Chapter 14

Integrating Functional Genomics Data

Insuk Lee and Edward M. Marcotte

Abstract

The revolution in high throughput biology experiments producing genome-scale data has heightened the challenge of integrating functional genomics data. Data integration is essential for making reliable inferences from functional genomics data, as the datasets are neither error-free nor comprehensive. However, there are two major hurdles in data integration: heterogeneity and correlation of the data to be integrated. These problems can be circumvented by quantitative testing of all data in the same unified scoring scheme, and by using integration methods appropriate for handling correlated data. This chapter describes such a functional genomics data integration method designed to estimate the "functional coupling" between genes, applied to the baker's yeast *Saccharomyces cerevisiae*. The integrated dataset outperforms individual functional genomics datasets in both accuracy and coverage, leading to more reliable and comprehensive predictions of gene function. The approach is easily applied to multicellular organisms, including human.

Key words: Data integration, function prediction, guilt-by-association, gene association, functional coupling, data correlation, data heterogeneity.

1. Introduction

The ultimate goal of functional genomics is to identify the relationships among all genes of an organism and assign their physiological functions. This goal is certainly ambitious, but seems somewhat more attainable when considering the remarkable advances in automation of molecular biology and innovations in high throughput analysis techniques, which are already producing enormous sets of functional data. Such data include micro-array analyses of gene expression *(1, 2)*, protein interaction maps using yeast two-hybrid *(3–8)*, affinity purification of protein complexes *(9–11)*, synthetic

Jonathan M. Keith (ed.), *Bioinformatics, Volume II: Structure, Function and Applications, vol. 453*
© 2008 Humana Press, a part of Springer Science+Business Media, Totowa, NJ
Book doi: 10.1007/978-1-60327-429-6 Springerprotocols.com

lethal screening *(12, 13)*, and many others, including computational methods that predict gene functions using comparative genomics approaches. These methods are introduced in earlier chapters of this book: prediction of gene function by protein sequence homology (*see* **Chapter 6**), Rosetta Stone proteins or gene fusions (*see* **Chapter 7**), gene neighbors (*see* **Chapter 8**), and phylogenetic profiling (*see* **Chapter 9**)—provide related information with relatively low cost. All of these data enable the prediction of gene function through guilt-by-association—the prediction of a gene's function from functions of associated genes. For example, if we observe a gene associated with other genes known to be involved in ribosomal biogenesis, we might infer the gene is involved in ribosomal biogenesis as well.

Although functional genomics data are accumulating rapidly, the assignment of functions to the complete genome or proteome is still far from complete. One complicating factor is the fact that all functional analyses (both experimental and computational) contain errors and systematic bias. For example, yeast two-hybrid methods can detect associations between physically interacting genes only, whereas genetic methods only occasionally do *(14, 15)*. Integration of diverse functional genomics data can potentially overcome both errors and systematic bias. In practice, data integration improves prediction of gene function by guilt-by-association, generally resulting in stronger inferences and larger coverage of the genome *(16–23)*.

Nevertheless, successfully integrating data is not trivial due to two major problems: heterogeneity and correlation among the data to be integrated. This chapter discusses an approach for integrating heterogeneous and correlated functional genomics data for more reliable and comprehensive predictions of gene function, applying the method to genes of a unicellular eukaryotic organism, the yeast *Saccharomyces cerevisiae*.

2. Methods

2.1. Standardization of Heterogeneous Functional Genomics Data

A major hurdle in data integration is the heterogeneity of the data to be integrated. All data to be integrated must be assessed for relevance and informativeness to the biological hypothesis that one wants to test. In practice, this means choosing a common quantitative test of relevance to apply to each dataset, allowing comparison of the datasets by a single unified scheme. After this data standardization, integration becomes a much easier task. Note that this process is not strictly necessary—many classifiers do not require it—but it greatly simplifies later interpretation of results.

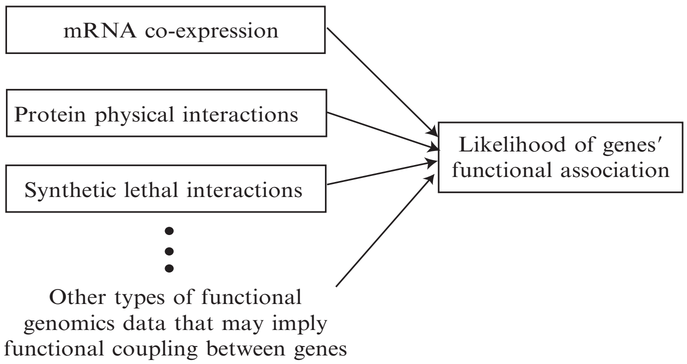
mRNA co-expression
Protein physical interactions
Synthetic lethal interactions
Other types of functional genomics data that may imply functional coupling between genes
Likelihood of genes' functional association

Fig. 14.1. Schematic for applying functional genomics data to estimate functional coupling between genes. Different functional genomics data imply different types of gene associations with varying confidence levels. To predict functional associations, these diverse data are re-interpreted as providing likelihoods of functional associations, and then combined into a single, integrated estimate of the observed coupling between each pair of genes.

The first step in a data integration scheme is choosing the biological hypothesis we want to test. In the prediction of gene functions by guilt-by-association, we are often interested in whether two given genes are functionally associated or not. Although we may be interested in a more specific type of relationship (e.g., protein physical interactions, genetic interactions, pathway associations), we illustrate here a more general notion of "functional" associations, which implicitly includes all of these more specific associations (**Fig. 14.1**). This can be defined more precisely as participation in the same cellular system or pathway.

2.2. Constructing a Reference Set for Data Evaluation

Standardization of heterogeneous data is carried out by evaluating them using a common benchmarking reference. Here, the reference set consists of gene pairs with verified functional associations under some annotation scheme (prior knowledge). The positive reference gene associations are generated by pairing genes that share at least one common annotation and the negatives by pairing annotated genes that do not share any common annotation. The quality of the reference set—both its accuracy and extensiveness—is critical to successful data evaluation. It is also important to keep in mind that the reference set must be consistent throughout the entire data evaluation and integration.

Several different annotations can be used to generate the reference set of functional associations. For example, reference sets might consist of gene pairs sharing functional annotation(s), sharing pathway annotation(s), or found in the same complex(es). The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations, Gene Ontology (GO), and Munich Information Center for Protein Sequences (MIPS) complex annotation (which is available only for *Saccharomyces cerevisiae*) are useful annotation sets to generate reference sets (*see* **Note 1**). A reference set generated from genes sharing or not sharing KEGG pathway annotation is used for the discussion in this chapter. We define two genes to be functionally associated if we observe at least one KEGG pathway term (or pathway code) annotating both genes.

2.3. A Unified Scoring System for Data Integration

One way to compare heterogeneous functional genomics data is to measure the likelihood that the pairs of genes are functionally associated conditioned on the data, calculated as a log-likelihood score:

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

where *P(I|D)* and *P(~I|D)* are the frequencies of functional associations observed in the given dataset *(D)* between the positive *(I)* and negative *(~I)* reference set gene pairs, respectively. *P(I)* and *P(~I)* represent the prior expectations (the total frequencies of all positive and negative reference set gene pairs, respectively). A score of zero indicates interaction partners in the data being tested are no more likely than random to be functionally associated; higher scores indicate a more informative dataset for identifying functional relationships.

2.4. Mapping Data-Intrinsic Scores into Log-Likelihood Scores

Many data come with intrinsic scoring schemes, which can easily be converted to log-likelihood scores. **Figure 14.2** describes such a mapping for using 87 DNA micro-array datasets measuring mRNA expression profiles for different cell cycle time points *(1)*. Genes co-expressed under similar temporal and spatial conditions are often functionally associated. The degree of co-expression can be measured as the Pearson correlation coefficient (*PCC*) of the two genes' expression profiles. Gene pairs are sorted by the calculated *PCC*, and then binned into equal-sized sets of gene pairs, starting first with the higher Pearson correlation coefficients (*see* **Note 2**). For each set of gene pairs in a given bin, *P(I|D)* and *P(~I|D)* are calculated. These probabilities correspond to a given degree of co-expression within this dataset. The log-likelihood score is calculated from these probabilities, along with *P(I)* and *P(~I)*, the unconditional probabilities calculated from the reference gene pairs. If the PCC provides a significant correlation with the log-likelihood score, we then define a regression model with which we can map all data-intrinsic scores (PCC scores) into the standardized scores (log-likelihood scores) (*see* **Note 3**). In this way, the learned relationship between co-expression and functional coupling can be extended to all un-annotated pairs. The datasets re-scored using log-likelihood scores will be used for integrating the datasets.

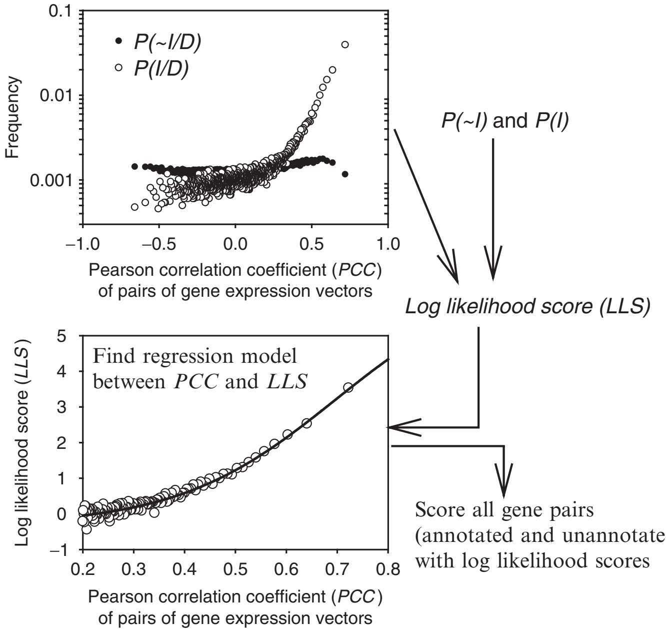

Fig. 14.2. An example of estimating functional coupling from DNA micro-array–based mRNA expression data. Many functionally associated genes tend to be co-expressed through the course of an experiment. Thus, the Pearson correlation coefficient of two genes' expression vectors shows a positive correlation with their tendency to share pathway annotation. Here, yeast mRNA expressions patterns across the cell cycle (1) are compared to their participation in the same KEGG (24) pathways, plotted for all annotated gene pairs as a function of each pair's Pearson correlation coefficient. The frequencies of gene pairs sharing pathway annotation *(P(I/D))* are calculated for bins of 20,000 gene pairs. In contrast, the frequencies of gene pairs not sharing pathway annotation *(P(~I/D))* show no significant correlation with the correlation in expression. The ratio of these two frequencies, corrected by the unconditional frequencies (*P(I)* and *P(~I)*), provides the likelihood score of belonging to the same pathway for the given condition. In practice, we calculate the natural logarithm for the likelihood score to create an additive score (log-likelihood score). Using a regression model for the relationship, we can score all gene pairs (not just the annotated pairs) with log-likelihood scores, indicating the normalized likelihood of functional coupling between genes. (Adapted from Lee, I., Date, S. V., et al. (2004) A probabilistic functional network of yeast genes. *Science* 306, 1555–1558.)

2.5. Integration of Correlated Data Using a Simple Weighted Sum Approach

If datasets to be integrated are completely independent, integration is simple: We can use a naïve Bayes approach, simply adding all available log-likelihood scores for a given gene pair to give the pair's integrated score. However, this assumption of independence among datasets is often unrealistic. We often observe strong correlations among functional genomics data. Integrating correlated datasets using formal models, such as Bayesian networks *(25)*, requires defining the degree of correlation among the datasets. The complexity of this correlation model increases exponentially with the number of datasets. Therefore, defining a correlation model among datasets is often computationally challenging or impractical. If the dataset correlation is relatively weak, the naïve Bayes approach provides reasonable performance in integration. As we accumulate more and more functional genomics data to be integrated, however, this convenient and efficient assumption becomes more troublesome.

An alternative approach that is simple but still accounts for data correlation is a variant of naïve Bayes with one additional parameter accounting for the relative degree of correlation among datasets. In this weighted sum method, we first collect all available log-likelihood scores derived from the various datasets and lines of evidence, then add the scores with a rank-order determined weighting scheme. The weighted sum (*WS*) score for the functional linkage between a pair of genes is calculated as:

$$WS = \sum_{i=1}^{n} \frac{L_i}{D^{(i-1)}},$$

where L represents the log-likelihood score for the gene association from a single dataset, D is a free parameter roughly representing the relative degree of correlation between the various datasets, and i is the rank index in order of descending magnitude of the n log-likelihood scores for the given gene pair. The free parameter D ranges from 1 to + ∞, and is chosen to optimize overall performance (accuracy and coverage, *see* **Note 4**) on the benchmark. When $D = 1$, *WS* represents the simple sum of all log-likelihood scores and is equivalent to a *naïve* Bayesian integration. We might expect D to exhibit an optimal value of 1 in the case that all datasets are completely independent. As the optimal value of D increases, *WS* approaches the single maximum value of the set of log-likelihood scores, indicating that the various datasets are entirely redundant (i.e., no new evidence is offered by additional datasets over what is provided by the first set). **Figure 14.3** illustrates the performance using different values of D in integrating datasets with different degrees of correlation. Datasets from similar types of functional genomics studies are often highly correlated. The integration of highly correlated DNA micro-array datasets from different studies *(1)* is illustrated in **Fig. 14.3A**. Here, the assumption of data independence ($D = 1$) provides high scores for a limited portion of the proteome. However, accounting for partial data correlation (e.g., $D = 2$) provides significantly increased coverage of the proteome in identifying functionally associated gene pairs for a reasonable cost of likelihood. In fact, the assumption of complete correlation ($D = +\infty$) among the different gene expression datasets provides a very reasonable trade-off between accuracy and coverage in identifying functionally associated gene pairs. In contrast, **Fig. 14.3B** shows the integration of 11 diverse functional genomics datasets, described in full in Lee et al. *(21)*. This integration is optimal with

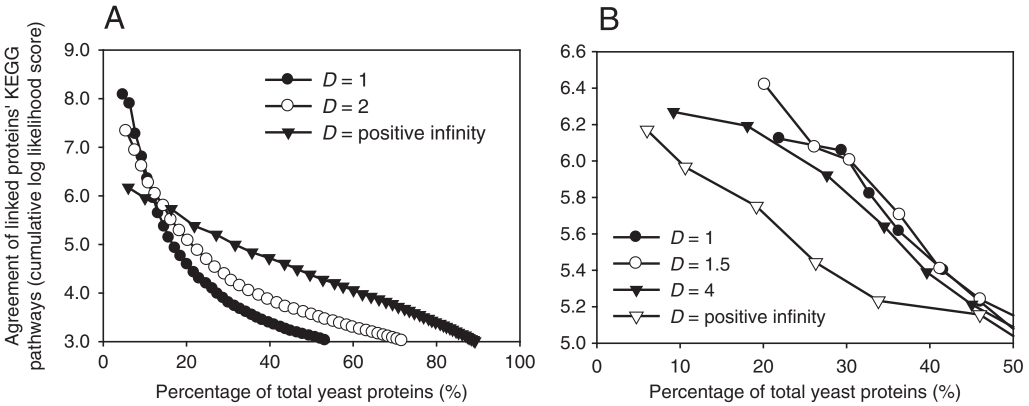

Fig. 14.3. Effects of data correlation on data integration. Here (**A**) 12 different DNA micro-array datasets or (**B**) 11 diverse functional genomics datasets from Lee et al. (21) are integrated by the weighted sum method, applying different values of *D*. The quality of the integration is assessed by measuring coverage (percentage of total protein-coding genes covered by gene pairs in the data) and accuracy (cumulative log-likelihood scores measured using a reference set of KEGG pathway annotations), with each point indicating 2,000 gene pairs. Integration of DNA micro-array datasets with naïve Bayes approach (*D* = 1) shows high likelihood scores for the top-scoring gene functional pairs, but a rapid decrease in score with increasing proteome coverage. In contrast, integration assuming higher (*D* = 2) or complete correlation (*D* = positive infinity) provides a dramatic improvement in coverage for a reasonable cost of likelihood. For the integration of 11 diverse functional genomics datasets, the naïve Bayes approach (*D* = 1) shows reasonable performance. However, the best performance is observed for *D* = 1.5. These examples illustrate that better performance in data integration is achieved by accounting for the appropriate degree of correlation among the datasets—similar types of datasets are often highly correlated, requiring high values of *D*, whereas more diverse types of data can be relatively independent, requiring low values of *D* or naïve Bayes for the optimal integration.

neither complete independence (*D* = 1) nor complete dependence (*D* = + ∞), integration with *D* = 1.5, accounting for intermediate dependence, achieves optimal performance.

Integrated data generally outperforms the individual datasets. A precision-recall curve (*see* **Note 5**) for the 11 individual datasets and the integrated set demonstrates that data integration improves performance of identifying functional associations in terms of both recall and precision (**Fig. 14.4**).

2.6. Inference of New Gene Function by Guilt-by-Association: A Case Study of PRP43

The gene associations arising from integrating different functional genomics datasets often generate new biological hypotheses. For example, PRP43 was initially implicated only in pre-mRNA splicing *(26)*. Interestingly, many genes involved in ribosomal biogenesis are strongly associated with PRP43 in an integrated gene network *(21)*. Among the 5 genes most strongly associated with PRP43, as ranked by the log-likelihood scores of their associations, three are known to be involved in ribosomal biogenesis (**Table 14.1**). Three recent experimental studies have validated this association, demonstrating that PRP43 is a regulator of both pre-mRNA splicing and ribosomal biogenesis *(27–29)*.

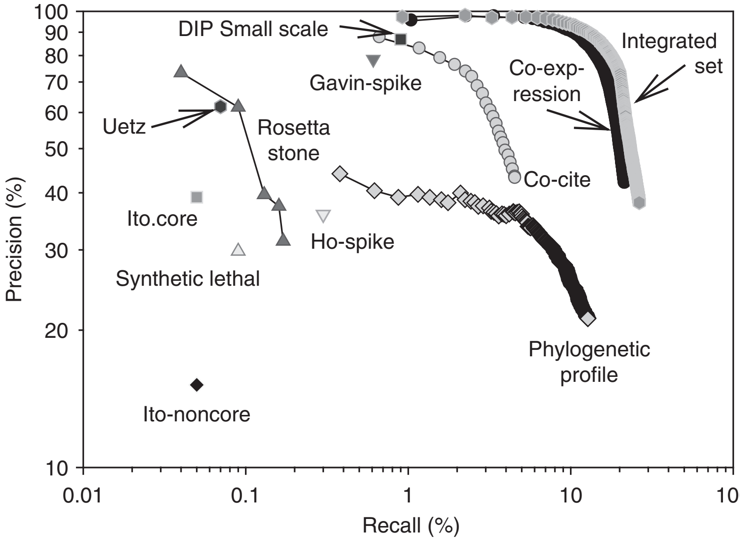

Fig. 14.4. A comparison of the quality of gene functional associations found using 11 diverse functional genomics datasets and an integrated dataset. The predictive power of 11 diverse functional genomics datasets and the integrated dataset *(21)* are assessed by a recall-precision curve (*see* **Note 5**). Measurements are carried out for bins of 2,000 gene pairs. The integrated dataset outperforms all individual datasets, with data integration improving the prediction of functional associations in both accuracy and coverage. Assessment curves are plotted with logarithm of both recall and precision for visualization purpose; thus, the predictive powers are significantly different between co-expression and integrated dataset with linear scale.

Table 14.1
The five genes most strongly associated with Prp43

Rank	Name	Cellular location[a]	Cellular function[b]
1	ERB1	nucleolus	rRNA processing
2	RRB1	nucleolus	ribosome biogenesis
3	LHP1	nucleus	tRNA processing
4	URA7	cytosol	CTP biosynthesis, phospholipid biosynthesis, pyrimidine base biosynthesis
5	SIK1	small nucleolar ribonucleo protein complex	rRNA modification, 35S primary transcript processing, processing of 20S pre-rRNA

[a] Annotated by Gene Ontology cellular component.
[b] Annotated by Gene Ontology biological process.

3. Notes

1. These annotations are hierarchically organized, and choosing different levels of the annotation hierarchy may generate quite different evaluations for the same dataset. Generally speaking, top-level annotations provide extensive coverage but low information specificity (resolution), whereas low-level annotations decrease coverage but increase information specificity. Therefore, the choice of appropriate levels of hierarchical annotation must be considered carefully in order to achieve the optimal trade-off between coverage and specificity. KEGG pathway and GO biological process annotations are available for yeast from ftp://ftp.genome.jp/pub/kegg/pathways/sce/sce_gene_map.tab and http://www.geneontology.org/ontology/process.ontology, respectively. CYGD (the comprehensive yeast genome database) functional categories from MIPS are available at ftp://ftpmips.gsf.de/yeast/catalogues/complexcat/). CYGD lists yeast cellular complexes and their member proteins. Similar data are available for many other organisms. It is striking that the current yeast annotation and reference sets are quite non-overlapping *(30)*. For example, fewer than half of the KEGG pathway database associations are also contained in the Gene Ontology (GO) annotation set. The low overlap is primarily due to different data mining methods and to inclusion bias among the annotation sets. However, this provides the opportunity to generate more comprehensive reference sets by combination of different annotation sets.

2. Here, an appropriate choice of bin size is important. Generally, a minimum of 100 annotated gene pairs per bin is recommended to obtain statistically reliable frequencies. Overly large bin sizes decrease the resolution of evaluation. Binning must start with the more significant values first. For Pearson correlation coefficient scores, positive values tend to be more meaningful than negative values. (We observe significant signals with Pearson correlation coefficient > 0.3 in **Fig. 14.2**.) Thus, we rank gene pairs with decreasing Pearson correlation coefficients (starting from +1) and bin in increasing increments of x pairs ($x = 20,000$ in **Fig. 14.2**).

3. Regression models may suffer from noisy data. For microarray data, gene pairs with negative Pearson correlation coefficient scores are often un-correlated with log-likelihood scores of gene functional associations. As most of the ~18 million of yeast gene pairs belong to this group of noisy data, taking only gene pairs with positive Pearson correlation coefficients generally gives an improved regression model.

4. The ability to identify functional associations is assessed by measuring accuracy for a given cost of coverage. To control for systematic bias (the dataset may predict well for only certain gene groups, e.g., associations between ribosomal proteins), we measure the coverage of total genes as the percentage of all protein-coding genes represented in the dataset. Accuracy is defined as the cumulative log-likelihood score of the dataset. The area under this coverage-accuracy curve line indicates relative performance. We select the value of *D* that maximizes the area under the coverage-accuracy curve.
5. One formal way to evaluate data coverage and accuracy is by plotting a recall-precision curve. *Recall* (defined as the percentage of positive gene associations in the reference set correctly predicted as positive gene associations in the dataset) provides a measure of coverage and *precision* (defined as the percentage of predicted positive gene associations in the dataset confirmed as true positive gene associations by the reference set) provides a measure of accuracy. The evaluation method should be able to identify any possible over-fitting during data integration, which occurs when the training process simply learns the training set, rather than a more generalized pattern. Over-fitting is tested using a dataset that is completely independent from the training set. The simple way of making an independent test set is to leave out some fraction (e.g., ~30%) of the original training set as a separate test set. However, this reduces the size of the training set, thus decreasing training efficiency. An alternative method is randomly splitting the original training set into *k* subsets, then iterating training *k* times, using one subset for the testing and all others for the training. The average measurements from this iterative process, called *k*-fold cross-validation, provide a fairly unbiased evaluation with minimal loss of training power (*see* **Chapter 15, Section 3.4.3** and **Note 3** for more details).

Acknowledgments

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

1. Gollub, J., Ball, C. A., Binkley, G., et al. (2003) The Stanford Micro-array Database: data access and quality assessment tools. *Nucleic Acids Res* 31, 94–96.
2. Barrett, T., Suzek, T. O., Troup, D. B., et al. (2005) NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res* 33, D562–566.
3. Uetz, P., Giot, L., Cagney, G., et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627.
4. Ito, T., Chiba, T., Ozawa, R., et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 98, 4569–4574.
5. Giot, L., Bader, J. S., Brouwer, C., et al. (2003) A protein interaction map of *Drosophila melanogaster. Science* 302, 1727–1736.
6. Li, S., Armstrong, C. M., Bertin, N., et al. (2004) A map of the interactome network of the metazoan *C. elegans. Science* 303, 540–543.
7. Rual, J. F., Venkatesan, K., Hao, T., et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178.
8. Stelzl, U., Worm, U., Lalowski, M., et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968.
9. Gavin, A. C., Bosche, M., Krause, R., et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147.
10. Ho, Y., Gruhler, A., Heilbut, A., et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415, 180–183.
11. Bouwmeester, T., Bauch, A., Ruffner, H., et al. (2004) A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat Cell Biol* 6, 97–105.
12. Tong, A. H., Evangelista, M., Parsons, A. B., et al. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294, 2364–2368.
13. Tong, A. H., Lesage, G., Bader, G. D., et al. (2004) Global mapping of the yeast genetic interaction network. *Science* 303, 808–813.
14. Wong, S. L., Zhang, L. V., Tong, A. H., et al. (2004) Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci USA* 101, 15682–15687.
15. Kelley, R., Ideker, T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 23, 561–566.
16. Mellor, J. C., Yanai, I., Clodfelter, K. H., et al. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res* 30, 306–309.
17. Troyanskaya, O. G., Dolinski, K., Owen, A. B., 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.
18. Jansen, R., Yu, H., Greenbaum, D., et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302, 449–453.
19. von Mering, C., Huynen, M., Jaeggi, D., et al. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 31, 258–261.
20. Bowers, P. M., Pellegrini, M., Thompson, M. J., et al. (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* 5, R35.
21. Lee, I., Date, S. V., Adai, A. T., et al. (2004) A probabilistic functional network of yeast genes. *Science* 306, 1555–1558.
22. Gunsalus, K. C., Ge, H., Schetter, A. J., et al. (2005) Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. *Nature* 436, 861–865.
23. Myers, C. L., Robson, D., Wible, A., et al. (2005) Discovery of biological networks from diverse functional genomic data. *Genome Biol* 6, R114.
24. Kanehisa, M., Goto, S., Kawashima, S., et al. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res* 30, 42–46.
25. Jensen, F. V. (2001) *Bayesian Networks and Decision Graphs.* Springer, New York.
26. Martin, A., Schneider, S., 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.
27. Lebaron, S., Froment, C., Fromont-Racine, M., et al. (2005) The splicing ATPase prp43p is a component of multiple preribosomal particles. *Mol Cell Biol* 25, 9269–9282.
28. Leeds, N. B., Small, E. C., Hiley, S. L., et al. (2006) The splicing factor Prp43p, a DEAH box ATPase, functions in ribosome biogenesis. *Mol Cell Biol* 26, 513–522.

29. Combs, D. J., Nagel, R. J., Ares, M., Jr., et al. (2006) Prp43p is a DEAH-box spliceosome disassembly factor essential for ribosome biogenesis. *Mol Cell Biol* 26, 523–534.

30. Bork, P., Jensen, L. J., von Mering, C., et al. (2004) Protein interaction networks from yeast to human. *Curr Opin Struct Biol* 14, 292–299.