REVIEW

Gene Expression Databases and Data Mining

Pascale Anderle[1], Manuel Duval[2], Sorin Draghici[3], Alexander Kuklin[4], Timothy G. Littlejohn[5], Juan F. Medrano[6], David Vilanova[7], and Matthew Alan Roberts[7]

[1]ISREC, Lausanne, Switzerland; [2]Genomics and Bioinformatics Group, Research and Technology Department, Fresnes Laboratories, Pfizer Global Research Development, Fresnes, France; [3]Department of Computer Science, Wayne State University, Detroit, MI, USA; [4]Transgenomic, Wayne, PA, USA; [5]BioLateral, Enfield Sth, NSW Australia; [6]Department of Animal Science, University of California, Davis, CA, USA; [7]Nestlé Research Center, Lausanne, Switzerland

BioTechniques 34:S36-S44 (March 2003)

ABSTRACT

The DNA microarray technology has arguably caught the attention of the worldwide life science community and is now systematically supporting major discoveries in many fields of study. The majority of the initial technical challenges of conducting experiments are being resolved, only to be replaced with new informatics hurdles, including statistical analysis, data visualization, interpretation, and storage. Two systems of databases, one containing expression data and one containing annotation data are quickly becoming essential knowledge repositories of the research community. This present paper surveys several databases, which are considered "pillars" of research and important nodes in the network. This paper focuses on a generalized workflow scheme typical for microarray experiments using two examples related to cancer research. The workflow is used to reference appropriate databases and tools for each step in the process of array experimentation. Additionally, benefits and drawbacks of current array databases are addressed, and suggestions are made for their improvement.

INTRODUCTION

Microarray studies, through the application of genetic and molecular biology information, have allowed biologists to study global gene expression in cells and tissues, over different temporal and experimental conditions, to discover key players in metabolic pathways and to assign probable function to genes. Other genomic scale technology aimed at capturing gene expression information includes serial analysis of gene expression (SAGE) and expressed sequence tag (EST) library sequencing. The common trait among these technologies is their capability to capture comprehensive biological information, in which all endpoints are measured simultaneously. In general, these global approaches for studying gene transcription have proven to be highly versatile with applications developing in basic research, high-throughput expression profiling in drug discovery, clinical diagnostics, and many others (1–3). In turn, these advances have triggered a conceptual shift in the scientific study, from single object (e.g., a gene, a protein) to system studies, aimed at capturing the true complexity of biological systems through global analysis.

A generalized workflow for microarray experimentation is depicted in Figure 1. Although initially restricted because of cost, nearly every molecular biology laboratory in the world will have access in the near future to high-throughput functional-genomic technology and be involved in capturing data sets consisting of tens of thousands of gene expression data points per sample measured. In fact, it is likely that the amount of expression data will soon surpass the amount of sequence data, since an organism can be characterized by a single sequencing effort but be represented by multiple expression profiles corresponding to any number of variables, such as tissue or cell type, age, sex, nutrition, disease status, etc. Therefore, there is an urgent need to update both the laboratory IT system architecture and the database management procedures in order to cope with such large and frequent additions of data. In this context, database repositories for gene expression become essential knowledge resources able to store data in a safe and yet easily retrievable manner. Finally, it is important that this information be archived according to standards, e.g., minimum information about a microarray experiment (MIAME) (http://www.mged.org/Workgroups/MIAME), which will then allow scientists to share common information and make valid comparisons among experiments. While availability of public and private data repositories has given individual scientists a potential wealth of information, it has also made it very difficult to find the answer to specific questions. More than 500 life science databases have been reported in the literature (4), which is exemplary of the difficulty related to where and how to query for information in a fast and efficient manner (5).

In an effort to guide the individual researcher approaching a global gene expression experiment, the following review walks through a generalized workflow of a typical microarray experiment, making reference to the databases and tools that are available for each step in the process. The paper will use two studies that have been published as examples of how our proposed workflow could be applied. One study identifies genes that are differently expressed in breast cancer using microarrays and SAGE (6). The other study uses gene expression profiles to predict the clinical outcome of breast cancer (7). While we use the

former to illustrate how deposited data could be used for the planning of a microarray study, we use the latter to illustrate how the obtained microarray data could be integrated in publicly available databases. The work flow is seen from a database perspective according to the following logic of a typical microarray research project: (*i*) screen existing data prior to the experiment (a search of microarray data repositories for experiments involving breast cancer-specific genes and breast cancer tissue); (*ii*) generate the experimental data (a high-density microarray experiment comparing nontumor tissue and/or a pool of reference tissues vs. tumor tissue); (*iii*) collect and manage data (microarray data management systems to analyze and compare the data); (*iv*) analyze interesting sequences (gene annotation through database references); and (*v*) deposit and/or archive new data into repositories. The data analysis and annotation step, so crucial to the interpretation of the experiments, is expanded upon in Figure 2. The paper principally discusses microarray databases, but also includes SAGE and EST libraries in order to give a review of database science generally applied to global gene expression studies.

TECHNOLOGY REVIEW FOR GLOBAL GENE EXPRESSION ANALYSIS

Two approaches have been proposed for large-scale analysis of mRNA expression levels: (*i*) analog methods, such as cDNA chips (8) and oligonucleotide chips (9,10), which provide a continuous analytical signal based on hybridization; and (*ii*) digital methods that are based on the generation of ESTs from nonnormalized cDNA libraries, which provide a specific count of the number of sequences for each gene and/or transcript. DNA-microarray technology for gene expression analysis can be further broken down into additional subcategories, such as cDNA vs. oligonucleotide probes spotted onto a planar surface of either glass, silicon, or membranes with either single or dual-color detection by either fluorescent or autoradiographic detection methods. Digital sequencing technologies include complete library sequencing, SAGE (11), massively parallel signature sequencing (MPSS) from Lynx Supply (Oak Ridge, TN, USA) (12), and others, which are only now emerging. It is important to note that microarray platforms tend to provide information about relative expression levels, whereas the EST sequencing methods provide information on absolute expression levels. This difference needs to be taken into account when comparing results in silico.

In brief, data from different technological platforms may be compared to develop confidence in measured differential gene expression. All of the different experimental platforms eventually return data sets consisting of pairs (gene identifier: expression value), which must be stored in databases in order to enable all the processes shown in Figures 1 and 2 to be performed properly. The appropriate presentation of this information must be carefully done so as to make comparisons both accurate, intuitive, and respective of the limitations and advantages of each technology.

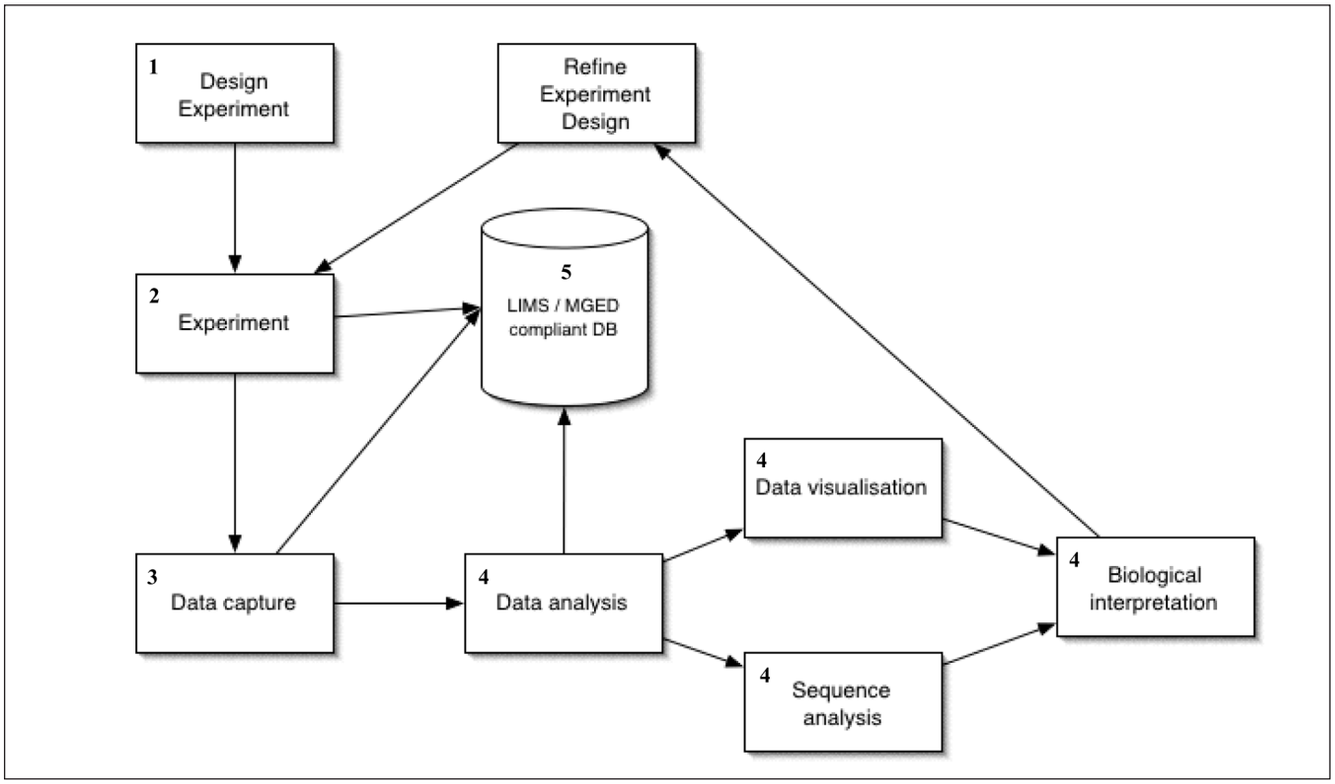

Figure 1. Microarray data analysis workflow. Existing data (repository) (1) → generate data (2) → collect and manage data (microarray data management systems) (3) → analyze interesting sequences (4) → depositing into repositories (5).

In the study by Nacht et al., SAGE data was obtained from independent primary cultures derived from normal mammary epithelial cell lines and breast cancer cell lines (6). Alternatively, one could extract this information from publicly available SAGE and EST library data, such as the Cancer Genome Anatomy Project (13). Secondly, a cDNA microarray experiment was performed comparing the same cell lines, but in addition, also comparing them to breast cancer tissue. In this case, the researcher would have two very different sets of data that both seek to answer the same question, "Which gene transcripts are differently regulated in a breast tumor compared to healthy tissue?". Clearly, many databases and gene analysis tools will be needed to answer this fairly simple experimental comparison that now involves tens of thousands of genes, none of which the principal investigator is an expert on and which are derived from different technologies in different laboratories at different times.

DATABASE REQUIREMENTS

The core function of a data management system is to store, process, and allow for data visualization. The first prerequisite of the system is its storage and archival capability. However, using a simple flat file system makes it difficult to maintain and link biological annotations that are essential for the interpretation of the data. Therefore, the second prerequisite is the development of a relational database. A microarray laboratory is usually run with several people participating in the common experimental workflow of array hybridization, scanning, data processing, and analysis. All members of the team will eventually need to visualize the data. Saving flat files at different locations after each step of the workflow breaks the information stream. Therefore, the third prerequisite is a system that centralizes the data, with an administration system that allows different users to act upon the data at different levels. The inference power of high-throughput expression data systems relies on comparing as many expression profiles as possible with each other and with other sources of information. Therefore, the fourth prerequisite is to be able to store the new data in a compatible format that allows such comparisons. Finally, the fifth prerequisite is a system that implements different protocols of data treatment, giving the investigators maximum flexibility to analyze data from different experimental designs. The final choice of which application to adopt may rely on the current system architecture of the laboratory. Specifications of current IT solutions to microarray data management are detailed in Table 1.

DATA REPOSITORIES FOR MICROARRAY, SAGE, AND EST DATA

Table 2 shows a few databases that can store, retrieve, and compare global gene expression information data (for a review of some of these see Reference 14). One emerging feature of such data repositories is the use of accepted scientific standards specific to microarray data. This capability is critical to import data from different sources, to quickly add gene annotation data, and to allow complex queries of the data using standard language or controlled vocabulary. This conceptual framework has been formalized by the Microarray Gene Expression Data Consortium (MGED) (http://www.mged.org/Workgroups/MIAME/miame.html) and has currently been adopted by prominent journals (15,16). Many data repositories are expected to follow soon. Another standard to emerge is microarray gene expression markup language (MAGE-ML) (16). This is a descriptive language widely adopted by several microarary database systems and applications. Both ArrayExpress (17) and the Gene-Expression Omnibus (GEO) (18), which are two of the most prominent microarray data repositories, intend to support the MIAME and MAGE-ML standards. These data, independent

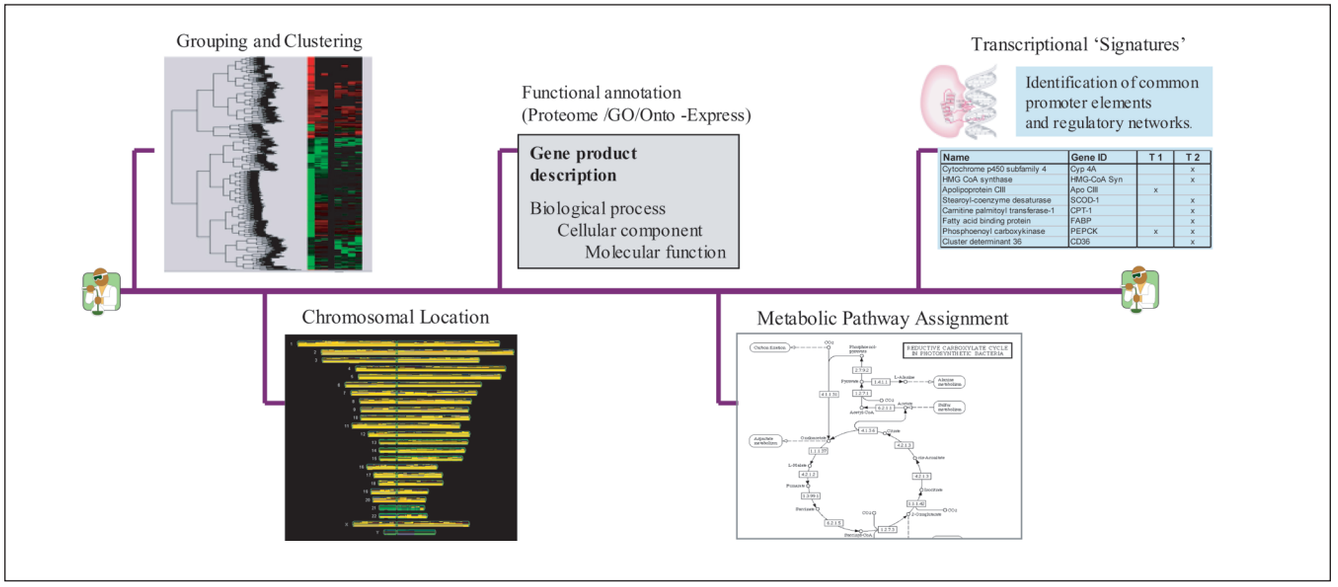

Figure 2. Bioinformatics processing of differentially expressed genes utilizing commercial software. Public and private databases are the sources of information to supply annotations of chromosomal location, GO, metabolic pathway assignment, and promoter binding sites.

Table 1. Microarray Software Specifications

Name	Minimum Hardware Requirements	Server OS	DBMS, Server Applications	GUI
Acuity	Single processor minimum 500 MHz, 1024 MB RAM, 10 GB hard disk space	Windows® 2000 or Windows 2000 Server operating system	MicroSoft® SQL Server	Internet Explorer
ArrayDB	Single processor minimum 1.0 GHz, 512 MB RAM, 40 GB hard disk space	Unix® (Linux or Irix)	Sybase®; Sybase client environment, Perl modules	CGI scripts and Java® applets: multi-experiment viewer
ArrayInformatics	Pentium® III 1GHz processor, 512 MB RAM, 40 GB hard disk space, 10/100 MB	Windows 2000 Professional	MicroSoft SQL Server	Internet Explorer
BASE	Single processor minimum 500 MHz, 256 MB RAM, 100 GB hard disk space	Unix (Solaris, Linux), MacOSX	MySql	Web browser
Expressionist	512 MB RAM	Unix	Oracle8i, Web server	Expressionist GUI
GeneDirector	Single processor minimum 1.5 GHz, 1024 MB RAM, 10 GB hard disk space	Unix	Oracle8i	X-Windows-based GUI
GeNet	Pentium III processor or equivalent or faster, 1 GB RAM recommended, 80 GB hard disk space	Windows 2000, XP or Unix (Linux or Solaris recommended)	Oracle8i	Web browser
GeneTraffic(Multi)	Biprocessor Pentium III, 4 or Xeon, 512 MB RAM, 2 GB hard disk space	Unix (Linux Red Hat recommended)	PostgreSQL, Apache Web, R statistical language	GeneTraffic GUI
GeneX	Single processor minimum 500 MHz, 256 MB RAM, 20 GB hard disk space	Unix (Linux)	PostgreSQL	Client-side Java application, X-Windows-based explorer (e.g., IBM data explorer), Web browser
maxdSQL	Single processor minimum 1.0 GHz, 512 MB RAM, 40 GB hard disk space	Unix (Solaris recommended)	MySql, Oracle8i, PostgreSQL	MaxdView in the ISYS environment
NOMAD	Single processor minimum 500 MHz, 256 MB RAM, 20 GB hard disk space	Unix (Linux recommended), MacOSX	MySql, Perl	Web browser
PartisanarrayLIMS	Single processor minimum 500 MHz, 256 MB RAM	Unix (Linux recommended)	Oracle8i or Sybase	Web browser
Resolver	2 UltraSPARC III Cu processors, 2 GB hard disk space	Unix (Solaris)	Oracle8i Enterprise	Rosetta Resolver system Image Viewer
SMD	Single processor minimum 500 MHz, 256 MB RAM, 10 GB hard disk space	Unix (Solaris recommended, Linux)	Oracle, Perl, and various modules, Java RTE	Web browser

Table 2. GeneExpression Repositories for Publicly Available SAGE Data and EST Library Data

Name	Data Type	Tissue Type	Description	Web address
GEO	Microarray/SAGE	Normal and tumor	Gene expression and hybridization array data repository.	http://www.ncbi.nlm.nih.gov/geo/
RAD	Microarray/SAGE	Normal and tumor	The ultimate goal is to allow comparative analysis of experiments performed by different laboratories using different platforms and investigating different biological systems.	http://www.cbil.upenn.edu/RAD2/
ExpressDB	Microarray/SAGE	Yeast	Collection of yeast RNA expression data sets.	http://arep.med.harvard.edu/cgi-bin/ExpressDByeast/EXDStart
CleanEx	Microarray/EST libraries	Normal and tumor	Gene expression and hybirdization array data repository. SAGE will be added.	http://www.epd.isb-sib.ch/cleanex/
Gene Expression Database	Microarray	Tumor	Data from 60 cancer cell lines based on Affymetrix (Santa Clara, CA, USA) and cDNA technology.	http://discover.nci.nih.gov/arraytools
SMD	Microarray	Normal and tumor	Extensive collection of cDNA microarray data.	http://genome-www.stanford.edu/microarray
SAGEmap	SAGE	Normal and tumor	Data from one hundred SAGE Cancer Genome Anatomy Project (CGAP) libraries.	http://www.ncbi.nlm.nih.gov/SAGE/
SAGE	SAGE	Normal and tumor	SAGE data from over 600 000 transcripts, including SAGE data from human, mouse, and yeast transcripts.	http://www.sagenet.org/SAGEData/sagedata.htm
UniGene	EST libraries	Normal and tumor	Collection of EST libraries from different species.	http://www.ncbi.nlm.nih.gov/UniGene/
CGAP/Tissue	EST libraries	Normal and tumor	Information on CGAP and other cDNA libraries.	http://cgap.nci.nih.gov/Tissues/xProfiler
BodyMap	EST libraries	Normal and tumor	Database of expression information of human and mouse genes in various tissues and cell types.	http://bodymap.ims.u-tokyo.ac.jp
TissueInfo	EST libraries	Normal	Information on tissue expression profile of a sequence by comparing the given sequence against the EST database. Each EST comes from a library derived from a specific tissue type.	http://icb.mssm.edu/services/tissueinfo/query

of platform, may be used as predictors or as references for future gene expression studies. The availability of such data is particularly important in fields such as cancer research, where access to raw material may be limited.

Our example studies would have at this stage generated high-density array data. These data would need to be characterized using MIAME descriptors and uploaded into a microarray repository. Whereas the data produced in our first example have not been deposited so far, the data from the second study have been made publicly available upon publication of the study (www.rii.com/publications/default.htm). Besides the Web page provided by the authors, the datasets have also been integrated in the CleanEX database. CleanEx is an example of a database that facilitates comparisons of expression profiles from different technologies. CleanEx makes all data concerning the same gene accessible under the same gene name (cf. Table 2). Whereas some studies publish only tables of some differently regulated genes (6), others make whole data sets publicly available (7,19). To protect intellectual property interests, this might be first done locally, but then later sent to the National Center for Biotechnology Information (NCBI)'s GEO or the European Bioinformatics Institute (EBI)'s ArrayExpress for public access. Alternatively, NCBI offers researchers up to 6 months of private access to support intellectual property applications or the review process in journals demanding data in advance (18); a trend which will likely continue. The advantage of depositing data publicly will come from new ideas and new comparisons that would not be possible within one's own institution. High quality database submissions will slowly become well known as more researchers use and reference the data.

MICROARRAY DATA MANAGEMENT SYSTEMS

Table 3 lists several major academic and commercial systems that comply with the list of criteria detailed in the section entitled Technology Review for

Table 3. Microarray Database Management Systems

Name	Supplier	License Type	Web Address
Acuity3.0	Axon Instruments	Commercial, perpetual license.	http://www.axon.com/GN_Acuity.html
ArrayDB	National Human Genome Research Institute (NHGRI)	Public domain.	http://genome.nhgri.nih.gov/arraydb/
ArrayInformatics	PerkinElmer Life Sciences	Commercial, perpetual license agreement.	http://lifesciences.perkinelmer.com/areas/microarray/arrayinfo1.asp
BASE	Lund University	GNU General Public License.	http://base.thep.lu.se/
Expressionist	GeneData	Commercial.	http://www.genedata.com/products/expressionist/
GeneDirector	BioDiscovery	Commercial, annual license, Per Seat Client Access License, with respect to modules use.	http://www.biodiscovery.com/genedirector.asp
GeNet	Silicon Genetics	Commercial, one off installation cost, including technical support and five Per Seat Client Access License.	http://www.silicongenetics.com/cgi/SiG.cgi/Products/GeNet/index.smf
GeneTraffic	Iobion Informatics LLC	Commercial perpetual license, per seat client access license.	http://www.iobion.com/products/products.html
GeneX	National Center for Genome Research (NCGR)	1. GNU Lesser General Public License. 2. Commercial license for commercial purposes.	http://www.ncgr.org/genex/
maxdSQL	University of Manchester	End-User License Agreement (EULA).	http://bioinf.man.ac.uk/microarray/maxd/maxdSQL/
NOMAD	UCSF, UCLA, Lawrence Berkeley National Laboratory	GNU General Public License.	http://ucsf-nomad.sourceforge.net/help/
PartisanarrayLIMS	Clondiag	Commercial, perpetual license, Per Seat Client Access License.	http://www.clondiag.com/products/sw/partisan/
Resolver	Rosetta Inpharmatics, Inc.	Commercial, perpetual or annual licenses, floating seat license.	http://www.rosettabio.com/products/resolver/default.htm
SMD	Stanford University	Royalty-free, nonexclusive, and nontransferable license, upon terms set by Stanford University.	http://genome-www5.Stanford.EDU/MicroArray/SMD/download/

GNU General Public License means that the source code is freely distributed and available to the general public. The EULA entitles the licensee to a royalty-free nonexclusive nontransferable end-user license to use the software solely for academic research and no other purposes.

Global Gene Expression Analysis, some of which have been reviewed previously (14). The applications listed in Tables 3 and 4 can be broken down into three subcategories: (*i*) data archives; (*ii*) integrated microarray laboratory information management systems (LIMS); and (*iii*) data analysis packages with file management capabilities. Most systems are designed using a three-tier architecture. The three-tiered architecture involves a database server (also known as back-end), an application layer (middle layer), and a graphical user interface (front-end GUI). This architecture is intended to accommodate many kinds of expression data at all steps within the generalized workflow shown in Figure 1. All systems listed support at least data storage, some data analysis, and data visualization. The main feature of this system architecture is the centralization of the data and of the computing intensive tasks into a central machine, the server. The minimum hardware and operating requirements for these databases are listed in Table 1 (summarized for lowest profile configuration).

A truly integrated solution should be self-contained. From both the end-user's and system administrator's point of view, this translates to a single point of entry, which is the front-end GUI and the server side configuration file, respectively. The GUI is either a common Web browser application or a small client application that communicates with both the server and the user. A fully integrated system allows the user to operate data capture, data filtering, and data normalization under the same application, no matter which data are coming in. This offers the great advantage that, at the end of the data preprocessing and normalization procedure, the users can interrogate different data sets with respect to a given property (e.g., log average difference greater than a given threshold) and ask

Table 4. A Technical Comparison of Commercially Available Microarray Storage and Analysis Packages

Name	Archival	Treatment	Visualization	Data Normalization Protocols and Data Analyses Modules
Acuity	Dual-color cDNA/oligo	Dual-color cDNA/oligo	Dual-color cDNA/oligo. Dendrograms, 2-D interactive plots, animated interactive 3-D plots, line graphs, scatter plots.	Global normalization, normalization on control spots, spike controls, or subset of spots. Hierarchical clustering, *k*-means, principal component analysis (PCA), self-organizing map (SOM).
ArrayDB	Dual-color cDNA/oligo	Dual-color cDNA/oligo	Dual-color cDNA/oligo.	Global mean or median ratio based normalization.
ArrayInformatics	Dual-color cDNA/oligo	Dual-color cDNA/oligo	Dual-color cDNA/oligo, Affymetrix, Scatter, line and series plots and a cluster image map is not supporting XML as of yet.	Normalization to Lowess, total intensity, median ratio or to a user generated gene list, graphing data trends after normalization enabling examination of data variability.
BASE	Dual-color cDNA/oligo, Affymetrix, SAGE	Dual-color cDNA/oligo, Affymetrix, SAGE	Dual-color cDNA/oligo, Affymetrix, SAGE.	Global mean or median ratio based normalization, Lowess, MDS module.
Expressionist	Affymetrix	Affymetrix	Affymetrix, dual-color cDNA/oligo.	Standard data processing and clustering.
GeneDirector	Dual-color cDNA/oligo	Dual-color cDNA/oligo	Dual-color cDNA/oligo, Affymetrix.	ImaGene and GeneSight packages.
GeNet	Dual-color cDNA/oligo, Affymetrix	Dual-color cDNA/oligo, Affymetrix	Dual-color cDNA/oligo, Affymetrix.	GeneSpring package.
GeneTraffic(Multi)	Filters, dual-color cDNA/oligo, Affymetrix	Filters, dual-color cDNA/oligo, Affymetrix	Filters, dual-color cDNA/oligo, Affymetrix.	Global normalization, z-score, Lowess normalization, full and sub-grid, for Affymetrix, alternative probe based protocol.
GeneX	Dual-color cDNA/oligo, Affymetrix	Dual-color cDNA/oligo	Dual-color cDNA/oligo, Affymetrix.	R routines are available to manipulate the data (normalization, clustering, etc.).
maxdSQL	Dual-color cDNA/oligo, Affymetrix	Dual-color cDNA/oligo, Affymetrix	Dual-color cDNA/oligo, Affymetrix, maxdView, expression data class which represents results from one or more hybridizations and any associated clusters of genes. Profiles viewers.	Filtering based on numerical values. 2-D correlation plot with overlay of cluster data, multidimensional plots.
NOMAD	Dual-color cDNA/oligo, Axon scanner outcome	Dual-color cDNA/oligo, Axon scanner outcome	Dual-color cDNA/oligo, Axon scanner outcome.	ScanAlyse package: global normalization.
PartisanarrayLIMS	Filters, dual-color cDNA/oligo, Affymetrix	Filters, dual-color cDNA/oligo, Affymetrix	Filters, dual-color cDNA/oligo, Affymetrix.	Global mean or median ratio based normalization.
Resolver	Affymetrix, Nylon filters, dual-color cDNA/oligo	Affymetrix, Nylon filters, dual-color cDNA/oligo	Affymetrix, Nylon filters. Table Viewer: *k*-means, *k*-medians clustering, and SOM algorithms.	Error models with any experimental replicates performed, *P*-values computed and error bars for every gene expression measurement, analysis of variance (ANOVA).
SMD	Dual-color cDNA/oligo	Dual-color cDNA/oligo	Dual-color cDNA/oligo.	ScanAlyse package: global normalization.

for entries that are present in all queries output. Even the initial data collection, independent of technology platform, can be automated in an integrated system through LIMS. Many LIMS options are supported by BioArray Software Environment (BASE) (20), PartisanarrayLIMS, and ArrayInformatics.

In the present example, the database and analysis tools listed in Tables 3 and 4 would be used to scale, normalize, cluster, manipulate, and visualize the global data sets to fairly compare gene expression between probe sets and to group genes with similar gene expression patterns across the distinct conditions. It might be discovered, here, that a new transcription factor has an opposite pattern of regulation with a tightly clustered set of 20 genes composed mainly of known or proposed oncogenes. This analysis might then lead to the hypothesis that the transcription factor is directly involved in the inhibition of oncogenes. Secondly, these same tools could also attempt to correlate the EST library data with those from the microarray. Finally, the database tools would track, archive, and backup all of the analysis and visualization steps for future reference.

ANNOTATION OF RESULTS USING DATABASE REFERENCES

Independently of the platform and the analysis methods used, the result of most microarray experiments is a list of differentially expressed genes. Most data analysis methods available in the tools surveyed concentrate on this aspect (21). Many researchers parse such lists of genes manually, using literature searches and browsing several public databases in an attempt to extract the relevant biological processes and pathways. This is an extremely tedious and error-prone process that usually takes many months. Thus, a major challenge is to translate these lists of differentially regulated genes into a better understanding of the underlying biological phenomena in an automated fashion.

Commonly, gene expression information is interpreted in the context of annotation information from a variety of sources. As shown in Figure 2, common sources of annotated information added in the BioIT (i.e., analysis process) include: (*i*) updated molecular information from databases such as SwissProt and Ensembl; (*ii*) Gene Ontology (GO) classification; (*iii*) metabolic pathway assignment; (*iv*) assignment of chromosomal location of genes; and (*v*) transcription factor activation, among others. The most common data sources and references shown in this scheme will be discussed below.

Gene annotation is one of the major efforts of the bioinformatics community today. Ensembl is a public project developed at the EBI that has managed to integrate resources from different databases such as SwissProt, Interpro, GO, Refseq, Locuslink, and other interesting annotation sources (22). A very interesting feature of Ensembl is the distributed annotation systems (DAS) server, which facilitates the integration of annotation from different sources (23). For example, two laboratories can share annotations and map them to Ensembl. In gene expression, one can map the probes to Ensembl and retrieve all the annotations for this specific probe. Although Ensembl is already widely used, much work is yet to be done to integrate other important data sources.

The study of genomics is contributing to a better understanding of diverse organisms through a unification of biology through comparative sequence analysis. This has led to progress in the way that biologists describe and conceptualize the shared biological elements, but has not kept pace with the growth of primary data. The exponential growth in the volume of accessible biological information has generated a confusion of voices surrounding the annotation of molecular information about genes and their products. Based on a number of model organisms, which constantly grows, a dynamic, structured, precisely defined, controlled vocabulary for describing the roles of genes and their products in any organism was formed under the GO Consortium. The goal of this effort is to address the deficits of the current, rather divergent, nomenclature schemes (24,25). To this end, three independent ontologies are being constructed, which are used as attributes for gene products, namely biological process, molecular function, and cellular component.

Until very recently, there were no tools available to classify genes according to the GO structure in an automated fashion. Onto-Express (OE) has been recently made available as a tool designed to mine the available functional annotation data and help the researcher find relevant biological processes (26). The result of this analysis is a functional profile of the condition studied. In the latest version of the OE software, this functional profile is accompanied by the computation of significance values for each functional category (27). Such values allow the user to distinguish between significant biological processes and biological processes affected by random events. OE's utility has been demonstrated by analyzing data from two recent breast cancer studies. In our second example (7), the functional annotation of genes provided insights into the biological mechanisms leading to rapid metastasis. The authors of the original paper showed that genes involved in the cell cycle, invasion and metastasis, angiogenesis, and signal transduction were significantly up-regulated in the poor diagnosis signature. The original work required many months of analysis. A tool such as OE is able to retrieve these processes in a few minutes. This tool and this sample dataset are available on line at (http://vortex.cs.wayne.edu/Projects.html).

Metabolic pathway information is also extremely important for gene expression profiling but has been highly under utilized. Although the primary causative factors in disease are altered protein activities or altered biochemical composition of cells and tissues, changes at the genetic level might be the ultimate cause for the disease. Thus, the link between the gene regulatory control and the primary causative factors will be crucial for application in drug development, medicine, nutrition, and other therapeutic courses of action. The identification of relationships between genes, transcripts, proteins, and metabolites are essential components to understand integrative metabolism (28,29). The annotation of genes with such information has been attempted by a number of public efforts, most notably from Kyoto Encyclopedia of Genes and Genomes (KEGG) (30–32). Software is now available to superimpose gene expression data onto said pathways, providing a powerful means to identify biological regulation of metabolism through the co-expression of gene data obtained from microarrays. GenMAPP is a useful tool for such purpose, allowing the user to link pathway information to gene expression data (33).

Lastly, as it is typically beyond the average researcher to be an expert on the functions of what may potentially be hundreds of

co-regulated genes, it will be important to annotate those genes with GO or keyword searches in the literature or link them directly to top literature hits themselves (34). This basic example points out that, by going through expectations in a systematic fashion, one will quickly be pointed to the appropriate database sources of information, i.e., Ensembl for molecular annotation, GO for cross-gene functionality, and PubMed for literature information. The intelligence and efficiency of viewing and understanding this information is the subject of modern database science and the various analytical tools being developed to sort through this information.

CONCLUSIONS

The ability to store and query genetic data has become crucial to the progress of scientific research in the life sciences. This paper reviewed a number of databases and tools addressing this vital issue. A number of requirements for such databases have been identified and discussed in the context of existing tools. The paper also reviewed a number of related tools able to analyze the vast amount of data stored in currently available repositories to better understand the results of the experiments undertaken.

REFERENCES

1. **Bustin, S.A. and S. Dorudi.** 2002. The value of microarray techniques for quantitative gene profiling in molecular diagnostics. Trends Mol. Med. *8*:269-272.
2. **Gill, R.T., S. Wildt, Y.T. Yang, S. Ziesman, and G. Stephanopoulos.** 2002. Genome-wide screening for trait conferring genes using DNA microarrays. Proc. Natl. Acad. Sci. USA *99*:7033-7038.
3. **Ye, S.Q., T. Lavoie, D.C. Usher, and L.Q. Zhang.** 2002. Microarray, SAGE and their applications to cardiovascular diseases. Cell Res. *12*:105-115.
4. **Baxevanis, A.D.** 2002. The Molecular Biology Database Collection: 2002 update. Nucleic Acids Res. *30*:1-12.
5. **Stein, L.** 2002. Creating a bioinformatics nation. Nature *417*:119-120.
6. **Nacht, M., A.T. Ferguson, W. Zhang, J.M. Petroziello, B.P. Cook, Y.H. Gao, S. Maguire, D. Riley, et al.** 1999. Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. Cancer Res. *59*:5464-5470.
7. **van't Veer, L.J., H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, et al.** 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*:530-536.
8. **Schena, M., D. Shalon, R.W. Davis, and P.O. Brown.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*:467-470.
9. **Kane, M.D., T.A. Jatkoe, C.R. Stumpf, J. Lu, J.D. Thomas, and S.J. Madore.** 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Res. *28*:4552-4557.
10. **Pease, A.C., D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes, and S.P. Fodor.** 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc. Natl. Acad. Sci. USA *91*:5022-5026.
11. **Velculescu, V.E., L. Zhang, B. Vogelstein, and K.W. Kinzler.** 1995. Serial analysis of gene expression. Science *270*:484-487.
12. **Brenner, S., M. Johnson, J. Bridgham, G. Golda, D.H. Lloyd, D. Johnson, S. Luo, S. McCurdy, et al.** 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. *18*:630-634.
13. **O'Brien, C.** 1997. Cancer genome anatomy project launched. Mol. Med. Today *3*:94.
14. **Gardiner-Garden, M. and T.G. Littlejohn.** 2001. A comparison of microarray databases. Brief Bioinform. *2*:143-158.
15. **Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, et al.** 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat. Genet. *29*:365-371.
16. **Spellman, P.T., M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, et al.** 2002. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. *3*:RESEARCH0046.
17. **Brazma, A., U. Sarkans, A. Robinson, J. Vilo, M. Vingron, J. Hoheisel, and K. Fellenberg.** 2002. Microarray data representation, annotation and storage. Adv. Biochem. Eng. Biotechnol. *77*:113-139.
18. **Edgar, R., M. Domrachev, and A.E. Lash.** 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*:207-210.
19. **Bates, M.D., C.R. Erwin, L.P. Sanford, D. Wiginton, J.A. Bezerra, L.C. Schatzman, A.G. Jegga, C. Ley-Ebert, et al.** 2002. Novel genes and functional relationships in the adult mouse gastrointestinal tract identified by microarray analysis. Gastroenterology *122*:1467-1482.
20. **Saal, L.H., C. Troein, J. Vallon-Christersson, S. Gruvberger, A. Borg, and C. Peterson.** 2002. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol. *3*:SOFTWARE0003.
21. **Draghici, S.** 2002. Statistical intelligence: effective analysis of high-density microarray data. Drug Discov. Today *7*:S55-S63.
22. **Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, et al.** 2002. The Ensembl genome database project. Nucleic Acids Res. *30*:38-41.
23. **Dowell, R.D., R.M. Jokerst, A. Day, S.R. Eddy, and L. Stein.** 2001. The Distributed Annotation System. BMC Bioinformatics *2*:7.
24. **Consortium TGO.** 2001. Creating the gene ontology resource: design and implementation. Genome Res. *11*:1425-1433.
25. **Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, et al.** 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*:25-29.
26. **Khatri, P., S. Draghici, G.C. Ostermeier, and S.A. Krawetz.** 2002. Profiling gene expression using onto-express. Genomics *79*:266-270.
27. **Draghici, S., P. Khatri, R.P. Martins, G.C. Ostermeier, and S.A. Krawetz.** Global functional profiling of gene expression. Genomics (In press.)
28. **Roberts, M.A., D.M. Mutch, and J.B. German.** 2001. Genomics: food and nutrition. Curr. Opin. Biotechnol. *12*:516-522.
29. **Berger, A., D.M. Mutch, J.B. German, and M.A. Roberts.** 2002. Dietary effects of arachidonate-rich fungal oil and fish oil on murine hepatic and hippocampal gene expression. Lipids Health Dis. *1*:2.
30. **Kanehisa, M., S. Goto, S. Kawashima, and A. Nakaya.** 2002. The KEGG databases at GenomeNet. Nucleic Acids Res. *30*:42-46.
31. **Kanehisa, M. and S. Goto.** 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*:27-30.
32. **Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa.** 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *27*:29-34.
33. **Dahlquist, K.D., N. Salomonis, K. Vranizan, S.C. Lawlor, and B.R. Conklin.** 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat. Genet. *31*:19-20.
34. **Chaussabel, D. and A. Sher.** 2002. Mining microarray expression data by literature profiling. Genome Biol. *3*:RESEARCH0055.

Address correspondence to:

Dr. Matthew A. Roberts
Nestle Purina Pet Care
Mail Zone 11T
Number One Checkerboard Square
St. Louis, MO 63164, USA
e-mail: mroberts@purina.com