Whole-genome expression analysis: challenges beyond clustering

Russ B Altman* and Soumya Raychaudhuri†

Measuring the expression of most or all of the genes in a biological system raises major analytic challenges. A wealth of recent reports uses microarray expression data to examine diverse biological phenomena – from basic processes in model organisms to complex aspects of human disease. After an initial flurry of methods for clustering the data on the basis of similarity, the field has recognized some longer-term challenges. Firstly, there are efforts to understand the sources of noise and variation in microarray experiments in order to increase the biological signal. Secondly, there are efforts to combine expression data with other sources of information to improve the range and quality of conclusions that can be drawn. Finally, techniques are now emerging to reconstruct networks of genetic interactions in order to create integrated and systematic models of biological systems.

Addresses
Stanford Medical Informatics, 251 Campus Drive, MSOB X-215, Stanford, California 95305-5479, USA
*e-mail: altman@smi.stanford.edu
†e-mail: sxr@smi.stanford.edu

Current Opinion in Structural Biology 2001, 11:340–347

0959-440X/01/$ – see front matter

Abbreviation
SOM self-organizing map

Introduction

The enthusiasm about microarray expression data analysis in the bioinformatics community has been remarkable. The peer-reviewed conference proceedings in the field have often provided the initial presentation of new methods, including the early application of clustering [1], linear decomposition [2] and algorithms to discern genetic networks [3•–5•]. (All references to the *Pacific Symposium on Biocomputing* can be found at http://www.smi.stanford.edu/projects/helix/psb-online/) The public release of expression data sets [6–8] created a *de facto* set of benchmarks for analysis by the bioinformatics community. There remains a risk, however, that the community has tuned these algorithms to perform well on this small set of training examples and that the algorithms will not perform well on entirely new data sets. Thus, the continued release of data from different groups using different detailed methods, and even measurements from redundant experiments, will be critical [9].

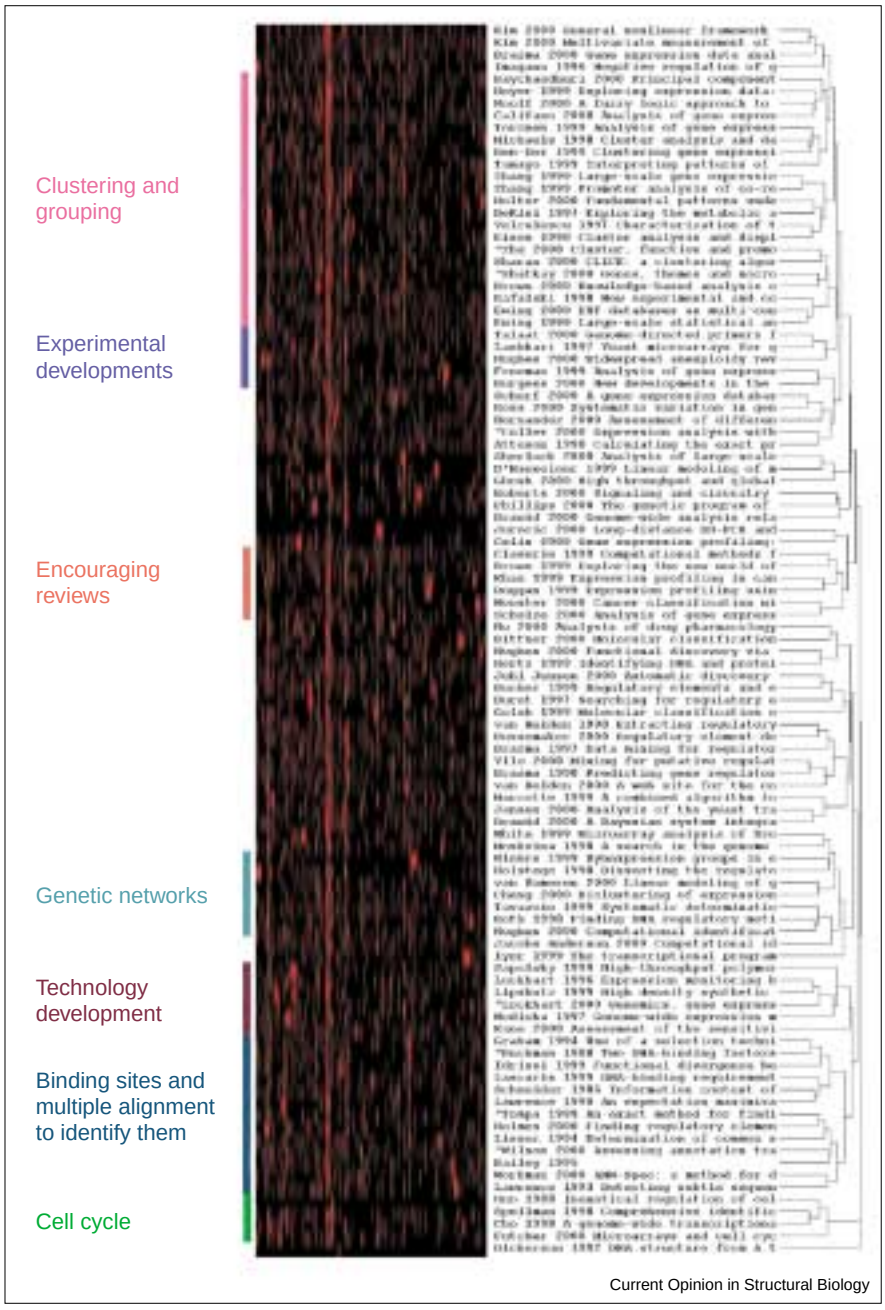
In a typical array experiment, many genes (frequently all known) in an organism are assayed under multiple conditions. The data can be represented as a matrix in which the rows are genes and the columns are conditions. These conditions may be different time points during a biological process, such as the yeast cell cycle [7,8] and *Drosophila* development [10], or they can be different tissue samples with some common phenotype, such as tissue type or malignancy. Although the amount of data generated in an expression experiment is tremendous, this is not yet a data-rich analytical task by statistical standards. The complexity of genomic systems, with N genes and thus N^2 potential pairwise interactions (not to mention higher order interactions), is even larger than the expression data sets and thus the ratio of data to unknown variables is still small. The major initial efforts at clustering and linear decomposition (such as principal components analysis) not only assist humans in understanding the data, but also demonstrate that the amount of independent new information may be much smaller than the number of raw data points suggests [2,11]. (Some microarray analysis tools are available at http://classify.stanford.edu/)

Whole-genome expression data affect structural biology by providing valuable functional information about when and where a protein is expressed, when it is degraded and with which other proteins it may interact. Early work has surveyed the ability of expression data to yield clues about common sequential/structural motifs for regulatory elements (as reviewed below). It also addresses issues such as protein localization or the justifiability of predicting function using 'guilt-by-association' techniques, whereby similar expression may be a component of the association (as reviewed below). Jansen and Gerstein [12] have analyzed the sequential and structural features of highly expressed genes and found biases (more alanine/glycine, less asparagine, shorter sequences, more TIM barrels) in a group of highly expressed proteins.

Although not the main focus of this review, there has been a satisfying focus on maximizing the reproducibility and analyzability of microarray experiments [9,13–15]. The 'fold difference' is widely used as a quantitative measure of the differential expression. The fold difference is the ratio of the expression in cells of interest versus the control cells. Genes expressed at low levels require higher fold differences in order to rise above the noise [16•,17] and duplicate measurements of identical experiments can be very valuable for reducing noise and simplifying subsequent analysis [9]. There are also emerging methods for assigning confidence to differentially expressed genes [18]. The noise in expression data can confound analyses and rank data are often more robust than absolute measured values because of the variation in methods for subtracting out background noise and quantifying expression levels [19,20]. Methods have emerged for imputing missing data in incomplete data sets (O Troyanskaya *et al.*, unpublished data; see Now in press). Finally, aneuploidy (and therefore the number of copies of a gene in the effective genome) has been shown to affect the expression level of a gene, either confounding the analysis or providing insight into the mechanisms of abnormal biology [21,22].

Figure 1

A clustering of 101 recent articles on whole-genome expression. For each article, the words in the titles and abstracts were extracted and filtered. 614 words that showed up in fewer than 90 and more than 3 articles were selected. Word vectors consisting of word counts for each article were created and normalized to avoid biases resulting from length. Complete linkage hierarchical clustering was used with an uncentered correlation metric and the tree was generated with TreeView [35]. Labels (left) indicating the subject of the paper were added manually based on our understanding of the contents of the papers that clustered together. The articles (right) are identified by first author, year and first words of title. Red spots indicate the presence of one of the 614 words in the associated article. The data used to create the figure, as well as a full online bibliography, are presented at http://www.smi.stanford.edu/projects/helix/pubs/cosb-01/.



The remainder of our review is organized around the result of a hierarchical clustering of the literature, in which the word counts are the features of articles used to cluster them, as shown in Figure 1.

A breadth of applications in biology and medicine

The number and diversity of microarray expression data measurements in the literature are impressive, and reports now appear in speciality journals in both biology and medicine. Initial data sets are often reported as genome-scale 'reviews' of a specific process, with subsequent analysis focusing on particular biological questions. Many reports, however, compare only a single pair of conditions and these are more difficult to evaluate because not all the differences between the two conditions are necessarily statistically or biologically significant.

The use of expression arrays to understand cancer has been attractive because most cancers are complex multigenic diseases and there is a natural 'control' group for the analysis — the noncancerous tissues. Initial studies have demonstrated the potential power of this technology for typing cancers and predicting prognosis. Golub *et al.* [23]

analyzed two subtypes of leukemia and created a classification algorithm that distinguished between the two subclasses, based only on expression patterns. They introduced self-organizing maps (SOMs) for clustering and rediscovered a known leukemia subclass. Alizadeh *et al.* [24] looked at diffuse B-cell lymphomas and identified a subtype with a distinct expression pattern correlating with particular clinical implications, such as the expected survival time. Bittner *et al.* [25••] looked at human genes in melanoma cells and found a group associated with lower invasive ability, reduced motility and (possibly) lower death rates. The genes that distinguished these groups are involved in the motility and invasion processes. This work is notable because the clusters were tested for robustness by assessing the sensitivity of the results to pertubations with random noise. Ross *et al.* [26] analyzed the expression of 8000 genes over 60 cancer cell lines from the National Cancer Institute (NCI) and showed that the cell lines clustered into groups that reflected the tissues of origin, suggesting that expression data may assist in assigning the primary tumor to metastases of unknown primary origin. Other classification techniques have an ability to distinguish between normal and malignant tissue using expression data at >90% accuracy [27]. Expression patterns may also explain mechanisms of sensitivity to drugs, as suggested by an analysis of sensitivity to L-asparaginase and 4-fluorouracil in the study of 60 NCI cell lines with known drug sensitivities [28••]. That report concluded with an intriguing table relating 1376 gene expression patterns to 118 drugs.

Other biological applications

Starting with the seminal observational papers [6–8] studying basic processes in yeast, such as metabolism, cell cycle and sporulation, there has been a new round of more directed studies involving wide-scale genetic manipulations. Holstege *et al.* [29] knocked out components of the transcription initiation machinery and studied essential cofactors and genes that modulate the response to environmental conditions. Roberts *et al.* [30••] perturbed elements of the mitogen-activated protein (MAP) kinase pathways and found interactions and shared elements between them. Their work is significant because it moves away from observational studies to a more hypothesis-driven mode of expression analysis — thus combining the strengths of traditional genetics with genome-wide high-throughput analysis. In an impressive study, Hughes *et al.* [31••] systematically studied the effect of 300 conditions, mostly gene deletions, on expression. They were able to assign the function of unknown genes by comparing the expression profiles from strains in which the gene is deleted with those from other deletion strains.

Other biologically significant studies include analysis of the fibroblast cell response to serum [32], the expression patterns following activation of the C-MYC helix-loop-helix protooncogene [33], and the expression program of hemapoietic stem cells [34]. White *et al.* [10] have followed the whole organism expression of *Drosophila* genes over time in order to understand the program of development and have found novel genes that appear to be associated with metamorphosis.

Clustering

The early uses of hierarchical clustering and SOMs on expression data provided a focal point for the introduction of alternative clustering methods. As with BLAST, clustering has become a basic tool for biologists in the field of expression analysis. Although there is a mature statistical literature about clustering, microarray data has sparked the development of multiple new methods. The initial excitement generated by the papers using hierarchical clustering [1,35] and SOMs (which arrange clusters spatially) [36,37] lead to a flurry of papers on fast and robust clustering methods [27,38–40]. The most promising innovations in this area may be the cluster methods that combine clustering of genes along with the conditions in a two-dimensional clustering. These address the limitation of some tree-based clusters that do not provide information about the degree of similarity between branches, and may be useful in recognizing reusable genetic 'modules' that are mixed and matched in order to create more complex genetic responses. For example, glucose metabolism may be invoked for a variety of otherwise disparate conditions (normal growth, stress, particular developmental stages) and so partial similarities among these conditions may be due to the shared glucose metabolism module, and not to a more general similarity. If such modules exist, cluster methods will need to associate genes in the context of particular conditions, in order to tease apart these associations. Thus, methods that can pull out subsets of genes associated with subsets of conditions are likely to be useful. Alon *et al.* [41] describe a two-way hierarchical clustering in which the order of subtrees is determined by the similarities of their associated conditions. Cheng *et al.* [42••] show a true biclustering method in which low-variance submatrices of the complete data matrix are found. These submatrices contain information about genes that may sometimes be correlated, but at other times are not. Califano *et al.* [43••] introduce a method to identify submatrices that differ with statistical significance from a set of control conditions.

Moving beyond clustering

After clustering is applied to an expression data set, we can examine those genes that cluster together and assign a function or value to the cluster. This approach may discover new associations, but in general rediscovers known associations and typically does not take full advantage of knowledge about known transcription factors, regulatory elements, sequence or structure information, or assigned gene functions. For example, there is interest in using information from genes with a common function to search for additional genes that share this function. Other efforts include the definition of regulatory elements using expression data and the combination of external data sources with expression data to validate new associations.

Using expression data to define regulatory elements

The co-expression of genes may imply that they share common regulatory mechanisms. This is a controversial hypothesis because regulatory mechanisms can be mixed and combined in ways that could lead to both convergent regulation (similar temporal expression patterns, different control strategies) and divergent regulation (similar control regions, put together in ways such that effect on expression is different). As in sequence analysis, expression can be similar (share significant features by some scoring method), but not homologous (common evolutionary origin). Nevertheless, this hypothesis underlies the study of upstream regions of genes and the search for regulatory elements guided by expression similarity. These methods are now routinely using expression clusters to guide the search for common motifs [44]. Notable approaches include that of Mandel-Gutfreund *et al.* [45••], who use 3D structural information about the protein–DNA binding site to analyze the effects of different mutations, and then evaluate the regions with a knowledge-based potential.

There are two general methods used for mining upstream regions to search for regulatory regions: first, oligomer-based methods; and second, statistical pattern-matching methods. Oligomer-based methods look for short patterns of nucleotides that occur in a statistically significant abundance, thus suggesting potential functional importance [46–51]. An automated pipeline for regulatory element discovery has been used to find potentially novel consensus patterns in yeast [52–54]. Juhl Jensen and Knudsen [55•] combine three sources of data (functional literature on a gene, short repeated subsequences found in upstream regions, and the expression behavior) to search for new regulatory sequences and find a new potential proteasomal upstream element.

Many statistical methods for finding regulatory elements are descendants of the pioneering work on Gibbs sampling, which constructs multiple sequence alignments using probabilistic models and local optimization [56], and the statistics of weight matrices for binding sites [57–59]. A new system can handle gapped motifs, motifs containing palindromes and imperfect input data sets, along with estimate of significance [60••]. Others have used similar technology, but focused on the location of the regulatory motifs relative to the coding regions, and have analyzed the entire yeast genome, finding 3311 motifs [61]. An interesting argument has been made that studying expression patterns first and then looking for regulatory elements may lose information, whereas the combined search for both clusters and promoters may be more efficient [62].

Combining expression data with other data sources

The most exciting work in the analysis of whole-genome expression has come with the combination of expression data with numerous other data sources, including the published literature, the DNA and protein sequence databases, the Protein Data Bank, and the functional taxonomies that are beginning to emerge.

Microarray expression data complements other data sources (including phylogenetic profiles, protein fusion in other organisms, metabolic function and annotated experimental functional studies) to allow functional predictions [63••]. Using expression alone to assign function has a relatively high false positive rate (36% of function assignments may not be accurate), but the volume of data still leads to many useful predictions. Yeast expression data also allows the classification (using Support Vector Machines, a general classification method) of the Munich Information Center for Protein Sequences (MIPS) yeast functional categories and their association with genes of unknown function [64••]. The justifiability of predicting function based on similar sequence, expression, location and other proxies should be carefully assessed. In the context of sequence identity, it seems that 40% identity implies close functional relationships, whereas 25% identity suggests more distant functional relationships [65]. Expression-based cell-cycle clusters provide a gold standard for evaluating a text-based assignment of genes into phases of the cell cycle [66•]. Expression patterns also provide information for creating rules that associate genes with functional categories [67], as provided by the Gene Ontology (GO; developed to give a standard set of terms for molecular and cellular functions, processes and compartments; http://www.geneontology.org/). Califano *et al.* [43••] use a database of conditions and associated phenotypes to build statistically significant expression patterns for each phenotype that are useful for understanding the phenotype and classifying new conditions.

Expression measurements form part of a data set that allows protein cellular localization to be predicted for yeast. In a data set including variables such as sequence signals, biophysical and structural features of molecules, as well as expression data, two of the top ten informative features are drawn from the expression data (absolute expression and standard deviation of expression). These features allow the assignment of about two-thirds of unlocated yeast proteins with about 75% accuracy [68••].

New directions: the reconstruction of genetic networks

A reductionist approach to studying model systems and isolating individual components is clearly the pillar upon which most biological knowledge rests. However, the understanding of interacting systems, for which approximations about isolation and crosstalk (normally made to simplify the systems) can no longer be made, constitutes a major challenge. Initial efforts in the representation and 'reverse engineering' of cellular networks containing genes, their regulators and their downstream targets have been demonstrated by McAdams and Shapiro [69] on lambda phage. The availability of detailed data about concentrations, binding constants, and regulatory relationships has, however, limited the applicability of these techniques. The arrival of expression data, particularly in the context of targeted mutation experiments [30••,31••], has raised

expectations that at least some of these data will make more modeling studies feasible. As discussed above, the abundance of data (compared to the number of parameters needed) is somewhat illusory, but the interest in regulatory and effector networks is clearly increasing.

The simplest methods for modeling the interactions of genes are Boolean networks, in which a 1 or 0 is used to express simply whether a gene is induced or not; the induction of each gene is a deterministic function of the state of a set of other genes. These representations are easy to compute with and require a minimum number of parameters to be estimated, but may be too simplified [70,71]. Similarly, it is possible to use linear modeling of gene interactions by representing the expression of a gene as a weighted linear combination of the expressions of all other genes. These methods are limited by the availability of data [4•,72]. An interesting new approach uses genetic programming techniques that have been successful in the design of computer logic chips to reverse engineer genetic networks, but results so far are on simulated data and relatively small networks [3•]. The most useful approach so far has been the use of expression data not to build a network (which requires more data than is available), but instead to evaluate two alternative network topologies. Friedman *et al.* [73] have explored the discovery of partial network information on the cell-cycle data using Bayesian belief networks — computer data structures that use probabilistic representations of discrete variables and their interdependencies to infer the most likely set of values for the variables. Hartemink *et al.* [5•] show that they can use Bayesian belief networks to distinguish between two competing models for galactose regulation in yeast, using data from 52 array experiments that were not designed to answer this question.

Conclusions

For some time, there were more review articles about the promises and problems with whole-genome expression analysis than there were primary research reports using the methods or introducing new analytic techniques. This imbalance is now being corrected and the community is thinking seriously about ways in which whole-genome expression data can be integrated with other biological knowledge to maximize its impact. The next few years may show major progress in our ability to understand the ways in which genomes implement their biological programs.

Update

Clustering methods are now more routinely being evaluated with respect to criteria such as robustness, computational cost, clarity of cluster definitions and reproducibility. A useful report by Yeung *et al.* [74•] introduces a leave-one-out type approach for testing cluster methods by evaluating their ability to predict the gene associations in a 'left out' data set. Herrero *et al.* [75] report and evaluate a self-organizing tree algorithm (SOTA) that shares features with SOMs, but imposes a binary tree structure on the data. Bussemaker *et al.* [76] showed that the expression of genes can be modeled without a preliminary clustering. Instead, they create a model of how any fragment (of length seven) in the upstream regions of genes can contribute (positively or negatively) to the overall expression of a gene. They create an additive model based on the sum of the logarithms of the expression and are able to explain 30% of the expression 'signal' with this simple model. Finally, Masys *et al.* [77] show the utility of interpreting expression data in the context of textual indexing terms in order to understand the biomedical significance of discovered clusters.

Acknowledgements

We would like to thank Josh Stuart, Olga Troyanskaya, Patrick Sutphin, Teri Klein and David Botstein for useful conversations. This work is supported by the Burroughs Wellcome Fund and grants from NIH GM-61374, LM-06422, GM-07365 and NSF DBI-9600637.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Michaels GS, Carr DB, Askenazi M, Fuhrman S, Wen X, Somogyi R: Cluster analysis and data visualization of large-scale gene expression data. *Pac Symp Biocomput* 1998:42-53.

2. Raychaudhuri S, Stuart JM, Altman RB: Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput* 2000:455-466.

3. • Koza JR, Mydlower JD, Lanza G, Yu J, Keanne MA: Reverse engineering of metabolic pathways from observed data using genetic programming. *Pac Symp Biocomput* 2001:434-445.
Genetic programming allows computer programs to evolve under selective pressure in order to maximize their performance on a given task. This paper is the first to apply these methods to genetic network reconstruction.

4. • van Someren EP, Wessels LF, Reinders MJ: Linear modeling of genetic networks from experimental data. *Ismb* 2000, 8:355-366.
Although expression data offer a wealth of data, the number of parameters for a genetic network may be far in excess. This group addresses this problem by first clustering the genes into a group of 'prototypical' patterns. This reduces the potential dimensionality of the problem. Subsequently, they solve a linear model to create a genetic network.

5. • Hartemink AJ, Gifford DK, Jaakkola TS, Young RA: Using graphical models and genomic expression data to statistically validate models of regulatory networks. *Pac Symp Biocomput* 2001:422-433.
Although large amounts of data are required to build a Bayesian network *de novo*, it is relatively easy to evaluate the compatibility of a network with a given set of data. The investigators encoded two models for galactose regulation and then scored them against experimental data. They were able to recover the correct network in yeast based on 52 expression arrays that were collected without this question in mind.

6. DeRisi JL, Iyer VR, Brown PO: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997, 278:680-686.

7. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ *et al.*: A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 1998, 2:65-73.

8. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the yeast ***Saccharomyces cerevisiae*** by microarray hybridization. *Mol Biol Cell* 1998, 9:3273-3297.

9. Butte A, Ye J, Niederfellner G, Rett K, Häring H, White M, Kohane I: Determining significant fold differences in gene expression analysis. *Pac Symp Biocomput* 2001:6-17.

10. White KP, Rifkin SA, Hurban P, Hogness DS: Microarray analysis of ***Drosophila*** development during metamorphosis. *Science* 1999, 286:2179-2184.

11. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV: Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci USA* 2000, 97:8409-8414.

12. Jansen R, Gerstein M: Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res* 2000, 28:1481-1488.

13. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ: Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* 2000, 28:4552-4557.

14. Talaat AM, Hunter P, Johnston SA: Genome-directed primers for selective labeling of bacterial transcripts for DNA microarray analysis. *Nat Biotechnol* 2000, 18:679-682.

15. Sengupta R, Tompa M: Quality control in manufacturing oligo arrays: a combinatorial design approach. *Pac Symp Biocomput* 2001:348-359.

16. • Tsien CL, Libermann TA, Gu X, Kohane IS: On the reporting of fold differences. *Pac Symp Biocomput* 2001:496-507.

This study addresses the issue of whether or not a particular expression value is truly meaningful, or just part of the noise. The authors created a tool to examine replicated data and then to mask insignificant fold differences in expression.

17. Claverie JM: Computational methods for the identification of differential and coordinated gene expression. *Hum Mol Genet* 1999, 8:1821-1832.

18. Manduchi E, Grant G, McKenzie S, Overton G, Surrey S, Stoeckert C: Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics* 2000, 16:685-698.

19. Park P, Pagano M, Bonetti M: A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac Symp Biocomput* 2001:52-63.

20. Raychaudhuri S, Stuart J, Liu X, Small P, Altman R: Pattern recognition of genomic features with microarrays: site typing of ***Mycobacterium tuberculosis*** strains. *Ismb* 2000, 8:286-295.

21. Klus G, Song A, Schick A, Wahde M, Szallasi Z: Mutual information analysis as a tool to assess the role of aneuploidy in the generation of cancer-associated differential gene expression patterns. *Pac Symp Biocomput* 2001:42-51.

22. Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ *et al.*: Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat Genet* 2000, 25:333-337.

23. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al.*: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999, 286:531-537.

24. Alizadeh A, Eisen M, Davis R, Ma C, Lossos I, Rosenwald A, Boldrick J, Sabet H, Tran T, Yu X *et al.*: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000, 403:503-511.

25. •• Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A *et al.*: Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000, 406:536-540.

On the basis of the gene expression patterns of clustered melanoma cell lines, the investigators proposed a subclass of cells that was not as invasive or mobile as other subclasses. This hypothesis was verified with *in vivo* studies. This paper is remarkable in its careful verification of the reproducibility of the clustered groups. The cluster was obtained with hierarchical clustering and verified with multidimensional scaling. Also, it is a prototypical microarray paper in that expression arrays are used to suggest an intriguing and unexpected hypothesis that is then verified in follow-up experiments.

26. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M *et al.*: Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000, 24:227-235.

27. Ben-Dor A, Shamir R, Yakhini Z: Clustering gene expression patterns. *J Comput Biol* 1999, 6:281-297.

28. •• Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT *et al.*: A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000, 24:236-244.

An impressive survey of the NCI60 cancer cell lines. The study compares the expression profiles of each of these cells to drug response and identifies examples of how variance in expression may relate to drug sensitivity and resistance.

29. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 1998, 95:717-728.

30. •• Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR *et al.*: Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 2000, 287:873-880.

This study utilized genetic manipulations in conjunction with expression array assays to specifically characterize interactions in protein kinase pathways. The deletional analysis provides evidence about which genes may control others. Similar to the study by Hughes *et al.* [31••], this study goes beyond observation of a process and actually genetically manipulates the organism to answer specific questions.

31. •• Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD *et al.*: Functional discovery via a compendium of expression profiles. *Cell* 2000, 102:109-126.

A powerful study in which 300 diverse mutations and chemical treatments of yeast are investigated using microarrays. The underlying hypothesis is that genes involved in the same process will elicit similar expression responses when rendered nonfunctional. The investigators are able to assign and confirm the function of eight uncharacterized genes that are involved in a variety of processes. This study (and those like it) demonstrates a departure from the early expression array studies, which were primarily observational. Along with this departure will no doubt come new analytical approaches that will fully exploit this development. Such data offer great potential for reconstructing the underlying genetic networks of an organism as deletions offer more directly causal information, instead of the more abundant correlative information.

32. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J Jr, Boguski MS *et al.*: The transcriptional program in the response of human fibroblasts to serum. *Science* 1999, 283:83-87.

33. Coller HA, Grandori C, Tamayo P, Colbert T, Lander ES, Eisenman RN, Golub TR: Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc Natl Acad Sci USA* 2000, 97:3260-3265.

34. Phillips RL, Ernst RE, Brunk B, Ivanova N, Mahan MA, Deanehan JK, Moore KA, Overton GC, Lemischka IR: The genetic program of hematopoietic stem cells. *Science* 2000, 288:1635-1640.

35. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998, 95:14863-14868.

36. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 1999, 96:2907-2912.

37. Toronen P, Kolehmainen M, Wong G, Castren E: Analysis of gene expression data using self-organizing maps. *FEBS Lett* 1999, 451:142-146.

38. Sharan R, Shamir R: CLICK: a clustering algorithm with applications to gene expression analysis. *Ismb* 2000, 8:307-316.

39. Sasik R, Hwa T, Iranar N, Loomis W: Percolation clustering: a novel algorithm applied to the clustering of gene expression patterns in ***Dictyostelium*** development. *Pac Symp Biocomput* 2001:335-347.

40. Heyer LJ, Kruglyak S, Yooseph S: Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* 1999, 9:1106-1115.

41. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 1999, 96:6745-6750.

42. Cheng Y, Church GM: Biclustering of expression data. *Ismb* 2000, •• 8:93-103.
This approach differs from other clustering approaches in that it considers both conditions and genes simultaneously. The algorithm finds subsets of conditions and genes (submatrices) that are homogeneous. Whereas traditional gene clustering approaches are mutually exclusive and try to identify genes that always behave identically, this approach seeks genes that for a set of conditions behave similarly, though they may have uncorrelated behaviors in the other conditions.

43. Califano A, Stolovitzky G, Tu Y: Analysis of gene expression •• microarrays for phenotype classification. *Ismb* 2000, 8:75-85.
The investigators introduce a novel method to identify patterns (submatrices) of interest within a given data set. The method requires two sets of array measurements: those taken on organisms, cell lines, and so on, with the phenotype of interest, and those without. The algorithm then proceeds to find all patterns for which the expression is significantly varied in the phenotype set compared to the nonphenotype set. These patterns can then be used to understand the phenotype and also to help classify unknown cases. An impressive array of demonstrations on several phenotypes including p53 mutation state and the drug response of a set of cancer cell-lines.

44. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: Systematic determination of genetic network architecture. *Nat Genet* 1999, 22:281-285.

45. Mandel-Gutfreund Y, Baron A, Margalit H: A structure-based •• approach for prediction of protein binding sites in gene-upstream regions. *Pac Symp Biocomput* 2001:139-150.
The investigators deviate from the more familiar multiple alignment sequence search approaches to identify novel binding sites. Rather, they use crystal structure information about the transcription factor and a knowledge-based potential to identify putative upstream binding regions. As more transcription factor crystal structures become available and our ability to predict unknown structures improves, approaches such as this one may become common.

46. Moskvina E, Schuller C, Maurer CT, Mager WH, Ruis H: A search in the genome of ***Saccharomyces cerevisiae*** for genes regulated via stress response elements. *Yeast* 1998, 14:1041-1050.

47. van Helden J, Andre B, Collado-Vides J: Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 1998, 281:827-842.

48. van Helden J, Andre B, Collado-Vides J: A web site for the computational analysis of yeast regulatory sequences. *Yeast* 2000, 16:177-187.

49. Zhu J, Zhang MQ: Cluster, function and promoter: analysis of yeast expression array. *Pac Symp Biocomput* 2000:479-490.

50. Bussemaker HJ, Li H, Siggia ED: Regulatory element detection using a probabilistic segmentation model. *Ismb* 2000, 8:67-74.

51. Jacobs Anderson JS, Parker R: Computational identification of cis-acting elements affecting post-transcriptional control of gene expression in ***Saccharomyces cerevisiae***. *Nucleic Acids Res* 2000, 28:1604-1617.

52. Brazma A, Jonassen I, Vilo J, Ukkonen E: Predicting gene regulatory elements ***in silico*** on a genomic scale. *Genome Res* 1998, 8:1202-1215.

53. Brazma A, Vilo J, Ukkonen E, Valtonen K: Data mining for regulatory elements in yeast genome. *Ismb* 1997, 5:65-74.

54. Vilo J, Brazma A, Jonassen I, Robinson A, Ukkonen E: Mining for putative regulatory elements in the yeast genome using gene expression data. *Ismb* 2000, 8:384-394.

55. Juhl Jensen L, Knudsen S: Automatic discovery of regulatory • patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics* 2000, 16:326-333.
This study proposes a statistic to identify whether a particular oligomer sequence is over-represented significantly in a positive sequence set compared to a negative sequence set. Presumably, if the positive set is a group of open reading frame upstream regions, these significant words may be binding sites. The use of a negative sequence set permits searching for significant words independent of expression data, for example, by using downstream sequences as a negative set. This method obtained a number of previously known binding sites and at least one potentially novel binding site when applied to yeast.

56. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 1993, 262:208-214.

57. Hertz GZ, Stormo GD: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 1999, 15:563-577.

58. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: Information content of binding sites on nucleotide sequences. *J Mol Biol* 1986, 188:415-431.

59. Workman CT, Stormo GD: ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput* 2000:467-478.

60. Liu X, Brutlag DL, Liu JS: BioProspector: discovering conserved •• DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 2001:127-138.
BioProspector is a modified version of the Gibbs sampler that permits the user to enter both a positive set to look for motifs and a negative set to create the background statistics in order to improve performance. BioProspector also contains a variety of other modifications that make it particularly suitable for binding-site searching; for example, it can look for palindromic sites and can identify motifs with variable-length gaps between them.

61. Hughes JD, Estep PW, Tavazoie S, Church GM: Computational identification of cis-regulatory elements associated with groups of functionally related genes in ***Saccharomyces cerevisiae***. *J Mol Biol* 2000, 296:1205-1214.

62. Holmes I, Bruno WJ: Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Ismb* 2000, 8:202-210.

63. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: •• A combined algorithm for genome-wide prediction of protein function. *Nature* 1999, 402:83-86.
A variety of information is used to predict protein function. 'Links' between proteins are generated form expression correlation, similar phylogenetic behavior, and identification of fused domain proteins. These links are differentially weighted and used to predict protein function. The study provides a quantification of the values of each of these different pieces of information in function prediction.

64. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, •• Ares M Jr, Haussler D: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 2000, 97:262-267.
This is the first published application of Support Vector Machines (SVMs) on microarray data. Supervised machine learning allows users to identify whether or not a gene belongs to some predefined group based on its expression pattern. The investigators used expression data to automatically determine whether unknown genes were ribosomal, histones, involved in the tricarboxylic acid (TCA) cycle, involved in aerobic respiration, or in the proteosome complex, with good results.

65. Wilson CA, Kreychman J, Gerstein M: Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* 2000, 297:233-249.

66. Shatkay H, Edwards S, Wilbur WJ, Boguski M: Genes, themes and • microarrays: using information retrieval for large-scale gene analysis. *Ismb* 2000, 8:317-328.
This group introduces a method for using the biomedical literature to rapidly and automatically identify the function of genes. The approach is an example of Natural Language Processing (NLP). The value of NLP approaches will increase as expression array studies are interpreted in the context of the published literature on individual gene function.

67. Hvidsten T, Komorowski J, Sandvik A, Loegreid A: Predicting gene function from gene expressions and ontologies. *Pac Symp Biocomput* 2001:299-310.

68. Drawid A, Gerstein M: A Bayesian system integrating expression •• data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol* 2000, 301:1059-1075.
Attempts to determine localization of a protein by combining expression data with a plethora of other protein information such as sequence signals and a variety of biophysical properties. This study demonstrates the promise of combining information from multiple sources to attack difficult problems.

69. McAdams H, Shapiro L: Circuit simulation of genetic networks. *Science* 1995, 269:650-656.

70. Maki Y, Tominaga D, Okamoto M, Watanabe S, Eguchi Y: Development of a system for the inference of large-scale genetic networks. *Pac Symp Biocomput* 2001:446-458.

71. Akutsu T, Miyano S, Kuhara S: Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics* 2000, 16:727-734.

72. D'Haeseleer P, Wen X, Fuhrman S, Somogyi R: Linear modeling of mRNA expression levels during CNS development and injury. *Pac Symp Biocomput* 1999:41-52.

73. Friedman N, Linial M, Nachman I, Pe'er D: Using Bayesian networks to analyze expression data. *J Comput Biol* 2000, 7:601-620.

74. • Yeung KY, Haynor DR, Ruzzo WL: Validating clustering for gene expression data. *Bioinformatics* 2001, 17:309-318.
The investigators propose a method to validate clustering without any external expert knowledge. Gene clusters are validated by their ability to predict expression in a 'left out' experimental condition. Six clustering methods are compared over four separate gene expression data sets.

75. Herrero J, Valencia A, Dopazo J: A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* 2001, 17:126-136.

76. Bussemaker HJ, Li H, Siggia ED: Regulatory element detection using correlation with expression. *Nat Genet* 2001, 17:167-171.

77. Masys DR, Welsh JB, Lynn Fink J, Gribskov M, Klacansky I, Corbeil J: Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics* 2001, 17:319-326.

Now in press

The work referred to in the text as (O Troyanskaya *et al*., unpublished data) is now in press:

78. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman R: Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001, in press.