Gene expression informatics—it's all in your mine

Douglas E. Bassett Jr[1], Michael B. Eisen[2] & Mark S. Boguski[3]

[1]Rosetta Inpharmatics, Kirkland, Washington 98034, USA. [2]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA. [3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. e-mail: boguski@ncbi.nlm.nih.gov

Technologies for whole-genome RNA expression studies are becoming increasingly reliable and accessible. However, universal standards to make the data more suitable for comparative analysis and for inter-operability with other information resources have yet to emerge. Improved access to large electronic data sets, reliable and consistent annotation and effective tools for 'data mining' are critical. Analysis methods that exploit large data warehouses of gene expression experiments will be necessary to realize the full potential of this technology.

Only a decade ago, the concept of being able to simultaneously measure the concentrations of every transcript in the cell in a single experiment would have seemed like science fiction to most researchers. However, the advent of new technologies has empowered genome researchers to do just that, and has also added new dimensions to our ability to leverage information from genome sequencing projects into a more comprehensive and holistic understanding of cell physiology. Despite the excitement generated by these technologies, only a handful of research laboratories (along with many pharmaceutical and biotechnology companies) are currently carrying out genome-scale expression studies. Certainly the high cost and technical expertise required is an obstacle to many investigators who are interested in pursuing such studies, leading some to use updated versions of more traditional systems (for example, filters[1,2]) that may be more economical in terms of capital equipment outlays, operating expenses and array and reagent costs. Furthermore, the scientific community has not determined how to cope with the massive amounts of data to be explored and interpreted in the context of other sources of biological knowledge. Lastly, there are no universal standards that define how the results can be shared and distributed most effectively within the scientific community.

The number of review articles on gene expression technologies probably exceeds the number of primary research publications in this field. This is not the result of any paucity of primary data; for example, at Stanford and Rosetta alone, more than 30 million independent gene expression measurements (one gene, one condition) have been made during the past two years. There are, however, a limited number of efficient, publicly available tools for data processing, storing and retrieving the information and analysing the results in the context of existing knowledge. In addition, there is no consensus on how to compare the results obtained using different technologies (for example, microarrays[3] versus oligonucleotide 'chips'[4] versus SAGE[5]) and how to communicate results using existing publication modalities and public database systems[6]. In this review, we will discuss some of the technical and intellectual issues involved in these processes, describe some of the ways in which they are currently being addressed and provide some thoughts about future directions.

Processes and information flow

Laboratory information management systems and databases. Large-scale, high-throughput experimental methods require material and information processing systems to match (Fig. 1). In addition to being used to locate and track physical resources (for example, clones, arrays or probes), computer systems must manage very large quantities of data both before and after an experiment. They may also need to interface with laboratory instrumentation—for example, controlling robotic 'printing' of microarrays (see pages 11 (ref. 7) and 16 (ref. 8) of this issue) and correlating array elements with specific microtitre plates and wells. Following hybridization of a microarray and the readout of gene expression levels, the data must be stored so that they are available for image processing[9] and statistical and biological analysis.

Image analysis. A variety of software tools have been developed for use in processing array images (Table 1). The basic goal is to reduce an image of spots of varying intensities into a table with a measure of the intensity (or, for multi-coloured fluorescence images, the ratio of intensities) for each spot. Although this is a relatively straightforward goal, there is as yet no common manner of extracting this information, and many research groups are still writing customized software for this purpose. Scanning and image processing are currently resource-intensive tasks, requiring human intervention to ensure that grids are properly aligned and that artefacts are flagged and properly excluded from subsequent analysis. Adoption of standard input/output formats, automation of feature identification and software identification of common artefacts are important goals for the next generation of array analysis software. In addition, routine quality assessment and the assignment of robust confidence statistics on gene expression data are critical. This quality assurance information should be transmitted with the primary data through subsequent analyses and database submission, as is done in X-ray crystallography and DNA sequence assembly[10].

Data integration. Genomic information resources can be highly synergistic, and public databases and tools such as GenBank, Entrez and BLAST provide biologists with integrated and linked information. Clearly a GenBank-like public database of gene expression measurements, integrated with MEDLINE, Entrez and other data and tools, would be a useful resource for

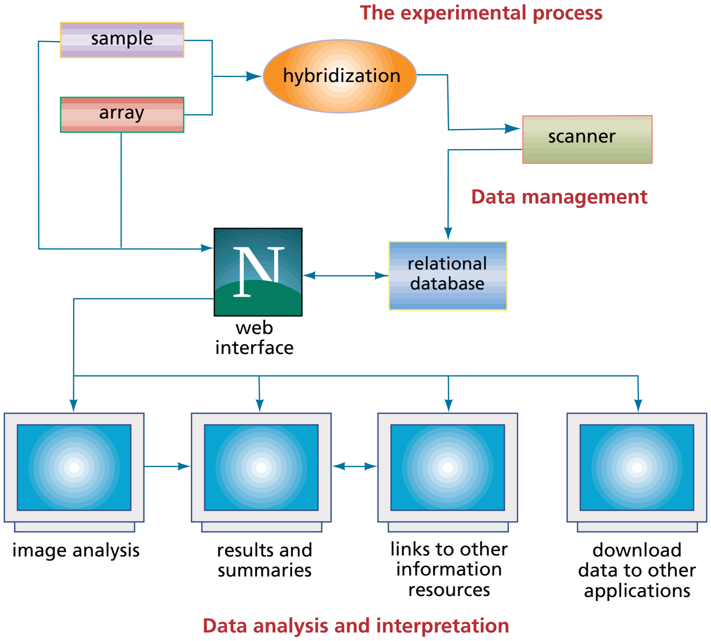

Fig. 1 Overview of the information system for large-scale gene expression experiments. It has previously been suggested that the most powerful and flexible design is that of a database-backed web site[6], although a number of groups are also developing self-contained or turnkey systems that run on dedicated workstations. Specialized software that addresses only a subset of the informatics requirements is also becoming available (Table 1). Laboratory information management systems (LIMS) must track both material and information flow throughout the experimental process as well as subsequent data processing steps. Critical information includes details on a target's physical and biological characteristics, that is, representing the hypothesis that an experiment was designed to test in the first place. Such information should be entered at the outset of an experiment, so that any results obtained may be easily linked back to starting hypothesis and materials. At the analysis and interpretation stage, standard analyses (for example, clustering) should be provided, along with results summaries that ideally would incorporate relevant information from external resources in an automatic or semi-automatic fashion. Lastly, because not all analytic needs and methods can be anticipated, it is important that the data can be easily downloaded from the database for importing into other applications such as existing statistical analysis packages or new programs and algorithms.

the biological community. Although existing database technologies are capable of managing such a resource, there are serious obstacles to establishing such a system. Fundamentally, what would be the form of expression data that such a database would store? The correlates of GenBank sequences or 3D atomic coordinates in the Protein Data Bank are difficult to specify for expression studies, as absolute values are hard to define. Methods for accurate, reliable normalization among multiple experiments and technologies are not available, and all gene expression methods suffer from both intra- and inter-experimental variability, making direct comparisons of raw intensity data between experiments prone to significant error. One goal of a public gene expression database would be to store data from diverse gene expression analysis technologies in a standardized, inter-operable form. Expression measurements in transcript copies per cell, or percentage of the total transcript pool, would be meaningful, useful and inter-operable among various technologies. Unfortunately, current hybridization technologies cannot measure these numbers; instead, they accurately measure the relative abundance of transcripts between two samples. Thus, conversion of typical gene expression data into a universal format requires that certain assumptions be made. If we assume that transcripts of one or more (ubiquitously expressed) 'housekeeping genes' have relatively stable, steady-state numbers, and if we assume that there is a linear relationship between average fluorescence intensities and transcript levels, the problem reduces quite nicely. Of course, such assumptions distort the primary data and may be unacceptable standards for a public gene expression database. Multicolour experiments, wherein fluorophore-labelled cDNA samples from each state being compared compete for binding to tethered hybridization probes, can generate highly accurate relative expression levels, allowing experiments with a common 'baseline' sample to be readily normalized. However, the use of standardized controls among laboratories employing a variety of gene expression technologies is an impractical and unlikely solution. Normalization among gene expression studies performed in multiple laboratories using varied technologies remains a difficult and pressing challenge to the establishment of a central, public domain resource of gene expression information.

Data mining

With the introduction of sophisticated laboratory instrumentation, robotics and large, complex data sets, biomedical research is increasingly becoming a cross-disciplinary endeavour requiring the collaboration of biologists, engineers, software and database designers, physicists and mathematicians. Techniques used in other fields can be extremely valuable if we can learn their proper applicability to biological problems. In this section, we describe data storage and analysis methods and requirements for gene expression studies, examine some approaches to 'data mining' drawn from other fields and consider how they might be applied for hypothesis testing and knowledge discovery.

Table 1 • Resources

Image Analysis		
BioDiscovery	http://www.biodiscovery.com/software.html	BioDiscovery's ImaGene Image Analysis Software
ScanAlyze	http://bronzino.stanford.edu/ScanAlyze	Brown Lab's Image Analysis software
Microarray Data Warehousing & Analysis		
Affymetrix	http://www.affymetrix.com/products/lims/lims.html	GeneChip LIMS data warehouse
Brown Lab, Stanford University	http://cmgm.stanford.edu/pbrown/explore/	Searchable database of published yeast microarray data
MicroArray Project, NIH	http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html	Database schema and software tools for analysis of high-throughput gene expression data
Rosetta Inpharmatics	http://www.rosetta.org/	Resolver data warehouse & analysis software
Silicon Genetics	http://www.sigenetics.com/GeneSpring/Overview.htm	GeneSpring data warehouse & analysis software

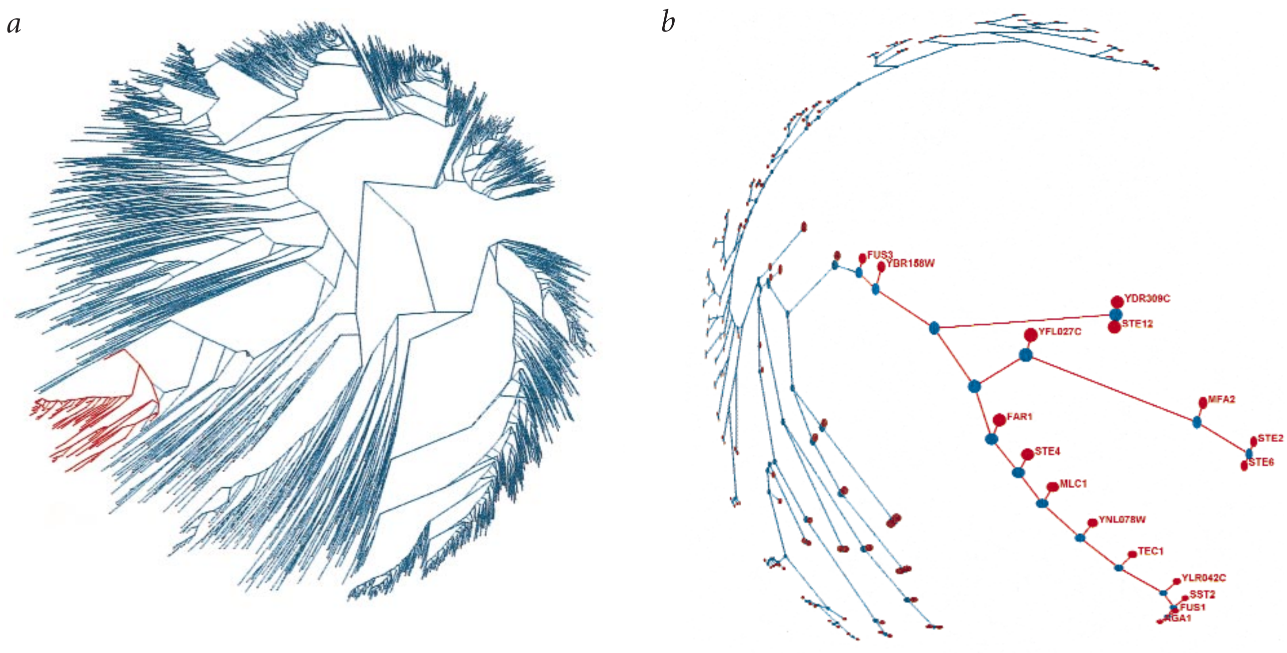

Fig. 2 Clustering high-throughput gene expression data can shed new light on biological pathways and processes. ***a***, clustering tree representing results from 3800 genes from the budding yeast *S. cerevisiae* which showed one or more significant gene expression changes across 365 experiments involving various genetic perturbations, drug treatments, and growth conditions. ***b***, *A* portion of this tree, highlighted in red, has been expanded. This branch contains a number of genes invoved in the yeast mating pathway, inluding yeast 'sterile' (STE) genes idenified in screens for mutants unable to undergo mating (*STE2, STE4, STE6, STE12*), the yeast a-factor mating pheromone precursor (*MFA2*), a gene involved in desensitization to mating pheromone (*SST2*), and other genes known to play a role in the yeast mating process (*FAR1*, *FUS1*, *FUS3*). Function-unknown open reading frames identified by genomic sequencing efforts are idenified which are co-regulated with these genes (YFL027C, YDR309C, YNL078W) and may play a role in the mating pathway. As with genes, microarray experiments can also be clustered to expose conditions, drug treatments, and genetic perturbations that yield similar expression profiles.

Data warehousing. Of central importance to the optimal utilitization of gene expression data is the development of some type of unified infrastructure for collecting, storing, retrieving and querying data, regardless of the technology used to generate it. The most significant contribution of gene expression arrays to our understanding of biological pathways and processes will derive not from the analysis of single experiments, but from libraries of experiments. Just as GenBank serves as an ever-improving classification space for each new gene and genome sequenced, gene expression analyses will benefit immensely from comparison and classification. The existence of such a resource for gene expression information would serve as a catalyst for the development of new and powerful tools that take advantage of large ensembles of these data.

The first step is to construct a database or 'data warehouse'. To cite a description of requirements drawn from other fields, "Data in the warehouse have already been cleaned and verified. Data from multiple sources have been integrated. A single data model ensures that similarly named fields have the same meaning throughout the database"[11]. Although these goals have largely been achieved for sequence, structure and bibliographic data in Entrez and its underlying ASN.1 data model[12], much work needs to be done to achieve similar results for gene expression data.

Errors and confidence levels of individual measurements are, in general, poorly understood. Most laboratories still base confidence levels on the magnitude of the ratios—typically deeming measurements with at least a two- or threefold deviation from a given intensity threshold 'significant'. This is a far-from-ideal standard, however, because high-intensity spots and/or those with highly reproducible ratios across multiple experiments are much more reliable than dim spots or those that display significant scatter across experimental repeats. As the quality of gene expression data can vary widely among experiments, reliable statistics for each expression measurement will be a necessity for a public repository of gene expression data.

In addition to variation among individual measurements in a single experiment, there are also variations or artefacts that can arise from the starting materials used for sample (or target) preparation (see page 38 of this issue (ref. 14)) that may make it difficult to compare even repetitions of the same experiment.

To stimulate discussion, we would like to propose a 'straw man standard' for annotated and normalized gene expression data in a public database. We suggest that the essential information could be represented in five categories:

Contact information: identifies the laboratory or investigator who submitted the data.

Hybridization targets: for each 'spot' on an array, there should be a public database identifier[15] for the DNA sequence present. For oligonucleotide arrays, the oligonucleotide sequences should identify a range in a reference sequence in a public database (for example, GenBank). For cDNA arrays, an additional clone identifier (such as an IMAGE clone_id) would be given. (Note that taxonomic information on the species from which the DNA target is derived could be accessed from the GenBank records.)

Target(s): (i) details of the cell types and/or tissues of origin using a controlled vocabulary. For mammalian tissues, histologic and histopathologic terms routinely used in diagnostic pathology could be employed. Standard terms for developmental stages in embryogenesis and organogenesis could also be used, depending on the experiment. (ii) Taxonomic names of the species of origin of the target (for example, *Saccharomyces cerevisiae*, *Homo sapiens*) should be provided. For some organisms such as rodents and microorganisms, strain information would also be identified. (iii) Information on the biological 'states' examined in the experiment, for example, drug-treated or untreated is crucial. Presumably, 'tumour versus normal' or separate developmental stages would be captured as in (i). (iv) The genetic background or genotype of the cell or organism would also be important in certain cases; yeast gene deletions and transgenic mice are obvious examples. Lastly, due to potential problems of reproducibility on the basis of variations in tissue handling, detailed information on the protocols should be provided.

mRNA transcript quantitation: for hybridization-based methods, level of induction or repression could be represented by increased or decreased intensity levels compared with an internal, universal control, perhaps on the basis of a set of empirically defined 'housekeeping' genes. The equivalent of fluorescence or radiation intensity measurements for the transcript tag-counting methods would be the number of observations of each gene tag. It does not seem any more practical to provide raw data for gene expression measurements than to supply the chromatograms or 'traces' underlying DNA sequence data (however, even these latter data currently lack a quality measure or confidence value for each base).

Statistical significance: some type of value expressing the confidence of the expression level changes described in the previous category is required. Ideally, it will be economically feasible to repeat an experiment a sufficient number of times so that the variance associated with each transcript level can be given.

The information in the first two categories is relatively straightforward to obtain and attach to expression records. The third category is more difficult because the information is more complex and there would need to be agreement on certain standards. However, none of these data attributes are unique to expression studies, and most problems have been successfully dealt with by the public sequence databases[16] or by more specialized resources[17,18]. The most challenging aspects of presenting gene expression data are represented by the last two categories.

Data exploration. The greatest intellectual challenge in using these new technologies is devising ways in which to extract the full meaning and implications of the data stored in large gene expression libraries. As expression databases grow, more sophisticated and user-friendly methods of analysis will be required. 'Data mining' has been defined as "the exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules"[11]. In biology, 'mining' of sequence databases has, of course, been going on for two decades and has reached a powerful new level in comparative genomics applications[19,20]. Indeed, there are some analogies between sequence similarity search technology and pattern recognition in gene expression arrays. The ultimate goal is to convert data into information and then information into knowledge.

There are two general approaches to data mining in large-scale expression analysis—hypothesis testing and knowledge discovery. Hypothesis testing is a 'top down' approach in which induction or perturbation of a biological process will lead to predicted results, although often also revealing new phenomena. Knowledge discovery by exploratory data analysis is a 'bottom up' approach in which the data are allowed to 'speak for themselves' after a statistical or visualization procedure is performed (see page 33 of this issue (ref. 21)).

Integration with other databases. In both cases, successful interpretation will rely on integrating experimental data with external information resources, such as those encompassed by NCBI's Entrez system[22,23]. Genes on arrays have been linked to the Entrez 'information space'[6], but this implementation is still very much a one-gene-at-a-time approach that needs to be streamlined by preprocessing the results of an experiment or series of experiments to produce an 'executive summary'. The idea is to employ a software agent to explore different *Entrez* nodes and to select the 'most important' records by user-defined or default criteria and then to summarize these results in a condensed overview of the findings relevant to genes that are upregulated or downregulated (or both). For example, one may only want to retrieve from MEDLINE review articles that were published more recently than a certain date, or only articles that include the term 'drug metabolism' in their titles, abstracts or MeSH headings. Or perhaps one would also want only those records for which 3D structures of the gene products were known or cases in which the genes have been mapped. Implementations of such strategies are clearly within the capabilities of existing data models and software tool kits[12] and are under development. Even more desirable would be a program that would be capable of suggesting possible explanations or hypotheses implied by the ensemble of information assembled by such a process.

Tools for exploring gene expression databases are in their infancy. In some cases, it is possible to understand what expression data are telling us about changing states of the organism by correlating transcriptional activity with known processes and pathways[24–28]. It is important to remember, however, that for even the best characterized organisms, functional information is usually incomplete and exists for only a fraction of the genes, and even less is known about the manner in which expression of these genes is regulated. Furthermore, electronic databases of pathway information[29] are currently limited in scope, computability, or both. A major focus of infrastructure development to support large-scale gene expression studies will be in the area of electronic biological pathway databases and resources.

Statistical analysis. Statistical methods can be applied to detect and extract internal structure in the data. It is a fundamental assumption of many gene expression studies that knowledge of where and when a gene is expressed carries important information about what the gene does; therefore, an obvious first step is to organize genes on the basis of similarities in their expression profiles. The idea of clustering genes on the basis of their expression patterns is well established (for example, the identification of gene expression at various points in the cell cycle) and has been applied to expression studies[30–32]. However, only recently have data become available to test the utility of this approach on a genomic scale (ref. 33; Fig. 2).

Although cluster analysis[34] has been the most widely used statistical technique applied to large-scale gene expression data, it is only one of several techniques that have been applied to data mining[11]. Others include affinity grouping or market basket analysis, memory-based reasoning, link analysis, decision trees and rule induction, self-organizing maps and other types of neural networks and genetic algorithms. Undoubtedly these other techniques will prove useful in gene expression analysis, particularly once standards in data exchange and confidence statistics are adopted.

Visualization. Another important component of genome-wide expression data exploration is the development of powerful data visualization methods and tools. Approaches have been developed that present clustering results in simple graphical displays to produce snapshots or overviews of large expression data sets that 'image' a transcriptional response without distorting the primary data[33]. Such visualization techniques, combined with integrated links to annotated sequence databases, provide very valuable tools that allow biologists to examine large expression data sets and develop new insights into and models of genome-wide transcriptional regulation. In other fields, such procedures are sometimes referred to as on-line analytic processing (OLAP), which is a data presentation methodology enabling efficient, manual knowledge discovery that depends on human intelligence and pattern recognition[11].

Perspectives

The overall transcriptional response of a cell to a given growth condition, drug treatment, or genetic perturbation is not unlike the sequence of a gene in some respects. Expression profiles can be 'aligned' with one another to identify similar cellular responses. As with global versus local sequence alignments, the entirety of an expression profile will not be important for some analyses and, in some cases, could even muddle or confuse a result. It will be important for us to begin to delve into the 'subsignatures' on array profiles that are the correlates of domains or motifs in protein sequence analysis. The analogue to multiple sequence alignment is statistical cluster analysis, which is being applied to gene expression data to an increasing degree. Cluster analysis of results from multiple experiments to identify genes that behave similarly can be used to identify co-regulated genes, and thereby reveal regulatory elements, transcription factors and even previously undiscovered players in a cellular pathway or process. On another level, clustering allows the grouping of growth conditions, mutations and drugs that elicit similar transcriptional responses in different experiments. Similar to sequence analysis a decade ago, the analysis of high-throughput gene expression data is in an early stage of development. With dominant technologies emerging for gene expression array construction and scanning, data analysis and integration techniques and tools will no doubt be the primary research focus in the future.

Despite the rather early stage of development of large-scale gene expression monitoring systems and methods, this new technology has already proven exceptionally useful in expanding our knowledge of even well-understood aspects of cellular biology. Studies on the mitotic cell cycle[24,35] and sporulation[25] in yeast and the serum response in human cells[36] reveal rich and coherent information contained in the expression patterns of genes. The use of whole-genome transcript analysis also has great potential for identifying *cis*-regulatory elements controlling expression networks[27], and for drug target validation and the identification of secondary drug target effects[37]. We look forward with excitement and confidence to the future of genome-wide expression experiments.

Acknowledgements

We thank A. Bondarenko, H. Dai, Y. He & R. Stoughton for their significant contributions to this work; and G. Church for valuable suggestions on the manuscript.

1. Chuang, S.E., Daniels, D.L. & Blattner, F.R. Global regulation of gene expression in *Escherichia coli*. *J. Bacteriol.* **175**, 2026–2036 (1993).
2. Gress, T.M., Hoheisel, J.D., Lennon, G.G., Zehetner, G. & Lehrach, H. Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm. Genome.* **3**, 609–619 (1992).
3. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
4. Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
5. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
6. Ermolaeva, O. *et al.* Data management and analysis for gene expression arrays. *Nature Genet.* **20**, 19–23 (1998).
7. Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J. Expression profiling using cDNA microarrays. *Nature Genet.* **21**, 10–14 (1999).
8. Cheung, V.G. *et al.* Making and reading microarrays. *Nature Genet.* **21**, 15–19 (1999).
9. Chen, Y., Dougherty, E.R. & Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomed. Optics* **2**, 364–374 (1997).
10. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
11. Berry, M.J.A. & Linoff, G. *Data Mining Techniques for Marketing, Sales and Customer Support* (John Wiley & Sons, New York, 1997).
12. Baxevanis, A. & Ouellette, B.F.F. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (John Wiley & Sons, New York, 1998).
13. Brownstein, M.J., Trent, J.M. & Boguski, M.S. Functional genomics. *Trends Guide to Bioinformatics (*eds Patterson, M. & Handel, M.) 27–29 (Elsevier, Oxford, 1998).
14. Cole, K.A., Krizman, D.B. & Emmert-Buck, M.R. The genetics of cancer—a 3D model. *Nature Genet.* **21**, 38–41 (1999).
15. Ouellette, B.F. & Boguski, M.S. Database divisions and homology search files: a guide for the perplexed. *Genome Res.* **7**, 952–955 (1997).
16. Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J. & Ouellette, B.F. GenBank. *Nucleic Acids Res.* **26**, 1–7 (1998).
17. Ringwald, M. *et al.* A database for mouse development. *Science* **265**, 2033–2034 (1994).
18. Ringwald, M. *et al.* The mouse gene expression database GXD. *Sem. Cell Dev. Biol.* **8**, 489–497 (1997).
19. Makalowski, W. & Boguski, M.S. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412 (1998).
20. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
21. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21**, 33–37 (1999).
22. McEntyre, J. Linking up with Entrez. *Trends Genet.* **14**, 39–40 (1998).
23. Schuler, G.D., Epstein, J.A., Ohkawa, H. & Kans, J.A. Entrez: molecular biology database and retrieval system. *Methods Enzymol.* **266**, 141–162 (1996).
24. Cho, R.J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* **2**, 65–73 (1998).
25. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
26. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
27. Roth, F.P., Hughes, J.D., Estep, P.W. & Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16**, 939–945 (1998).
28. Velculescu, V.E. *et al.* Characterization of the yeast transcriptome. *Cell* **88**, 243–251 (1997).
29. Kanehisa, M. Databases of biological information. *Trends Guide to Bioinformatics (*eds Patterson, M. & Handel, M.) 24–26 (Elsevier, Oxford, 1998).
30. Carr, D.B., Somogyi, R. & Michaels, G. Templates for looking at gene expression clustering. *Statistical Computing and Graphics Newsletter* **8**, 20–29 (1997).
31. Michaels, G.S. *et al.* Cluster analysis and data visualization of large-scale gene expression data. *Pac. Symp. Biocomput.* 42–53 (1998).
32. Wen, X. *et al.* Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA* **95**, 334–339 (1998).
33. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* (in press).
34. Kaufman, L. *Finding Groups in Data: An Introduction to Cluster Analysis* (John Wiley & Sons, New York, 1990).
35. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* (in press).
36. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* (in press).
37. Marton, M.J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4**, 1293–1301 (1998).