004). Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinform.* **5**, 38.
Zhu, J. and Zhang, M. Q. (1999). SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* **15**, 607–611.