Practical computational approaches to inferring protein function

Edward M. Marcotte

A quick search through high-throughput proteomics and genomics data can reveal information on many aspects of protein function, such as mutant phenotypes, protein interactions, mRNA expression patterns, transcriptional regulation, and even protein structure. The computational integration of such data is proving to be the most effective route to protein function.

Department of Chemistry and Biochemistry, Institute for Cellular and Molecular Biology, Center for Computational Biology and Bioinformatics,
2500 Speedway, MBB 3.232,
University of Texas at Austin,
Austin, TX 78712, USA
e-mail: marcotte@icmb.utexas.edu

As the availability of raw data about protein function grows continuously, investigators are scrambling to convert these data to knowledge. Datasets describing deletion mutant phenotypes, protein and mRNA expression profiles, genome sequences, and protein interactions, to name but a few, have opened several new routes to protein function. Each set of data, typically collected on as large a scale as is practical, tells something of the functions of many proteins. As a consequence, few of the current computational approaches for inferring protein function derive from first-principles models of protein function; rather, they represent varied approaches for 'mining' functional inferences from diverse proteomics and functional-genomics data. The best routes to function involve integrating the partial functional inferences from many types of these data at once.

Approaches for the integration of different types of data relating to protein function generally take what might be termed the 'genome-down' rather than 'protein-up' approach. They exploit the principle that a protein's function can be determined more easily in the context of the other proteins with which it works in the cell. Genome-down approaches systematically analyze the entire set of proteins encoded by a genome, and only focus in on specific proteins after completing this holistic analysis. One of the most compelling arguments in support of this strategy is that the proteins with known function act as cases to test the approach's effectiveness, and the overall accuracy of the approach can be measured. Such assessments of accuracy show that these methods are often still a bit hit-and-miss, and there is no guarantee that data will exist for a particular protein. Despite these caveats, a tremendous amount of functional information has been found in this manner, much freely available for public consumption. This review discusses recent computational approaches for inferring protein function, several successful integrated approaches for analyzing both functional genomics and proteomics data, and tools for effectively navigating these complex datasets. A flowchart describing the use of these tools is provided in Fig. 1; internet links to the major resources available at present are listed in Boxes 1–3.

Protein function from comparative genomics

Two major trends are emerging in the use of genomics data to infer protein function. The first approach relies upon discovering the information about gene function that is intrinsic in genomes. This information can be revealed by finding contextual cues shared by genes that interact or perform a given function [1,2]. The second approach, which relies upon the completeness of genome sequences, is to match a gene with its equivalent genes in well-characterized model organisms. This approach allows the investigator to profit from the rich functional datasets that exist for model organisms.

When using the first approach, several contextual trends have proved to be useful for finding protein function. These include searching for evidence of fusions between the gene of interest and other functionally related genes [3–5]; finding functionally linked genes because of their tendency to be 'co-inherited' [6–9]; identifying proteins that physically interact by looking for the conservation of their phylogenetic tree structures [10–12]; and

Box 1. **Servers for computationally predicting protein function**

Prediction of protein function, interactions, and networks

Bioverse	http://bioverse.compbio.washington.edu
In silico two hybrid	http://www.pdg.cnb.uam.es/i2h
InterDom	http://InterDom.lit.org.sg
Magic	http://genome-www.stanford.edu/magic
Predictome	http://predictome.bu.edu
ProtFun	http://www.cbs.dtu.dk/services/ProtFun
ProteinFunction	http://www.aber.ac.uk/compsci/Research/bio/ProteinFunction
Protein Link Explorer (PLEX)	http://bioinformatics.icmb.utexas.edu/plex
STRING	http://www.bork.embl-heidelberg.de/STRING

Predicting prokaryotic operons to find functionally linked proteins

Gene Neighbors	http://bioinformatics.icmb.utexas.edu/operons
STRING	http://www.bork.embl-heidelberg.de/STRING
TUpredictions	http://www.cifn.unam.mx/moreno/pub/TUpredictions
WIT	http://wit.mcs.anl.gov/WIT2

computationally identifying operons, either by virtue of their conservation across organisms [13–15] or the physical separation between the genes along the chromosome [16]. Each of these approaches produces a set of candidate proteins that are functionally linked to a protein of interest, with a score that indicates the confidence of the linkages.

Comparisons of phylogenetic trees or analyses of the numbers of nucleotides that separate genes can be performed easily. The practical implementation of the remaining methods, however, requires the systematic comparison of large sets of protein sequences from many organisms, followed by statistical analysis of the many comparisons. For example, calculation of phylogenetic-profile-based linkages involves comparing the amino-acid sequence of each protein encoded by a genome with the complete protein complement of all organisms with sequenced genomes. From the results of these comparisons, profiles are constructed that indicate the organismal distribution of each protein's homologs. Comparison of the profiles against each other reveals proteins with similar phylogenetic distributions, which are frequently functionally linked. In one such analysis, carried out recently on 57 genomes, approximately 31 billion sequence comparisons were made during the construction of a database of phylogenetic profiles that could be searched for functionally linked genes [8]. The scale of such analyses means that their results must be analyzed statistically to minimize the inevitable false-positive linkages that arise. Nevertheless, over the past year, these approaches have begun the shift in status from specialized research topics to publicly accessible research tools. Several internet servers have been created where these methods can be explored and where functional linkages for a protein of interest can be found, accompanied by estimates of the confidence in the predictions. Several such web servers are listed in Box 1.

Computational genetics approaches have proven useful for several 'real-world' cases. One such case is the computational identification of the archaeal exosome, which was achieved using a combined analysis of gene sequence homology and gene order [17]. Another recent example is the discovery of functional displacements of thiamin biosynthesis genes [18]. In this study, candidates for gene displacements in thiamin biosynthetic pathways were identified using comparative genomics. Pairs of genes that might substitute for each other in these pathways were first identified by their anti-correlated phylogenetic distributions, the involvement of these genes in the biosynthetic pathway were validated experimentally.

The second approach, which is poised for more widespread adoption, is the 'borrowing' of function from orthologs in better-characterized model organisms [19], such as the yeast *Saccharomyces cerevisiae*, the nematode worm *Caenorhabditis elegans*, the fly *Drosophila melanogaster*, and the bacterium *Escherichia coli*. Given a protein of interest, it is worth attempting to identify the equivalent protein in a model organism and then searching for available functional data among the rich functional genomics and proteomics databases available for model organisms. The identification of the equivalent

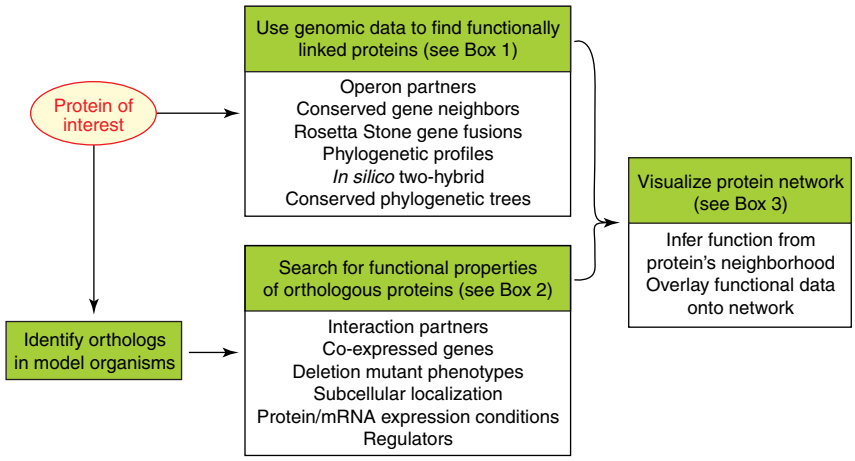

Figure 1. A flowchart describing the general 'genome-down' steps for identifying protein function computationally. Two parallel strategies exist: comparative genomics approaches for identifying linkages to other proteins, and mapping the protein of interest into an organism from which abundant functional genomics data are available and then assigning function. When using either strategy, it is often useful to examine the local network of proteins around the protein of interest, allowing the neighbors' functions to clarify the central protein's role.

protein in a model organism may not be trivial — for example, the highest-scoring BLAST match in the genome may not actually be to a protein of equivalent function — and so one typically would wish to find an orthologous, rather than simply a homologous, protein.

Orthologous genes, defined as homologous genes that are separated by speciation events [20], are typically more functionally equivalent than paralogs, defined as homologous genes separated by a duplication event. Thus, it can often be important to distinguish between these two categories of homologous genes. Even paralogs can give strong hints as to function, providing the basis for the usefulness of protein-domain databases, with the caveat that the precise functions of paralogs occasionally differ and may therefore mislead if relied upon exclusively. Identifying orthologs and paralogs requires the calculation of rooted phylogenetic trees, with outgroups, from which one can distinguish gene duplication events from speciation events. This approach is difficult to automate, and hence hard to scale to complete genomes, so several heuristic approaches have been developed that approximately identify orthologous genes.

One such heuristic approach for finding orthologs in an imperfect, yet rapid and easy, fashion has been developed by Remm and colleagues [21]. The approach is an improvement on the notion of finding 'bi-directional best hits' (BBHs). BBHs are proteins from two genomes, each of which is the top-scoring BLAST match of the other when searched in the appropriate genome [14]. Remm *et al.*'s improvement, termed InParanoid, is to recognize that many genes have been duplicated and that the duplications blur the ability to identify orthologs. InParanoid therefore searches for BBHs, but then also identifies proteins from the two genomes that are as similar to the BBH proteins as the two BBH proteins are to each other. In this manner, two or more potential orthologs for a protein in the other genome may be identified within one organism

With a potential ortholog in hand, one can now search for its associated functional information. Recommended functional databases are listed in Box 2 and encompass model organism mRNA expression profiles, gene deletion phenotypes, protein subcellular localization, transcription-factor specificity, genetic interactions and protein interactions.

Integrating functional genomics and proteomics data

Data derived from DNA microarrays are one of the richest sources of information about protein function. Literally thousands of microarray datasets exist in the public domain, spawning an entire field of research in interpreting the data and distilling out functional information. Much effort in analyzing the data focuses on finding groups of genes (clusters) that have tended to co-express across a variety of experiments (reviewed by Slonim [22]).

This approach has been useful in suggesting functions for several uncharacterized genes (see [23,24] for examples). Without additional data, however, the results often tend to be coarse-grained and uncertain. This uncertainty is primarily caused by the inherent ambiguity in the relationships between the genes' expression patterns, which result in alternate clusterings of more-or-less equivalent quality. Furthermore, many genes may be found in a

Box 2. **Functional genomics data for proteins of model organisms, which are useful for estimating the functions of orthologous proteins from other systems**

Prediction of orthologs	
InParanoid	http://inparanoid.cgb.ki.se
Clusters of Orthologs (COGs)	http://www.ncbi.nlm.nih.gov/COG
mRNA expression profiles	
dbEST	http://www.ncbi.nlm.nih.gov/dbEST
SAGEmap	http://www.ncbi.nlm.nih.gov/SAGE
Stanford Microarray Database	http://genome-www5.stanford.edu/MicroArray/SMD
UniGene	http://www.ncbi.nlm.nih.gov/UniGene
Protein interaction data	
Bind	http://www.bind.ca
BRITE	http://www.genome.ad.jp/brite
Database of Interacting Proteins	http://dip.doe-mbi.ucla.edu
GRID	http://biodata.mshri.on.ca/grid
MIPS	http://mips.gsf.de/proj/yeast/CYGD/db/index.html
PIMRider (*Helicobacter pylori* protein interactions)	http://pim.hybrigenics.com
Regulatory network data	
BIOCYC/METACYC	http://biocyc.org
GeneNet	http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet
KEGG	http://www.genome.ad.jp/kegg
Promoter Database of *S. cerevisiae*	http://cgsigma.cshl.org/jian
RegulonDB	http://www.cifn.unam.mx/Computational_Genomics/regulondb
TRANSFAC	http://transfac.gbf.de/TRANSFAC
Yeast transcription factor targets	http://web.wi.mit.edu/young/regulator_network
Model organism mutant phenotypes	
Comprehensive yeast genome database	http://mips.gsf.ed/genre/proj/yeast/index.jsp **[Could you please check this link, it's not working for me.]**
FlyBase	http://flybase.bio.indiana.edu
Saccharomyces Genome Database	http://www.yeastgenome.org
WormBase	http://www.wormbase.org
TRIPLES (yeast disruption phenotypes)	http://ygac.med.yale.edu/triples/triples.htm
Protein subcellular localization and mRNA *in situ* hybridization data	
TRIPLES (yeast protein localization)	http://ygac.med.yale.edu/triples/triples.htm
Yeast green fluorescent protein (GFP) localization database	http://yeastgfp.ucsf.edu
C. elegans mRNAs	http://nematode.laboratory.nig.ac.jp **[Could you please check this link, it's not working for me.]**
D. melanogaster mRNAs	http://www.fruitfly.org/cgi-bin/ex/insitu.pl
Xenopus laevis mRNAs	http://www.dkfz-heidelberg.de/molecular_embryology/axeldb.htm

single cluster, complicating the precise definition of their relationships. For these reasons, recent efforts to interpret these data have sought to increase the accuracy of the clusters. For example, Wu and colleagues [25] increased accuracy by testing alternate clustering methods and keeping track of those genes that consistently associated together. They then assigned functions to genes according to the well-characterized genes that they consistently clustered with. In this manner, Wu *et al*. predicted the involvement of five genes in rRNA processing and verified these predicted functions experimentally.

By integrating microarray data with other functional genomics data, the quality of the discovered relationships **[between clustered proteins?]** has recently been improved considerably.

Statistical approaches for assigning protein function from disparate sorts of data have been explored in the past (e.g. [26–29]), but Troyanskaya and colleagues [30] took this analysis a step further by having experts in the field (mostly curators of the *Saccharomyces* Genome Database) estimate the accuracy of the different classes of functional genomics data. Rather than directly comparing different clustering results (as in [25]), Troyanskaya and colleagues [30] then integrated the functional genomics data according to the expert-assigned accuracies using a probabilistic approach. On the

Box 3. **Software for visualizing complex protein networks**

General tools for network visualization	
Graphlet	http://www.infosun.fmi.uni-passau.de/Graphlet/
Graphviz	http://www.research.att.com/sw/tools/graphviz/
Pajek	http://vlado.fmf.uni-lj.si/pub/networks/pajek/
Tools customized for biological networks	
BioLayout	http://maine.ebi.ac.uk:8000/services/biolayout/
Cytoscape	http://www.cytoscape.org/
InterViewer3	http://wilab.inha.ac.kr/protein/
Large Graph Layout (LGL)	http://bioinformatics.icmb.utexas.edu/lgl
Osprey	http://biodata.mshri.on.ca/osprey

basis of these weighted data, genes were assigned to the most appropriate functional categories from the Gene Ontology project, with results available on the internet (Magic; listed in Box 1).

A second recent improvement in clustering builds on the reasonable assumption that groups of co-expressed genes are often controlled by the same set of regulatory systems. Segal and colleagues [31] exploited this assumption and took a non-obvious approach to finding gene systems: they simultaneously searched both for sets of co-expressed genes and their corresponding regulatory networks. Doubling the search problem, which at first seems to add difficulty, may actually simplify the problem by requiring mutually consistent networks and gene clusters. The algorithm works as follows: initially, the genes are simply clustered according to their expression profiles. Then, a repetitive procedure begins in which known regulatory genes are assembled into simple networks of activators and repressors that can best explain the gene expression patterns within each gene cluster. After the best set of networks has been found, the original genes are redistributed among the clusters, assigning each gene to a cluster according to how well the cluster's associated regulatory network predicts the gene's expression. Then, these two processes are alternated: first constructing the optimal regulatory network for each cluster then reassigning genes among the clusters according to the networks. The program eventually converges upon sets of co-expressed genes and their candidate regulatory networks. Unlike simply clustering the genes, this approach produces interesting, and more important, testable hypotheses that potentially explain why the genes cluster as they do.

Visualizing and navigating complex proteomics data

The most intimidating aspect of working with proteomics and genomics data is often the inherent large scale and complexity of the task. Nowhere is this more apparent than in attempts to unravel networks of proteins or genes, whose tangled sets of connections are complex in the extreme. To take full advantage of the data requires that a protein be viewed in context, and that at least the most relevant interaction partners be organized into a single coherent view.

A useful technique for viewing networks, which was originally derived from an algorithm in computer sciences [32], has been to model proteins as an abstract objects at some location in two- or three-dimensional space. These objects are connected by springs whenever proteins are known to be linked together, and positioned so as to minimize the spring energies. It is important to realize that only the network connections represent real observations, and that the layout represents an attempt to summarize all of these perhaps conflicting associations. Therefore, the resulting distances between proteins in the network can be variable, and may even vary stochastically if the network layout is repeated. However, the tendency is for proteins that function together to be positioned close in space, whereas those that are not intimately linked tend to be further apart. In this manner, two proteins that are linked to the same set of other proteins will often be positioned adjacently, even if the two proteins are not themselves directly linked. This approach allows layouts of very complex networks, but is also famous for producing incomprehensibly complicated 'spaghetti-like' diagrams.

The difficulty of visualizing these complex datasets has led several groups to develop computer programs that allow us to visualize, and even interactively navigate, these networks more effectively. Several new network-visualization tools are listed in Box 3. Most of these new tools retain the spring-based layout approach, but use modifications to improve the visual esthetic and interpretability. Such

modifications include indicating different types of proteins or linkages with different symbols, increasing the separation between major components of the network, allowing interactive navigation or manipulation of the visual field of view, and even collapsing proteins with more-or-less equivalent interactions into single objects in the network [33] to simplify the resulting network.

Some of the newest visualization tools, such as Cytoscape and Large Graph Layout (LGL), also allow the overlay of other forms of functional genomics data onto the network. In this manner, protein and mRNA expression levels can be simultaneously viewed together with the relationships between the proteins, allowing an investigator a visual summary of the behavior of the system. Such a visualization can allow a much more logical analysis of both the expression data and the interaction data by allowing the investigator to see the changes in gene expression in the light of the regulatory and physical interactions between the genes. A researcher can also use such tools search for connected regions of the interaction network that show coordinated changes in gene expression patterns [34]. These tools tend to be most useful as part of a genome-down approach for the simple reason that a protein's final position in the network map relies on maximally satisfying all of the relationships in which it participates. In this manner, the dominant trends in the protein's relationships tend to be reinforced and suppress the less-confident or less-well-observed relationships, in effect providing some filtering to an otherwise very complex set of relationships.

Conclusions and future outlook: is the time ripe for a central repository of protein function?

The approaches discussed here provide general frameworks for discovering protein function by computationally integrating many distinct types of data. Many types of data exist that have yet to be extensively incorporated into these approaches. Such datasets include protein structures and metabolite and protein expression data. Protein structures provide rich information about molecular aspects of protein function, and it should be reasonably straightforward to begin to incorporate these functional inferences with those derived from functional genomics and protein interaction data. Little metabolite expression data exist in the public domain, as useful as they would be, for example, in more precisely characterizing knockout phenotypes. Similarly, protein expression data are accumulating rapidly in public and private laboratories, yet few of this data are publicly available, curtailing the development of algorithms for data analysis.

We expect that protein expression data will be invaluable for many of the same reasons that DNA microarray data are useful: they provide systematic measurements of the major changes in the cell and allow direct characterization of a large fraction of expressed proteins. The field of functional genomics has benefited tremendously from publicly accessible genome sequence data and from centralized DNA microarray databases (e.g. the Stanford Microarray Database), and it is unfortunate that no equivalent exists for proteomics. No doubt the field of proteomics would profit greatly from an extensive public database of protein expression data contributed to by the community of proteomics scientists.

Perhaps an equally pressing need is that for a central repository of protein function data, storing both experimentally determined and computationally predicted functions. The biological community has long had the luxury of community databases that archive primary sequence, structure and mRNA expression data. By contrast, the distillation of functional information from these data is scattered through a myriad of separate publications and web servers. Several more specialized databases, notably the model organism databases listed in Box 2 and open format sequence databases such as SwissProt, have made admirable strides towards cataloging this functional data, but only a small fraction of computational functional analysis has been included. The centralization of this information, with uniformity of formats and access, would open up the work of computational biologists and functional genomicists to the community as a whole. Most importantly, this would allow the full weight of evidence for each function to be examined at once. It would seem the time is ripe for systematically acquired protein functions to be archived systematically.

Acknowledgements

This work was supported by a Packard Fellowship and grants from the National Science Foundation (0219061,0241180), the Welch Foundation (F-1414), and the Texas Advanced Research Program.

References

1 Huynen, M. *et al.* (2000) Exploitation of gene context. *Curr. Opin. Struct. Biol.* 10, 366–370

2 Marcotte, E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.* 10, 359–365

3 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753
4 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90
5 Yanai, I. *et al.* (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl. Acad. Sci. U. S. A.* 98, 7940–7945
6 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288
7 Huynen, M. *et al.* (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* 10, 1204–1210
8 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
9 Wu, J. *et al.* (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* 19, 1524–1530
10 Goh, C-S. *et al.* (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* 299, 283–293
11 Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as an indicator of protein–protein interaction. *Protein Eng.* 14, 609–614
12 Ramani, A.K. and Marcotte, E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* 327, 273–284
13 Dandekar, T. *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328
14 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901
15 Wolf, Y.I. *et al.* (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* 11, 356–372
16 Moreno-Hagelsieb, G. and Collado-Vides, J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* 18(Suppl. 1), S329–S336
17 Koonin, E.V. *et al.* (2001) Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. *Genome Res.* 11, 240–252
18 Morett, E. *et al.* (2003) Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat. Biotechnol.* 21, 790–795
19 Eisen, J.A. and Wu, M. (2002) Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor. Popul. Biol.* 61, 481–487
20 Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113
21 Remm, M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052
22 Slonim, D.K. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.* 32(Suppl.), 502–508
23 Smith, J.J. *et al.* (2002) Transcriptome profiling to identify genes involved in peroxisome assembly and function. *J. Cell Biol.* 158, 259–271
24 Cheung, K.J. *et al.* (2003) A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. *Genome Res.* 13, 206–215
25 Wu, L.F. *et al.* (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31, 255–265
26 Clare, A. and King, R.D. (2002) Machine learning of functional class from phenotype data. *Bioinformatics* 18, 160–166
27 Jansen, R. *et al.* (2002) Integration of genomic datasets to predict protein complexes in yeast. *J. Struct. Funct. Genomics* 2, 71–81
28 Jensen, L.J. *et al.* (2002) Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.* 319, 1257–1265
29 Brown, M.P.S. *et al.* (2000) Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc. Natl. Acad. Sci. U. S. A.* 97, 262–267
30 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. U. S. A.* 100, 8348–8353
31 Segal, E. *et al.* (2003) Module networks: identifying regulatory modules and their condition specific regulators from gene expression data. *Nat. Genet.* 34, 166–176
32 Eades, P. (1984) A Heuristic for graph drawing. *Congressus Numerantium* 42, 149–160
33 Ju, B.H. and Han, K. (2003) Complexity management in visualizing protein interaction networks. *Bioinformatics* 19(Suppl. 1), I177–I179
34 Ideker, T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18(Suppl. 1), S233–S240