Data management and analysis for gene expression arrays

Olga Ermolaeva[1,2], Mohit Rastogi[3], Kim D. Pruitt[2], Gregory D. Schuler[2], Michael L. Bittner[1], Yidong Chen[1], Richard Simon[4], Paul Meltzer[1], Jeffrey M. Trent[1] & Mark S. Boguski[2,3]

Microarray technology makes it possible to simultaneously study the expression of thousands of genes during a single experiment. We have developed an information system, ArrayDB, to manage and analyse large-scale expression data. The underlying relational database was designed to allow flexibility in the nature and structure of data input and also in the generation of standard or customized reports through a web-browser interface. ArrayDB provides varied options for data retrieval and analysis tools that should facilitate the interpretation of complex hybridization results. A sampling of ArrayDB storage, retrieval and analysis capabilities is available (www.nhgri.nih.gov/DIR/LCG/15K/HTML/), along with information on a set of approximately 15,000 genes used to fabricate several widely used microarrays. Information stored in ArrayDB is used to provide integrated gene expression reports by linking array target sequences with NCBI's Entrez retrieval system, UniGene and KEGG pathway views. The integration of external information resources is essential in interpreting intrinsic patterns and relationships in large-scale gene expression data.

Our modern concept of gene expression dates to 1961, when messenger RNA was discovered, the genetic code deciphered and the theory of genetic regulation of protein synthesis described[1–3]. The first attempts at global surveys of gene expression were undertaken in the mid-1970s. Kinetic studies of the hybridization of mRNA pools with radioactively labelled cDNA produced the general concepts of varying mRNA abundance classes that are related to the functional class (structural, catalytic and so on) of the translated proteins[4,5]. These experiments also provided insight into: (i) the number of members of these classes; (ii) the presence of a large number of ubiquitously expressed ('housekeeping') genes thought to be necessary for the structural and functional integrity of all cell types; and (iii) the existence of significant numbers of genes that are apparently cell-type–specific. This period coincided with the establishment and popularization of the phrase 'gene expression' through its usage in the titles of a series of influential books[6–9]. Interest in gene expression increased steadily during the 1980s, as shown by the fact that the frequency of usage of the phrase increased more than 10-fold in the titles of publications over this decade (unpub. obs.).

In the 1990s, a new era of gene expression studies has unfolded as a result of data sufficiency (that is, complete genomes or comprehensive cDNA surveys) and technological advances[10–12]. As a consequence of large-scale DNA sequencing activities, there are now more DNA sequences in GenBank than there are related publications in the literature (Fig. 1). Thus, we have reached a turning point in biomedical research: in the past we have had many publications about a relatively small number of genes, whereas now, and in the future, single publications will begin to encompass aspects of thousands of genes[12–17]. Large-scale study of gene expression is a hallmark of the transition from 'structural' to 'functional' genomics[18], where knowing the complete sequence of a genome is only the first step in understanding how it works.

There are several new technologies for studying the simultaneous expression of large numbers of genes. These technologies may be generally divided into serial and parallel methods. The serial methods involve direct, large-scale sequencing of cDNA (for review, see ref. 19); the parallel approaches are based on hybridization to cDNA immobilized on glass (termed 'microarrays'; ref. 11) or to synthetic oligonucleotides immobilized on silica wafers or 'chips' (termed 'probe arrays'; refs 10,20). In both parallel methods, hybridized probes are detected using incorporated fluorescent nucleotide analogs. These methods are the conceptual descendents of filter-immobilized targets detected by radioactive probes[21,22], and filter-based technology is undergoing a renaissance as a low-cost alternative to the newer methods. Regardless, arrays of hybridization targets, generated at high density in small areas (for example, 10,000 cDNAs on a 2×2-cm filter or glass slide) are now commonly referred to as microar-

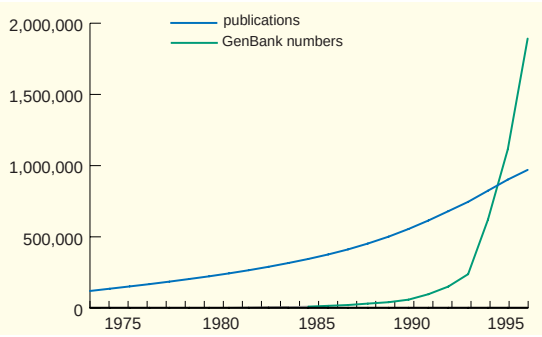

Fig. 1 Cumulative growth of molecular biology and genetics literature (blue) compared with DNA sequences (green). Articles in the 'G5' (molecular biology and genetics) subset of MEDLINE are plotted alongside DNA sequence records in GenBank over the same time period. The former data was obtained with the help of R.M. Woodsmall of NCBI and the latter data is available (ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt). No attempt has been made to eliminate data redundancy among either the DNA sequence records or information contained in the literature.

[1]Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. [3]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [4]Biometric Research Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. Correspondence should be addressed to M.S.B. (e-mail: boguski@ncbi.nlm.nih.gov).

Box 1 • The 10K/15K gene sets

The initial resources required to design and fabricate gene expression microarrays include cDNA sequence data, cDNA clones, or both. Identification of genes and clones of interest is problematic due to the quantity and redundancy of sequence data available. Some problems associated with the large-scale application of genome resources have been faced before in the context of building a transcript map of the human genome[24], and databases consisting of non-redundant collections of human and mouse genes and ESTs have been developed[25]. The UniGene collection of human sequences (http://www.ncbi.nlm.nih.gov/UniGene/) currently represents more than 45,000 genes and it is possible to fabricate arrays containing this entire collection. Initial work in our laboratories focused on a smaller, but still significant subset of approximately 10,000–15,000 transcribed human sequences referred to as the 10K and 15K sets, originally conceived by P. Brown, J.M.T. and M.L.B., developed by G.S. and arrayed by J. Hudson. Detailed information on the composition of these sets is available (http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/). Briefly, the sets were designed to include a selection of human genes of known function, ESTs on the human transcript map[24], ESTs with significant similarities to genes in other organisms and some hand-picked genes of specific research interest.

rays. Detailed discussion of these technologies is beyond the scope of this article (see http://www.ncbi.nlm.nih.gov/ncicgap/-expression_tech_info.html and http://www.nhgri.nih.gov/), but we note that bioinformatics needs are similar and equally essential for all methods.

Although a great deal of effort has gone into the development of the enabling technologies, relatively little attention has been paid to the computational biology underlying data analysis and interpretation. We describe here some general aspects of gene expression informatics as well as our specific implementation of an integrated data management and analysis system (ArrayDB), designed as a database-backed web site[23]. Informatics plays an important role at every step in the process, from the design of arrays, through laboratory information management, to the processing and interpretation of experimental results. We also discuss the role of the public database in this new era of biomedical research.

Design considerations

Array-based experiments aim to simultaneously catalogue the expression behaviour of thousands of genes in a single experiment. It is also expected that comparisons will be carried out across tissues, developmental and pathological states, or as temporal responses following a defined alteration to cells or their environment. Such experiments require the ability to manage large quantities of data both before and after the experiment. The design and construction of arrays that will detect gene expression requires direct access to all sequences, annotations and physical DNA resources for the genes of an organism (Box 1).

Following hybridization and readout of relative expression levels observed in various sites on an array, the data collected must be stored and preserved in a way to make it readily available for image processing[26] and statistical and biological analysis. The latter includes identifying the changing and unchanging levels of expression and correlating these changes to identify sets of genes with similar profiles. Easy access to existing biological knowledge of gene function and interaction is necessary to fully interpret the biological implications of the observed patterns. An information system must also be flexible enough to accommodate new statistical data mining tools as they become available.

Laboratory information management systems (LIMS)

The successful use of large-scale functional genomics technologies depends on robust and efficient systems for tracking and managing material and information flow. An overview of the types of practical problems addressed by our LIMS is shown (Fig. 2). The individual components and detailed design of LIMS is connected with specific laboratory environments, particularly for those technologies still under development, but some general principles have guided our work. These include the use of an industry standard relational database management system combined with platform-independent web browser interfaces for data entry and retrieval[23].

The microarray LIMS, ArrayDB, was developed to store, retrieve and analyse microarray experiment information. The ArrayDB system integrates the multiple processes involved in microarray expression experiments, including data management, user interface, robotic printing, array scanning and image processing[26]. Data stored in the ArrayDB system includes information about the experimental resources, experimental parameters and conditions, and raw and processed hybridization results.

The relational database underlying the ArrayDB system stores extensive information pertaining to each clone in the microarray, including a brief gene description, GenBank accession number, IMAGE clone identifier (http://www-bio.llnl.gov/bbrp/image/image.html), metabolic pathway identifier and internal clone identifier (Fig. 3). ArrayDB also stores information related to microarray fabrication and experimental conditions and outcomes. This information includes data pertaining to printing the arrays (such as the printer robot parameters), environmental conditions (temperature, humidity, tip wash conditions) and the GIPO ('gene in plate order') data that relates the clones and their relative order on the array. In addition, the database stores extensive information about the probes and experimental conditions, including the investigator's name, the purpose of the experiment and textual descriptions of the conditions, tissues or cell types in which the 'red' (Cy5-labelled) and 'green' (Cy3-labelled) hybridization probes originated. Information on hybridization results includes the scanned ('raw') images, hybridization intensity data, intensity ratios and background values.

The design of ArrayDB allows for flexibility in the exact nature of the data stored. This design strategy permits data input from different sources. Most clone information stored in the ArrayDB is extracted from UniGene (for example, sequence definition and accession number). However, the design accommodates addition of newly isolated clones for which accession numbers or meaningful names are not yet available. Many data input and processing tasks are automated. Software automatically scans a directory for new intensity data that are uploaded into the database without requiring an operator's assistance. Additional automated

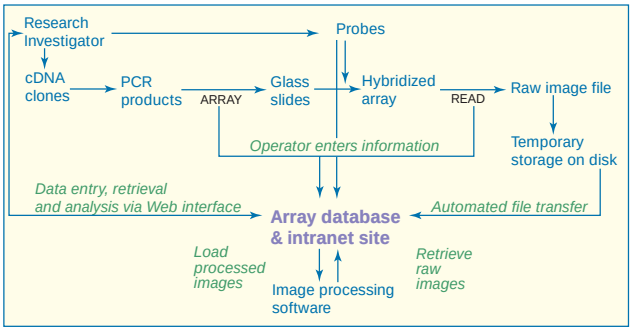

Fig. 2 Schematic overview of the ArrayDB information management system. The basic information in the database consists of arrays, 'probes' and images. Arrays of specific cDNA clone inserts (and accompanying annotation) are as described in the text, Box 1 and the legend to Figure 3. The section labelled 'probes' signifies details of a particular experiment as described in the text. Details regarding image processing are provided[26]. An *ad hoc* raw image format is used for processing, but this is converted to standard formats (JPEG, GIF) for subsequent analysis and display. A complete relational schema of the database is available on request.

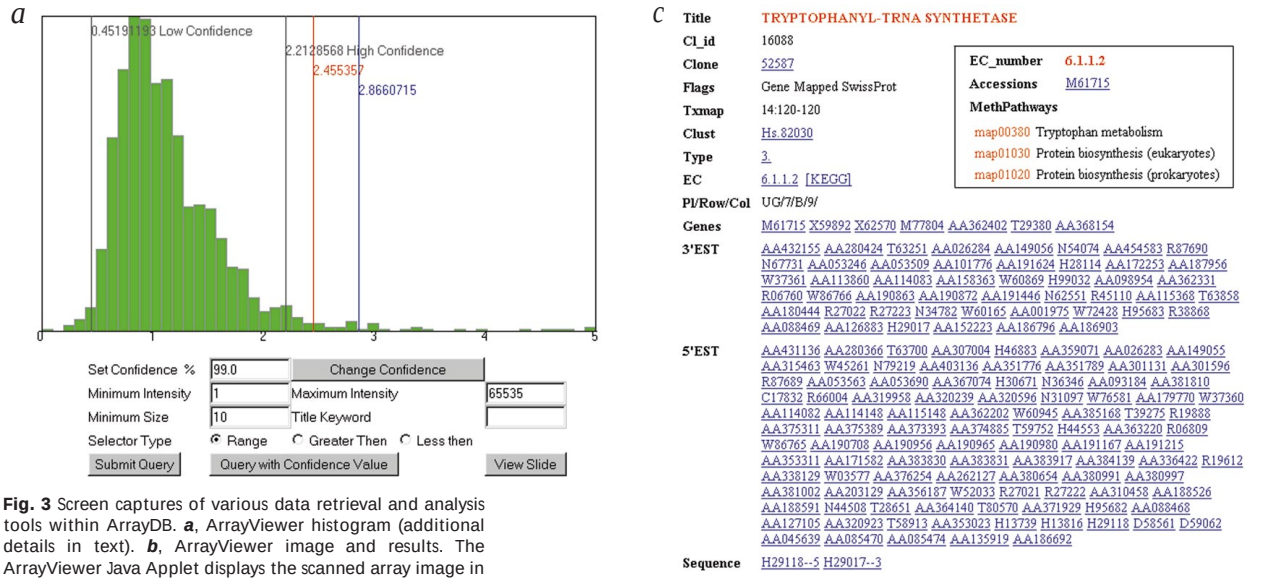

Fig. 3 Screen captures of various data retrieval and analysis tools within ArrayDB. ***a***, ArrayViewer histogram (additional details in text). ***b***, ArrayViewer image and results. The ArrayViewer Java Applet displays the scanned array image in the top window. Boxes and the ranking number are overlaid on the image for clones that have satisfied the query criteria; clones are ranked according to ascending ratio value. The boxed clones, and related quantitative data, are listed under the image. Quantitative data presented in the lower window include: the ranking number, IMAGE clone ID the ratio, probe A intensity, probe B intensity, probe A pixel size, probe B pixel size and the clone title. ***c***, ArrayViewer cluster report. The example shows a report for Tryptophanyl-tRNA synthetase. 'Cl_id' is an internal database identifier. The 'Clone' field contains the IMAGE clone identifier and is hyperlinked to the dbEST records containing the sequences of this clone. 'Flags' summarizes the criteria by which this sequence was included in the 10K/15K sets. 'Txmap' refers to the location of an STS derived from this sequence on the human transcript map[24] and 'Clust' indicates the UniGene cluster containing this sequence (http://www.ncbi.nlm.nih.gov/UniGene). 'EC' contains the enzyme commission nomenclature number for this enzyme and 'KEGG' links it to the biochemical pathway reports available through the KEGG web site (http://www.genome.ad.jp/kegg/)[27]. 'Pl/Row/Col' refers to the microtitre plate and well from which the original clone was obtained. The 'Genes' field contains GenBank accession numbers for annotated (non-EST) versions of the sequence and the '3' EST' and '5' EST' fields contain GenBank accession numbers for all ESTs corresponding to the cDNA sequence. Lastly, the 'Sequence' field contains only those accession numbers referring to those EST sequences derived from the actual IMAGE clone insert selected for inclusion in the array.

processes were developed to facilitate integration of intensity data with clones data; for example, ArrayDB maintains the association between a spot on an image and all the data related to the clone located at that position on the microarray.

The web-based user interface to the ArrayDB system allows convenient retrieval of distinct types of information, ranging from clone data to intensity data to analysis results. ArrayDB supports database queries by different fields, such as clone ID, title, experiment number, sequence accession number, or microtiter plate number, with a resulting report of the relevant clone(s). Additional information about each clone is available through hypertext links to other databases such as dbEST, GenBank or UniGene. Furthermore, metabolic pathway information is also available through links to the Kyoto encyclopedia of genes and genomes (KEGG) web site[27].

The inconsistency inherent in gene nomenclature makes it more efficient and accurate to search for a gene of interest by doing a sequence similarity search. ArrayDB supports BLASTN searches against the 10K/15K set so that anyone can quickly determine if a gene of interest is included on our arrays. Matches against individual sequences are linked to a 'cluster report' (Fig. 3*c*), and from there to further annotation in external databases *via* hypertext links as described above.

Data analysis

The ultimate goal of ArrayDB is to identify patterns and relationships among intensity ratios both in individual and across multiple experiments. The ArrayViewer tool supports retrieval and analysis of single experiments; MultiExperiment viewer supports analysis of data from multiple experiments. In addition, the option to download intensity data, images and some analysis results to a local disk adds flexibility to the end-users analysis options: once downloaded, intensity data can be imported into other software packages for analysis.

ArrayViewer facilitates identification of statistically significant hybridization results in single experiments. The data set for a single experiment includes intensity ratio data for two fluorescent hybridization probes. However, the inherent flexibility in the ArrayDB design strategy is compatible with results derived from single intensity (for example, radioactive probe) data. In the case of radioactive probes, a single image consists of the intensity data from two separate hybridization experiments using two different

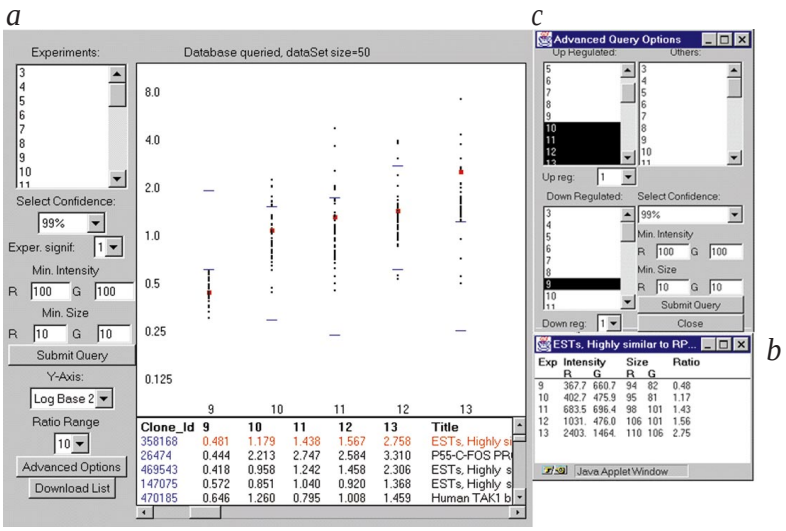

Fig. 4 MultiExperiment viewer window. ***a***, The main panel of the MultiExperiment viewer is divided into three sections. The left side is composed of the control panel where the query criteria are selected. One also selects the experiments to analyse and other filters such as keywords, minimum intensities and minimum pixel sizes. The data returned from a query can be downloaded in a tab delimited text file by selecting download list in this panel. The control panel can also be used to alter the y-axis format and scale of the data represented in the window on the right side. This window is a dot plot of the experimental data returned from the query. Selecting particular 'dots' with a mouse highlights the ratio data for that clone across all selected experiments in both the dot plot and the quantitative data in the lower right window. The lower right window displays the calibrated ratio of the returned genes (clones). Selecting the ranking number highlights that data in the dot-plot. The IMAGE consortium Clone Id is linked to the cluster reports (Fig. 3, legend). Selecting ratio and title will launch a new window (***b***) that displays the red and green intensities and sizes for that clone. By selecting advanced options in the control panel, a new window (***c***) is launched that allows greater flexibility and control in defining a query. Greater precision is achieved by allowing one to specify experiments where only up-regulated clones or only down-regulated clones are of interest.

probes. Ratios of the intensity values obtained with each probe, for each clone, are determined and stored in the database. (The mathematical basis for our image analysis approach is reported elsewhere[26].) The basic premise of ArrayViewer is that significant hybridization results can be determined from the ratio values. Therefore, ArrayViewer initially displays a histogram that is created on demand using the ratios stored in ArrayDB (Fig. 3*a*).

From the ArrayViewer histogram, there are three basic ways to query the data and return information on the nature and expression of specific genes. The first method uses a confidence algorithm[26]. Querying by confidence values will return a list of those genes with statistically significant ratio values that are less than a lower confidence limit and greater than an upper confidence limit. The default confidence value is 99%, but this can be changed and the lower and upper confidence limits re-calculated. The second method allows the user to select a range of ratios on the histogram and will return information on genes with expression ratios in this range. The last method is to simply view the image of the hybridization results and select spots in the array using a computer mouse or other pointing device. One can further refine the ArrayViewer query by adjusting optional filters for minimum intensity, maximum intensity, minimum size, or keyword.

Query results are provided in a new window that displays the array image and a list of clones with their associated intensity data (Fig. 3*b*). Additional information about each data point or clone can be obtained by clicking on the ranking number (red) or the clone Id number (blue), respectively. Selecting the ranking number opens a new window presenting a ×10 magnification of the hybridized target spot plus a reiteration of the hybridization result. Selecting the clone Id number opens a new window containing a cluster report for that clone (Fig. 3*c*). Lastly, the data in the results window can be downloaded to a tab delimited text file by clicking on 'Download List'.

To realize the full potential of microarray expression analysis, MultiExperiment viewer was developed. This web-based tool enables users to query the database across multiple experiments to identify clones that share some pattern of expression across those experiments. For example, one can use this tool to identify genes that are up-regulated or down-regulated across a series of experiments. In addition, the user can track the behaviour of a particular gene or genes of interest by specifying key words from gene descriptions in the 15K set. Analysis results are presented in both a graphical and tabular format. Also provided is a download option of the result table to facilitate storage of results for future reference and/or additional analysis.

The MultiExperiment viewer window (Fig. 4) provides a control panel for selecting the query criteria, an area to display a dot plot of the query results and a section where the table of quantitative information is displayed. To develop the query, one must first select the experiments from the list in the upper left corner; several filters are also provided which enables the user to 'fine-tune' the query. The MultiExperiment viewer then queries the database to identify clones exhibiting ratios that meet the query requirements, returns the ratio for each clone and draws a dot-plot of the results for each experiment selected. This provides a convenient method to identify clones with particularly high or low ratios in an experimental series, such as a time course. There are two ways to visualize the expression pattern shown by an individual clone across the selected set of experiments. The position of the clone is highlighted in the dot plot diagram (Fig. 4, red boxes) for each experiment by either clicking on a desired spot in the diagram or by clicking on the ranking number (left column) of a clone with interesting quantitative data. As previously described, additional information about each gene product is readily available in the clone's cluster report (Fig. 3*c*) *via* the hyperlinked clone ID column.

The comparison of data across multiple experiments requires a way of normalizing ratio results between experiments; to date,

Box 2 • Public access to expression array data

As large-scale gene expression data accumulates, public data access becomes a critical issue. What is the best forum for making the data accessible? Summaries and conclusions of individual experiments will, of course, be published in traditional peer-reviewed journals, but electronic access to full data sets is essential. There are three models for data publication: first, authors can make data available on their own web sites (for example, http://cmgm.stanford.edu/pbrown/explore[14]); second, journals that publish the results of these studies can provide the complete data sets as electronic supplements (this approach fulfills the traditional archival responsibility of the literature); and the third approach is to submit the data to a centralized public data repository such as GenBank. The primary disadvantages of the first two models are that data is widely dispersed and lacks uniform structure and retrieval modalities. In addition, the first case is complicated further by an uncertain life span for the data and the second case incurs new expenses for curating and maintaining this data that journals may not wish to bear. Clearly, the successful history of public sequence databases provides an attractive model for the most efficient management of and convenient access to large-scale expression data. However, it would be highly desirable to arrive at some type of data format standards that are independent of a particular expression technology.

this has only been possible by using a single reference state as the source of one of the hybridization probe mixtures for all of the experiments to be compared. For example, such an approach has been used in comparing points along a time course, and in comparing multiple samples of a particular type of tumour (unpublished observations). In diauxic shift experiments[14], the reference sample was cDNA prepared from yeast cells harvested at the first interval after inoculation. Although the use of such a reference comparator allows ratio comparisons within a series of experiments, there is clearly a need for a more broadly applicable reference standard to serve as a benchmark for all expression experiments. A number of microarray laboratories have given thought to formulating such a standard. An ideal standard would provide modest signals for every human gene, so that expression of any gene in the experimental probe could be assigned a reliable ratio value. The standard would also need to be readily and reproducibly generated and easily disseminated. Efforts to produce such a reference standard are underway.

Discussion

Given the great potential of large-scale expression analysis, and biologists' desire to exploit this new technology, we anticipate a deluge of data soon. The capacity to ask questions and perform analyses across hundreds, thousands, or tens of thousands of experiments should dramatically enhance our ability to identify 'fingerprints' of gene expression that exemplify particular diseases or other biological states. But first we will need to empirically define 'housekeeping' genes, identify reproducible artifacts and detect subtle patterns through the application of powerful statistical analysis techniques.

This potential cannot be fully realized without efficient data management and analysis systems. ArrayDB provides a first-generation, convenient, flexible and extendable microarray data management and analysis system. Planned future extensions to the ArrayDB include more sophisticated links between the database and external data sources and more powerful data mining capabilities. Currently, querying multiple databases such as NCBI's PubMed, GenBank, or dbEST databases can assemble a great deal of valuable information, but it can be a tedious and time-consuming process to repeatedly query each database for information on even a small number of genes. However, by fully exploiting the applications programming interfaces in the Entrez system, sophisticated 'executive summaries' can, in principle, be generated.

Although these types of reports can be generated by the thoughtful integration of external data resources, the larger problem is identifying intrinsic patterns and relationships in the data itself. In the world outside of biological databases, the term 'data mining' has been applied to this type of knowledge discovery[28]. Because of the complexity of the data, data mining tools are essential to fully exploit the power of microarray expression analysis. Data mining tools, similar to mathematical techniques that identify patterns in complex data sets, will enable identification of multiple expression profiles in complex biological processes. This will provide a means to identify genes that share an expression profile, genes that are expressed in succession, or genes showing opposing expression profiles. For instance, cluster analysis[29] of a time course experiment can identify different expression profiles exhibited by groups of genes. We are currently developing a data mining tool for the ArrayDB system to help address this need.

Methods

The microarray relational database management system (ArrayDB) is implemented in Sybase®; details of the schema are available on request. Briefly, the database was designed to store information on hybridization targets (cDNA clones) that may or may not have sequence information available and may or may not be publicly available resources. Particular microarrays consist of subsets of all potential targets in the database. Data stored for individual experiments include the composition of the array, specific combinations of probes, experimental conditions and hybridization results, including fluorescence intensities for raw and processed images. Further details of image processing are available[26].

Interactions between Sybase SQL server and HTTP servers (web browsers) are managed by web.seql 'middleware' (http://www.sybase.com/products/internet/websql/) combined with the Perl and Java programming languages. For example, our ArrayViewer and MultiExperiment viewer are interactive Java applets that use a custom object that is an extension of a publicly available JavaCGIBridge Class developed by G. Birznieks and S. Sol (http://www.gunther.web66.com/JavaCGI/).

Acknowledgements

We thank M. Eisen, P. Brown and J. Hudson for stimulating discussions. We also thank J. Hudson and Research Genetics for re-arraying the 10K/15K clone sets from their original IMAGE cDNA libraries.

1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
2. Nirenberg, M.W. & Matthaei, J.H. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl Acad. Sci. USA* **47**, 1588–1602 (1961).
3. Taylor, J.H. *Selected Papers on Molecular Genetics*. (Academic Press, New York, 1965).
4. Bishop, J.O. & Smith, G.P. The determination of RNA homogeneity by molecular hybridization. *Cell* **3**, 341–346 (1974).
5. Galau, G.A., Britten, R.J. & Davidson, E.H. A measurement of the sequence complexity of polysomal messenger RNA in sea urchin embryos. *Cell* **2**, 9–20 (1974).
6. Lewin, B. *The Molecular Basis of Gene Expression*. (Wiley-Interscience, London, 1970).
7. Lewin, B. *Gene Expression-1* (John Wiley, New York, 1974).
8. Lewin, B. *Gene Expression-2* (John Wiley, New York, 1974).
9. Lewin, B. *Gene Expression-3* (John Wiley, New York, 1977).
10. Fodor, S.P. *et al.* Multiplexed biochemical assays with biological chips. *Nature* **364**, 555–556 (1993).
11. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–70 (1995).
12. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
13. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* **14**, 457–460 (1996).
14. 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).
15. Wodicka, L., Dong, H., Mittmann, M., Ho, M.H. & Lockhart, D.J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* **15**, 1359–1367 (1997).
16. Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
17. de Saizieu, A. *et al.* Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nature Biotechnol.* **16**, 45–48 (1998).
18. Hieter, P. & Boguski, M. Functional genomics: it's all how you read it. *Science* **278**, 601–602 (1997).
19. Adams, M.D. Serial analysis of gene expression: ESTs get smaller. *Bioessays* **18**, 261–262 (1996).
20. Fodor, S.P. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773 (1991).
21. Lennon, G.G. & Lehrach, H. Hybridization analyses of arrayed cDNA libraries. *Trends Genet.* **7**, 314–317 (1991).
22. 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).
23. Greenspun, P. *Database Backed Web Sites: The Thinking Person's Guide to Web Publishing* (Ziff-Davis, Emeryville, California, 1997).
24. Schuler, G.D. *et al.* A gene map of the human genome. *Science* **274**, 540–546 (1996).
25. Boguski, M.S. & Schuler, G.D. ESTablishing a human transcript map. *Nature Genet.* **10**, 369–371 (1995).
26. Chen, Y., Dougherty, E.R. & Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics* **2**, 364–374 (1997).
27. Kanehisa, M. A database for post-genome analysis. *Trends Genet.* **13**, 375–376 (1997).
28. Berry, M.J.A. & Linoff, G. *Data Mining Techniques for Marketing, Sales, and Customer Support* (John Wiley, New York, 1997).
29. Kaufman, L. & Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis* (John Wiley, New York, 1990).