Exploiting big biology: Integrating large-scale biological data for function inference

Edward M. Marcotte and Shailesh V. Date

Date received (in revised form): 20th September 2001

Edward Marcotte
is Assistant Professor of Chemistry and Biochemistry in the Institute for Cellular and Molecular Biology at the University of Texas at Austin. His research interests include large-scale methods, both experimental and computational, to determine gene function and the study of the organisation of genes and proteins into complex biological networks.

Shailesh Date
is a graduate student in the Institute for Cellular and Molecular Biology at the University of Texas at Austin. His interests include the study of protein interaction networks in normal and pathological states in cells.

Keywords: *bioinformatics, DNA microarrays, gene expression, functional inference*

Abstract

The amount of data produced by molecular biologists is growing at an exponential rate. Some of the fastest growing sets of data are measurements of gene expression, comparable in quantity only to gene sequences and the vast biological literature. Both gene expression data and sequence data offer hints as to the functions of thousands of newly discovered genes, but neither give complete answers. Therefore, much effort is being focused on integrating these large data sets and combining them with all available functional data to draw inferences about the functions of uncharacterised genes. This review discusses the most pertinent functional data for genome-wide functional inference and describes several methods by which these disparate data types are being integrated.

INTRODUCTION

The many ongoing genome sequencing projects have led to an explosion of sequence data, leading to the discovery of thousands of genes of as yet unknown function. In an effort to assign function to these uncharacterised genes, biologists are developing a number of methods to collect functional data on a genome-wide scale. While these new technologies offer tantalising suggestions of gene function, it is becoming evident that, instead of relying on the results of any one method, large-scale functional inference would benefit immensely if the variety of data being generated were to be unified in some manner. Here we discuss some of the large data sets available to biologists, and some of the ways in which these data are being integrated.

A BRIEF SURVEY OF THE LARGEST BIOLOGICAL DATA SETS

To introduce efforts at integrating data for inferring gene function, let us first examine the available sources of genome-wide data. The biological literature itself represents one of the largest sets of biological data, with about 12 million catalogued research papers and abstracts available through Medline. Representing the current state of biological knowledge, these abstracts and a smaller number of complete research papers available on-line, growing through the efforts of PubMed Central and the Public Library of Science, are proving to be useful for functional inference and automated interpretation of other data.

In spite of the extraordinary amount of literature data, the number of known nucleotide and amino acid sequences may even exceed the literature in quantity. With more than 60 fully sequenced genomes in the public domain, and many more in the pipeline, the amount of sequence data contained in databases such as Genbank (see Table 1) has been growing exponentially. As of April 2001, Genbank contained more than 16 billion base pairs of DNA sequence, which includes expressed sequence tags (ESTs), sequence tagged sites (STS), genome survey sequences (GSS) and complete genome sequence data.

E. M. Marcotte,
Department of Chemistry and Biochemistry and Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin,
TX 78712, USA

Tel: +1 512 471 5435
Fax: +1 512 232 3432
E-mail: marcotte@icmb.utexas.edu

Table 1: Online sources of large-scale biological data discussed

Database	Records	Address
dbEST	8,281,203 public entries	http://www.ncbi.nlm.nih.gov/dbEST/
DIP	9,746 Interactions	http://dip.doe-mbi.ucla.edu/
EcoCyc/MetaCyc	164 pathways	http://ecocyc.pangeasystems.com/ecocyc/
Genbank	11,546,000 sequence records	http://www.ncbi.nlm.nih.gov/Genbank/index.html
KEGG	Most known pathways, in 100 graphical diagrams and 60 orthologue group tables	http://www.genome.ad.jp/kegg/
Medline	>11 million references	http://www4.ncbi.nlm.nih.gov/PubMed/
SAGEmap	189,004 sage sequences; 375,011 unique tags	http://www.ncbi.nlm.nih.gov/SAGE/
SGD	6,000 yeast genes	http://genome-www.stanford.edu/Saccharomyces/
Stanford Microarray Database	313 public array experiments	http://www.dnachip.org/
SWISS-PROT (release 38.0)	~80,000 curated sequence entries from 7,478 organisms	http://www.expasy.ch/
SWISS-2D PAGE	31 2D-PAGE gels; >700 indexed proteins	http://ca.expasy.org/ch2d
TRANSFAC (release 5.0)	9,009 sites; 3,504 factors	http://www.gene-regulation.de/
YPD	6,281 proteins; 20,285 refs	http://www.proteome.com/databases/index.html

Hundreds of sets of DNA microarray data are freely available

Publicly available gene expression data, such as EST libraries, Serial Analysis of Gene Expression (SAGE) libraries and DNA microarray data, are also growing in quantity extremely rapidly. Currently, more than 8 million EST measurements and 2 million SAGE measurements are available from the dbEST and SAGEmap databases respectively, as well as hundreds of publicly available DNA microarray data sets for several different organisms, including yeast and human, from the Stanford Microarray Database.

About ten thousand protein interactions are available in the Database of Interacting Proteins

Other growing sources of genome-wide data are protein-interaction measurements, phenotypic data and protein expression measurements. Though these data sets are not available in the same large volumes as literature, sequence and gene-expression data, they are proving to be extremely useful for construction of large-scale interaction networks, ultimately leading to more accurate functional annotation. Protein interaction data are largely derived from high-throughput yeast two-hybrid experiments in which interactions are measured between all gene pairs in a genome. Over 4,000 unique protein interactions were observed between yeast proteins in three large-scale experiments.[1–3] Using a similar strategy, over 1,200 interactions have been identified between proteins of the human gastric pathogen *Helicobactor pylori*, connecting by interactions about 47 per cent of the proteins encoded in the *H. pylori* genome.[4] Similar interaction screens are being performed for proteins of *Caenorhabditis elegans*.[5]

Databases combining such large-scale interaction data with the hundreds of individual interactions reported in the literature are also available. The Database of Interacting Proteins (DIP) currently contains over 9,500 interactions between approximately 5,700 proteins, a majority of which are from yeast and human.[6] Beyond cataloguing protein interactions, several groups have attempted to gather together all known pathways into databases, such as the metabolic pathway databases KEGG,[7] EcoCyc/MetaCyc[8] and the regulatory database TRANSFAC.[9] Like interaction data, phenotypic data are accumulating for mutants of thousands of genes in a few model organisms. Recent data of this sort include the measurement of disruption phenotypes of approx. half of the genes of

yeast,[10,11] the transposon mutagenesis of most of the genes of *Mycoplasma genitalium*,[12] and the RNAi silencing of many of the *C. elegans* genes.[13–15]

Protein expression data are accumulating much more slowly than interaction and phenotype data. Currently, few sets of protein expression data are publicly available, although more than 30 two-dimensional polyacrylamide gel electrophoresis (2D PAGE) analyses, each an observation of the expression level of thousands of proteins, are publicly available from the Swiss 2D-PAGE web server. However, a great deal of effort is being focused in this direction, and it is likely that these data will rapidly become available. Recent developments in mass-spectrometric techniques have allowed the first large-scale quantitative measurements of protein expression patterns, using two main approaches. In the first approach, tandem mass spectrometry of cysteine-containing peptides purified with sulphydryl-specific affinity reagents has yielded expression measurements of hundreds of proteins.[16,17] In a second approach, peptides generated from cell extracts were identified based upon capillary electrophoresis elution times and high-resolution mass measurements from Fourier transform ion cyclotron resonance mass spectrometry.[18,19] As with the first technique, expression levels were measured for thousands of peptides; each peptide was then mapped to its parent protein by database matching. Curiously enough, measured protein expression levels correlate poorly with mRNA expression levels in the two experiments in which this has been tested on a large scale, suggesting that this information will be a valuable complement to mRNA expression measurements[20,21] (Figure 1).

Large-scale quantitative measurements of protein expression levels are increasingly available

Measurements of metabolite concentrations as a function of cell condition, and spatial expression patterns of genes (the measurement of the locations in a cell or tissue where a particular gene is expressed) provide two additional types of data useful for functional annotation. Such data are currently either only sporadically available or have not been sufficiently developed to demonstrate their full potential.

Given sufficient data, genes with similar expression vectors can be assumed to function together

Spatial expression patterns of either

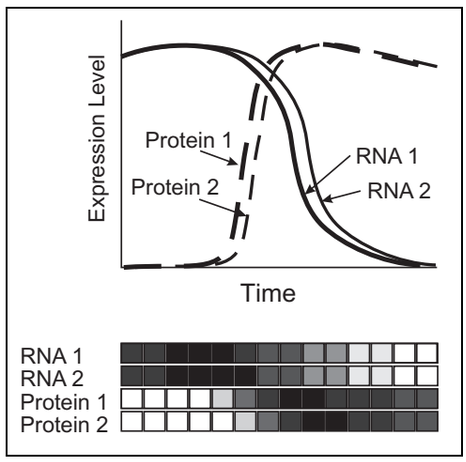

Figure 1: Protein and mRNA expression levels of several hundred genes have been only poorly correlated in two large-scale experiments.[20,21] This raises the question of whether it would be better to infer functional relationships between genes from correlations in their protein or correlations in their mRNA levels. However, this poor correlation distracts from the fact that both mRNA and protein expression patterns, while being largely uncorrelated with each other, are likely to both provide similar functional linkages between genes. In this illustration, two genes with related function have closely related mRNA expression patterns as well as closely related protein expression patterns. However, the mRNA patterns little resemble the protein expression patterns. In spite of this disagreement, a functional relationship can be inferred between genes 1 and 2 by analysis of either protein or mRNA co-expression; the functional link inferred between genes 1 and 2 is even stronger when both methods are used. The corresponding expression vectors are shown at the bottom of the figure, with grey-scale colouring indicating the level of expression at successive time points. Typically, a measure of the correlation between such expression vectors is calculated. Genes with correlated expression vectors can be assumed to function together provided the vectors are sufficiently complex

mRNA or proteins are technically difficult to collect and to compare quantitatively. However, one such project has been quite successful in collecting measurements of the cellular distribution of gene expression in *Xenopus* embryos.[22] Screening with randomly chosen cDNAs, the expression patterns of nearly 300 unique mRNAs have been detected by whole mount *in situ* hybridisation. The patterns were documented and collected in a database[23] (AxelDB). Early analysis of these data confirms that genes with common spatial expression patterns often have related functional roles.

The concentrations of hundreds of different metabolites are measured in the technique known as metabolic profiling

Metabolite expression patterns, collected in a process termed *metabolic profiling*, hold a great deal of promise for functional dissection of metabolic pathways. More generally, metabolic profiling appears to be a relatively straightforward method by which to build a quantitative description of the state of a cell. Metabolic profiling carries no spatial information, instead measuring the concentration of metabolites under a given set of cellular conditions. Two approaches seem most promising for measuring cellular metabolite concentrations: gas chromatography/mass spectrometry (GC/MS; eg see Fiehn *et al.*[24]) and nuclear magnetic resonance spectroscopy (NMR; eg see Raamsdonk *et al.*[25]). In both methods, cells are lysed and the filtered crude cell extract is analysed directly with minimal separation. In NMR, a proton spectrum of the crude extract is used as a data vector describing the state of the cells. Here, the NMR spectrum, corresponding to the composite spectrum of all metabolites in the cell, provides a quantitative measurement of the phenotype of the cell. This approach abrogates the need to identify individual metabolites. In GC/MS, the crude extract is first derivatised, then analysed to produce a data set of molecular masses and gas chromatographic elution times, which can be analysed in a fashion analogous to the analysis of microarray or NMR metabolite data. Using GC/MS techniques, Fiehn and co-workers quantified 326 distinct compounds from *Arabidopsis thaliana* leaf extracts and assigned metabolic phenotypes using the data collected during their experiments.[24]

INFERRING GENE FUNCTION FROM EXPRESSION DATA

Each of the types of experimental data discussed above can be used to infer functional linkages between genes. Specifically, the fact that two genes are co-expressed under many different conditions allows us to infer that the genes work together in the same pathway and therefore have related function.[26,27]

The microarray expression data for each gene can be written as a vector of expression levels measured under different conditions, such as the sample expression vectors at the bottom of Figure 1. Pairs of genes are considered co-expressed if their expression vectors are more similar than might be expected at random. For instance, if two genes, gene A and gene B, show similar expression levels over a range of time points or similar experimental conditions, and gene A is known to be involved in protein synthesis, then gene B might also be assigned a role in protein synthesis with some degree of statistical confidence.

DNA microarray experiments from multicellular organisms can reveal cell-specific expression patterns

Interestingly, this approach seems to work for multicellular organisms, even when cell-specific expression data have not been collected. In an analysis of 553 microarray experiments from *C. elegans*[28] genes exhibiting cell-specific expression, such as intestinal or muscle-specific expression, were discovered to be co-expressed even though only whole organism microarray data were collected. In this case, it is likely that the very large number of microarray experiments compensated for the lack of cell-specific data.

Non-normalised EST and SAGE expression libraries can be treated in a fashion analogous to microarray expression measurements, treating separate libraries as separate experiments measuring gene expression levels, and

approximating the expression of a gene as the frequency of observations of that gene in the library.[29] Protein expression data are currently sparse, but should be interpretable in much the same general manner.

Metabolic expression data require a somewhat different treatment. In microarray experiments, the expression of every gene is measured in one experiment. However, metabolite expression data would seem to require one experiment per gene, as the data vector associated with a gene often consists of the metabolite concentrations measured when that gene is disrupted. However, given measurements of metabolic profiles for sufficient gene disruptions, genes could be clustered by their metabolic profiles much as described above.[25] Other types of gene disruption phenotypes can be treated similarly: genes producing unusually similar knockout phenotypes are assumed to be functionally related.

Non-homology methods analyse the physical context in which genes occur

Spatial expression patterns appear to be the most difficult type of expression data to handle. Observed patterns must be converted into formalised descriptions of pattern morphology, which can then be compared with each other. In the case of *Xenopus* embryos, such a formal description was created for spatial gene expression patterns, allowing identification of co-expressed genes.[30]

Non-homology methods can be used to predict functional linkages between genes

INFERRING GENE FUNCTION FROM GENOME SEQUENCE DATA

Many functional inferences can be made purely on the basis of sequence data. At the most basic level, functional inferences from sequence data come in two general flavours: homology- and non-homology-based inferences. Homology-based inferences are those that derive from the straightforward comparison of sequences to find groups of similar sequences, such as by using the powerful BLAST[31] or Smith–Waterman[32] algorithms. When one of the members of a sequence group has been characterised, that gene's general function can be extended to the rest of the members of that sequence family (eg the sequence of gene *X* is similar to a serine kinase, therefore gene *X* is probably also a serine kinase). This approach is by far the dominant analysis performed on large gene sets, such as in the comparison of worm and yeast proteins by Chervitz and co-workers.[33]

Perhaps more pertinent for comparison with expression data are the non-homology methods. These methods operate on the principle that functionally linked genes will share common aspects of the contexts in which they occur.[34,35] Context in this sense refers to the physical locations of the genes, the nature and identity of the adjacent sequences, the organisms in which the genes occur, and so on. Genes whose contexts are more similar than expected by random chance can be linked by this approach.

Non-homology methods include the calculation of phylogenetic profiles describing the spectrum of organisms in which a gene is present and absent,[36] the search for conserved gene neighbours[37,38] or Rosetta Stone fusion proteins,[39,40] and the identification of genes that are physically closer to one another than expected.[41] Application of these analyses leads to prediction of functional linkages between genes that can be integrated easily with expression data.[27] Gene function can then be inferred on the basis of these linkages.

INTEGRATING EXPRESSION DATA WITH OTHER GENOME-WIDE DATA FOR FUNCTIONAL ANNOTATION

Expression data have been analysed largely by clustering genes into co-expression groups.[26] However, several other methods of handling expression measurements exist that allow integration of expression data with the sorts of large-scale data introduced above. Perhaps the simplest treatment is the superposition of gene or protein expression levels onto

networks of functionally linked proteins.[21,42] In this approach, illustrated in Figure 2, relationships between genes are first established by another method, such as measuring protein–protein interactions using large-scale yeast two-hybrid interaction assays or identifying metabolic pathways. Each protein in the network is labelled with its expression level under some standard conditions. Gene expression levels that were measured following a perturbation to the system can be interpreted more easily using this network model for visualisation. This approach can produce an understanding of which of the alternate pathways through a network are preferred by the cell[21,42,44] and has resulted in refinements to known biochemical pathways, such as the addition of several elements to the galactose utilisation pathway.[21]

A second method integrates the expression data more explicitly with other data types for the purpose of assigning genes into specific functional categories. In this approach, a 'feature vector' is calculated for each gene. The feature vector is a numerical list of quantitative properties of that gene, such as its expression under different conditions,[26,27,45] its conservation in different genomes,[36] the presence of various regulatory sites upstream of the

Superimposing expression data onto gene networks helps in interpreting complex gene expression patterns

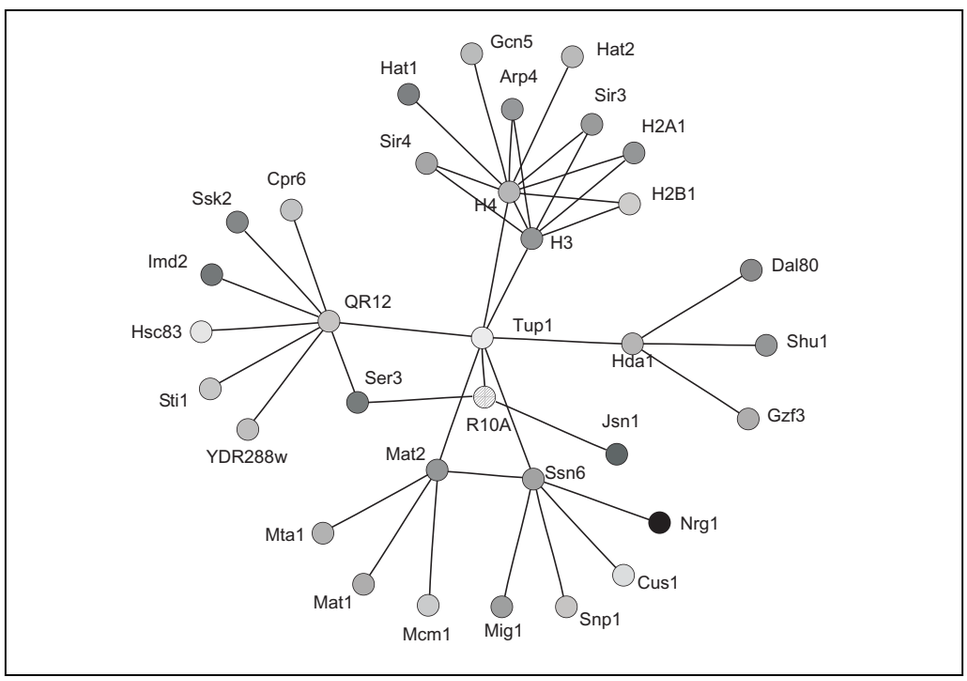

Figure 2: One approach to combining expression data with other genome-wide data has been simply to map expression data onto genetic or interaction networks derived by other methods. Here, a fraction of a protein interaction network of yeast[6] is illustrated centred on the transcriptional co-repressor protein Tup1. Each protein is illustrated as a shaded circle, and each experimentally observed protein–protein interaction is drawn as a line connecting the interacting partners. The intensity of shading in a circle reflects the level of expression of the gene measured by DeRisi and co-workers in a partial TUP1 knockout relative to gene expression in the wild type yeast.[43] Relative levels of expression are encoded as degrees of grey-scale, with white representing a strong decrease in expression and black representing a strong increase in expression of TUP1 cells relative to wild-type cells. For example, the expression of TUP1 can be seen to be decreased in the partial TUP1 knockout, while the expression of NRG1 is strongly increased. The network was plotted using the program Neato (AT&T Labs)

Gene function can be discovered by classification algorithms that learn to associate specific functions with patterns in the feature vectors

gene,[46,47] and so on. Given sufficiently informative feature vectors, genes with dissimilar functions can be distinguished from each other using standard pattern classification algorithms.[48] For the purpose of assigning genes into functional categories, an algorithm is trained to learn which features in the feature vector allow it to discriminate between a set of positive and negative examples. As illustrated in Figure 3, a properly trained algorithm could be applied to a new gene to predict if its function matches that of the genes in the training sets. One such discrimination algorithm is called a support vector machine[49] (SVM). SVMs rely on a geometric interpretation of the feature vectors, each vector describing the coordinates of a point in some high-dimensional space. The algorithms can be thought to work by finding a plane in an abstract high-dimensional space that separates the positive examples from the negative examples. SVMs have been applied with success on both gene expression data and phylogenetic profiles to assign genes into functional categories (eg 'cytoplasmic ribosome' or 'respiration').[45,50] A second type of discrimination algorithm working on entirely different principles is that of inductive logic programming and rule learning. Such algorithms suggest specific traits of the feature vector that provide the discrimination. A rule derived by this method, from the work of King *et al.*,[51] is 'If the percentage of lysine in the encoded protein is >6.5 per cent, then the gene functions in *macromolecular metabolism*'.

A third method provides an even more general mode for integrating disparate data for functional annotation. In this approach, pairwise relationships between genes are inferred based upon whatever data are available. Ideally, each linkage is ascribed a confidence or statistical significance. When taken all together, these pairwise linkages describe the organisation of genes into a network.[27] Figure 4 shows an example of such a network calculated for the genes of yeast. Function can then be assigned to uncharacterised genes based upon their linkages to characterised genes and pathways. Inferences from different methods can be easily combined in this framework, provided the measures of confidence are in some way comparable from method to method. This method also allows easy integration of protein interaction and pathway data, as these are naturally expressed as linkages between genes.

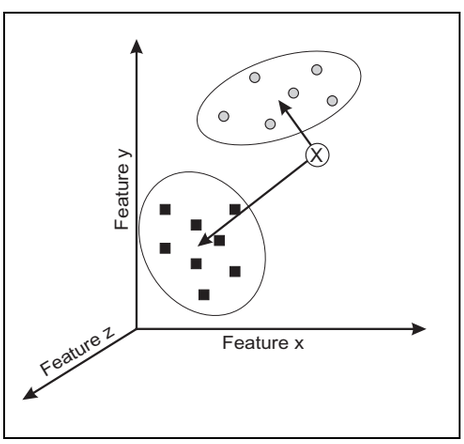

Figure 3: An illustration of a discrimination algorithm. Genes are mapped into an abstract high dimensional space by virtue of a vector of quantitative features, such as the expression vectors in Figure 1. A new gene (X) can be assigned a function by testing whether its position in this space is closer to the positive (grey circles) or negative (black squares) examples in a the training set of genes of known function. In this case, the features of the new gene show a better match to the features of the positive set, and the new gene would be assigned their function

Networks of genes can be reconstructed from pairwise functional inferences

MULTIPLE OBSERVATIONS LEAD TO ROBUST INFERENCES

Unfortunately for molecular biologists, few of these computational analyses, when taken in isolation, provide complete functional information for large numbers of genes. The computational functional inferences discussed above are largely independent of one another. Therefore, the inferences can be combined

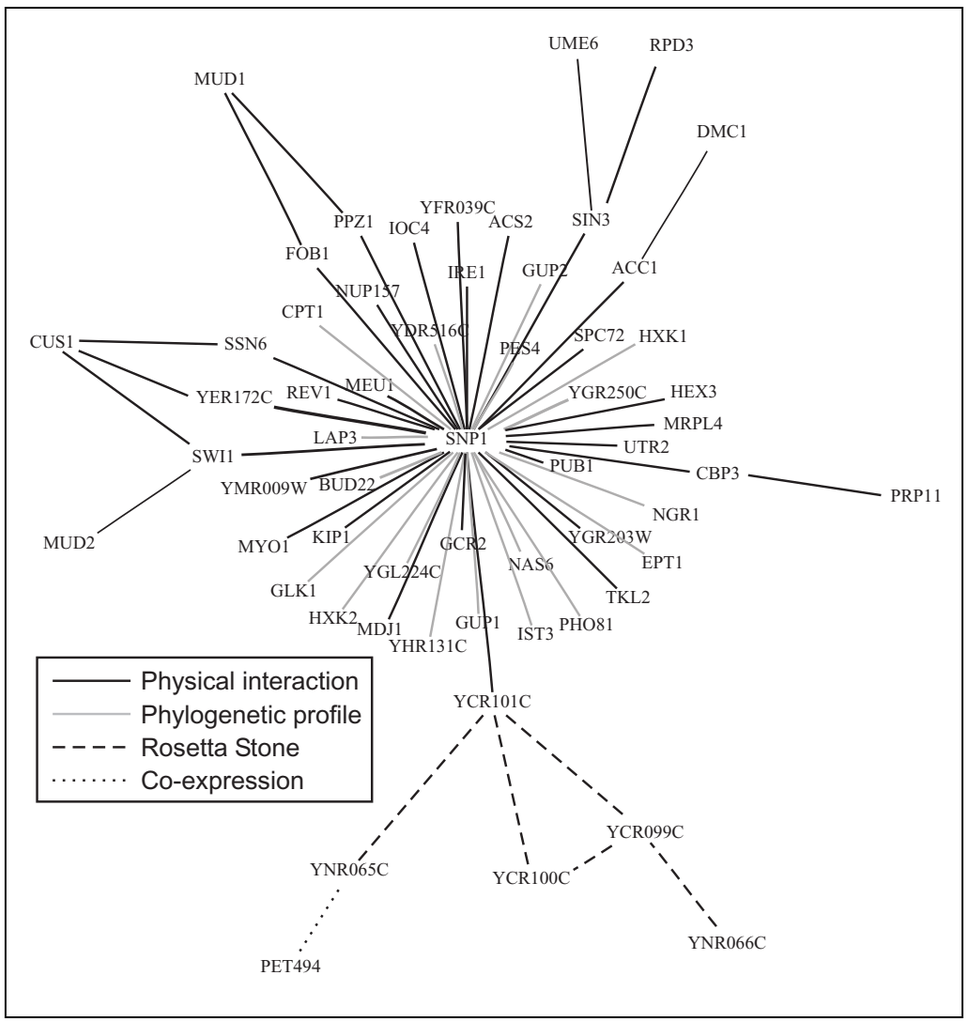

Figure 4: In this figure, expression data are integrated with data generated from other methods by expressing functional inferences from each method as a functional linkage between a pair of genes, then gathering together all pairwise linkages and plotting the resulting network. This network shows a subset of functional linkages between genes of yeast generated by non-homology methods such as phylogenetic profiles[36] and Rosetta Stone linkages,[39] as well as links between co-expressed yeast genes[27] and physically interacting proteins.[6] Here, the *SNP1* gene is linked to genes whose functions vary from RNA splicing to metabolism. Links generated by different methods tend to reinforce each other. For example, links predicted between *SNP1* and the hexokinase and glucokinase genes *HXK1* , *HXK2* and *GLK1* are supported by the observed physical interaction between *SNP1* and *GCR2* , a transcriptional activator involved in glycolytic gene expression

Redundant functional linkages derived using independent methods help to strengthen predictions

synergistically. If a linkage is derived from several independent methods, with each method *i* returning a probability of the link occurring by random chance $p(\text{link} \mid \text{method } i)$, the confidence in the linkage can be calculated as the product of the probabilities from each method, $\Pi_i\, p(\text{link} \mid \text{method } i)$.

In truth, the methods discussed above are not strictly independent. Nonetheless, they are generally derived from different underlying principles (eg 'functionally related genes have related expression' versus 'functionally related genes are co-inherited'), so combinations of these methods provide extremely robust

functional inferences.[27,50] Algorithms are available that explicitly measure the dependencies between different data sets and then weight the information accordingly. This process is particularly easy for data organised into feature vectors, in which the statistical covariance between each pair of columns in the matrix of feature vectors can be calculated to describe the relative independence of the features.

EXPRESSION ORTHOLOGUES

One possibility for increasing the strength of functional inferences derived from expression data is to combine expression data from different organisms. For example, the yeast glycolytic genes glyceraldehyde-3-phosphate dehydrogenase, phosphoglycerate mutase, phosphoglycerate kinase and enolase are strongly co-expressed across many conditions; the mouse homologues show a similar coexpression.[26,52] However, systematic analysis of expression and orthology is lacking. By definition, this approach will only work for conserved systems of genes. However, the potential exists for substantially reducing the number of false positive functional partners of a gene by imposing the simultaneous restrictions of sequence conservation of both partners across two or more organisms and co-expression of both partners in each of these organisms (Figure 5). This is a very strict requirement, and as more expression data become available from differing organisms, this idea can be put to the test more thoroughly.

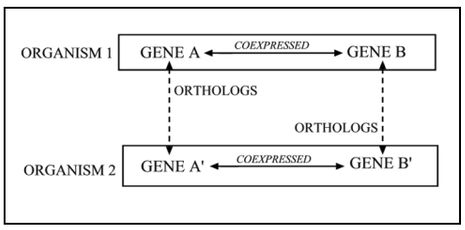

Figure 5: One approach for increasing the accuracy of functional inferences from expression data is to calculate expression orthologues. In this case, genes A and B co-express in one organism; the orthologues of A and B co-express in a second organism. Such pairs of genes have multiple constraints: they must be present in multiple organisms and co-express in each. The probability of satisfying such multiple constraints by random chance is quite low. As a consequence, such genes are expected to be tightly functionally coupled

Yeast and mouse glycolytic genes show similar co-expression patterns

INTEGRATING EXPRESSION DATA WITH BIOMEDICAL LITERATURE IN AN AUTOMATED FASHION

Co-citation or bibliometric approaches aim to link genes whose names co-occur in literature

Finally, many biologists wish to integrate the biological literature with the other genome-wide data sets. Working with biological literature, a very large corpus of text with more than 25 years worth of abstracts and papers online, can in fact prove quite difficult. Problems arise because of the sheer variation in language and descriptions: terms change over time, synonyms exist for many genes and techniques, and authors often have individual preferences for describing their work. Lack of standardisation, while making the literature much more interesting to read, makes extracting information automatically a tedious task. Many groups are focusing on how to extract data efficiently, as witnessed by the devotion of a special session to this topic at the coming 2002 Pacific Symposium of Biocomputing. These efforts have led to the establishment of a number of databases of gene and protein function, many of which are described in the annual January database issue of *Nucleic Acids Research*.

However, one very simple approach has been used to extract gene function from Medline without the difficult step of reading the literature. This approach, termed '*co-citation*' or '*bibliometrics*', creates links between genes whose names are mentioned in the same article.[53] The more frequently the genes are co-cited rather than cited independently, the

tighter the association inferred. This clever approach allows very easy extraction of known functional linkages between genes. The links are reasonably accurate, given their unusual derivation, and so provide one with means to capture the thousands of known relationships for integration with linkage data produced by other methods. Recently, Jenssen and co-workers[54] used this method to create a gene-to-gene co-citation network for 13,712 named human genes by analysis of titles and abstracts in over 10 million Medline records, and were able to show that literature co-occurrence associates biologically related genes. Analysis of the links suggests an accuracy of approximately 60–70 per cent. These co-citation links can be used to benchmark the pairwise links generated by other methods or can even be added to the other sets of links to produce more complete gene networks.

Literature co-citation analyses are generally consistent with known biological relationships between genes

CONCLUSIONS

It is an obvious fact that the increasing amounts and sources of biological data require creative methods of integration:[55] each of the sources of data, from expression measurements to genome sequences, gives only partial clues to the functions of genes. The inference of function by these methods may have an intrinsic error rate, as even functional assignment by direct sequence homology shows an appreciable error rate.[56] Even the existing annotation for genes of known function is likely to be error-prone.[57] Compounding this inherent difficulty in inferring function is the fact that genes are probably likely to play multiple roles in the cell, as confirmed by large-scale protein interaction assays showing that the majority of cellular systems are very closely interlinked.[1–3] It is therefore likely that large-scale functional inference will only come from integrating all of the disparate data available, each piece of data allowing some small inferences to be made. Though these seem to be small, faltering steps, each takes us closer to our final goal of deciphering the genome.

Error rates of functional links can be reduced by integrating data from many different sources

References

1. Uetz, P., Giot, L., Cagney, G. *et al.* (2000), 'A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*', *Nature*, Vol. 403(6770), pp. 623–627.
2. Ito, T., Tashiro, K., Muta, S. *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*, Vol. 97(3), pp. 1143–1147.
3. 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*, Vol. 98(8), pp. 4569–4574.
4. Rain, J. C., Selig, L., De Reuse, H. *et al.* (2001), 'The protein–protein interaction map of *Helicobacter pylori*', *Nature*, Vol. 409(6817), pp. 211–215.
5. Walhout, A. J. M., Sordella, R. Lu, X. *et al.* (2000), 'Protein interaction mapping in *C. elegans* using proteins involved in vulval development', *Science*, Vol. 287(5450), pp. 116–122.
6. Xenarios, I., Fernandez, E., Salwinski, L. *et al.* (2001), 'DIP: The Database of Interacting Proteins: 2001 update', *Nucleic Acids Res.*, Vol. 29(1), pp. 239–241.
7. Kanehisa, M. and Goto, S. (2000), 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Res.*, Vol. 28(1), pp. 27–30.
8. Karp, P. D., Riley, M., Saier, M. *et al.* (2000), 'The EcoCyc and MetaCyc databases', *Nucleic Acids Res.*, Vol. 28(1), pp. 56–59.
9. Wingender, E., Chen, X., Fricke, E. *et al.* (2001), 'The TRANSFAC system on gene expression regulation', *Nucleic Acids Res.*, Vol. 29(1), pp. 281–283.
10. Winzeler, E. A., Shoemaker, D. D., Astromoff, A. *et al.* (1999), 'Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis', *Science*, Vol. 285(5429), pp. 901–906.
11. Ross-Macdonald, P., Coelho, P. S., Roemer, T. *et al.* (1999), 'Large-scale analysis of the yeast genome by transposon tagging and gene disruption', *Nature*, Vol. 402(6760), pp. 413–418.
12. Hutchison, C. A., Peterson, S. N., Gill, S. R. *et al.* (1999), 'Global transposon mutagenesis and a minimal *Mycoplasma* genome', *Science*, Vol. 286(5447), pp. 2165–2169.
13. Fraser, A. G., Kamath, R. S., Zipperlen, P. *et al.* (2000), 'Functional genomic analysis of *C. elegans* chromosome I by systematic RNA

interference', *Nature*, Vol. 408(6810), pp. 325–330.

14. Gonczy, P., Echeverri, G., Oegema, K *et al.* (2000), 'Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III,' *Nature*, Vol. 408, pp. 331–336.
15. Maeda, I., Kohara, Y., Yamamoto, M. and Sugimoto, A. (2001), 'Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi', *Curr. Biol.*, Vol. 11(3), pp. 171–176.
16. Gygi, S. P., Rist, B., Gerber, S. A. *et al.* (1999), 'Quantitative analysis of complex protein mixtures using isotope-coded affinity tags', *Nature Biotechnol.*, Vol. 17(10), pp. 994–999.
17. Conrads, T. P., Alving, K., Veenstra, T. D. *et al.* (2001), 'Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling', *Anal. Chem.*, Vol. 73(9), pp. 2132–2139.
18. Conrads, T. P., Anderson, G. A., Veenstra, T. D. *et al.* (2000), 'Utility of accurate mass tags for proteome-wide protein identification', *Anal. Chem.*, Vol. 72(14), pp. 3349–3354.
19. Jensen, P. K., Pasa-Tolic, L., Peden, K. K. *et al.* (2000), 'Mass spectrometric detection for capillary isoelectric focusing separations of complex protein mixtures', *Electrophoresis*, Vol. 21(7), pp. 1372–1380.
20. Gygi, S. P., Rochon, Y., Franza, B. R. and Aebersold, R. (1999), 'Correlation between protein and mRNA abundance in yeast', *Mol. Cell. Biol.*, Vol. 19(3), pp. 1720–1730.
21. Idekar, T., Thorsson, V., Ranish, J. A. *et al.* (2001), 'Integrated genomic and proteomic analyses of a systematically perturbed metabolic network', *Science*, Vol. 292, pp. 929–934.
22. Gawantka, V., Pollet, N., Delius, H. *et al.* (1998), 'Gene expression screening in *Xenopus* identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning', *Mech. Dev.*, Vol. 77(2), pp. 95–141.
23. Pollet, N., Schmidt, H. A., Gawantka, V. *et al.* (2000), 'Axeldb: a *Xenopus laevis* database focusing on gene expression', *Nucleic Acids Res.*, Vol. 28(1), pp. 139–140.
24. Fiehn, O., Kopka, J., Dormann, P. *et al.* (2000), 'Metabolite profiling for plant functional genomics', *Nature Biotechnol.*, Vol. 18(11), pp. 1157–1161.
25. Raamsdonk, L. M., Teusink, B., Broadhurst, D. *et al.* (2001), 'A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations', *Nature Biotechnol.*, Vol. 19(1), pp. 45–50.
26. 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*, Vol. 95(25), pp. 14863–14868.
27. Marcotte, E. M., Pellegrini, M., Thompson, M. J. *et al.* (1999), 'A combined algorithm for genome-wide prediction of protein function', *Nature*, Vol. 402(6757), pp. 83–86.
28. Kim, S. K., Lund, J., Kiraly, M. *et al.* (2001), 'A gene expression map for *Caenorhabditis elegans*', *Science*, Vol. 293(5537), pp. 2087–2092.
29. Walker, M. G., Volkmuth, W. and Klingler, T. M. (1999), 'Pharmaceutical target discovery using guilt-by-association: schizophrenia and Parkinson's disease genes', in 'Proc. 7th International Conference on Intelligent Systems for Molecular Biology', August 6–10, AAAI Press, Menlo Park, CA, pp. 282–286.
30. Pollet, N., Schmidt, H. A., Gawantka, V. *et al.* (2000), '*In silico* analysis of gene expression patterns during early development of *Xenopus laevis*', *Pac. Symp. Biocomput.*, pp. 443–454.
31. Altschul, S. F., Gish, W., Miller, W. *et al.* (1990), 'Basic local alignment search tool', *J. Mol. Biol.*, Vol. 215(3), pp. 403–410.
32. Smith, T. F. and Waterman, M. S. (1981), 'Identification of common molecular subsequences', *J. Mol. Biol.*, Vol. 147(1), pp. 195–197.
33. Chervitz, S. A., Aravind, L., Sherlock, G. *et al.* (1998), 'Comparison of the complete protein sets of worm and yeast: Orthology and divergence', *Science*, Vol. 282(5396), pp. 2022–2028.
34. Marcotte, E. M. (2000), 'Computational genetics: finding protein function by nonhomology methods', *Curr. Opin. Struct. Biol.*, Vol. 10(3), pp. 359–365.
35. Huynen, M., Snel, B., Lathe, W. and Bork, P. (2000), 'Exploitation of gene context', *Curr. Opin. Struct. Biol.*, Vol. 10, pp. 366–370.
36. Pellegrini, M., Marcotte, E. M., Thompson, M. J. *et al.* (1999), 'Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles', *Proc. Natl Acad. Sci. USA*, Vol. 96(8), pp. 4285–4288.
37. Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998), 'Conservation of gene order: a fingerprint of proteins that physically interact', *Trends Biochem. Sci.*, Vol. 23(9), pp. 324–328.
38. Overbeek, R., Fonstein, M., D'Souza, M. *et al.* (1999), 'The use of gene clusters to infer functional coupling', *Proc. Natl Acad. Sci. USA*, Vol. 96(6), pp. 2896–2901.
39. Marcotte, E. M., Pellegrini, M., Ng, H. L. *et al.* (1999), 'Detecting protein function and protein–protein interactions from genome sequences', *Science*, Vol. 285(5428), pp. 751–753.

40. 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*, Vol. 402(6757), pp. 86–90.
41. 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*, Vol. 97(12), pp. 6652–6657.
42. Zien, A., Kuffner, R., Zimmer, R. and Lengauer, T. (2000), 'Analysis of gene expression data with pathway scores', in 'Proc. 8th International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 407–417.
43. DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997), 'Exploring the metabolic and genetic control of gene Expression on a genomic scale', *Science*, Vol. 278(5338), pp. 680–686.
44. Marcotte, E. M. (2001), 'The path not taken', *Nature Biotechnol.*, Vol. 19, pp. 626–627.
45. Brown, M. P., Grundy, W. N., Lin, D. *et al.* (2000), 'Knowledge-based analysis of microarray gene expression data by using support vector machines', *Proc. Natl Acad. Sci. USA*, Vol. 97(1), pp. 262–267.
46. Pavlidis, P., Furey, T. S., Liberto, M. *et al.* (2001), 'Promoter region-based classification of genes', *Proc. Pac. Symp. Biocomput.*, Jan. 3–7, pp. 151–163.
47. Tavazoie, S., Hughes, J. D., Campbell, M. J. *et al.* (1999), 'Systematic determination of genetic network architecture', *Nature Genet.*, Vol. 22, pp. 281–285.
48. Duda, R. O. and Hart, P. E. (1973), 'Pattern Classification and Scene Analysis', John Wiley & Sons, New York.
49. Cortes, C. and Vapnik, V. (1995), 'Support vector networks', *Machine Learning*, Vol. 20, pp. 273–293.
50. Pavlidis, P., Weston, J., Cai, J. and Grundy, W. N. (2001), 'Gene functional classification from heterogeneous data', in 'Proc. 5th Int. Conf. Comp. Mol. Biol.', 21–24 April, ACM Press, New York, pp. 242–248.
51. King, R. D., Karwath, A., Clare, A. and Dehaspe, L. (2000), 'Accurate prediction of protein functional class from sequence in the *Mycobacterium tuberculosis* and *Escherichia coli* genomes using data mining', *Yeast*, Vol. 17(4), pp. 283–293.
52. Miki, R., Kadota, K., Bono, H. *et al.* (2001), 'Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays', *Proc. Natl Acad. Sci. USA*, Vol. 98(5), 2199–2204.
53. Stapley, B. J. and Benoit, G. (2000), 'Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts', *Pac. Symp. Biocomput.*, pp. 529–540.
54. Jenssen, T. K., Laegreid, A., Komorowski, J. and Hovig, E. (2001), 'A literature network of human genes for high-throughput analysis of gene expression', *Nature Genet.*, Vol. 28(1), pp. 21–28.
55. Vidal, M. (2001), 'A biological atlas of functional maps', *Cell*, Vol. 104, pp. 333–339.
56. Devos, D. and Valencia, A. (2000), 'Practical limits of function prediction', *Proteins*, Vol. 41(1), pp. 98–107.
57. Brenner, S. E. (1999), 'Errors in genome annotation', *Trends Genet.*, Vol. 15(4), pp. 132–133.