Progress in the Application of DNA Microarrays
Edward K. Lobenhofer, Pierre R. Bushel, Cynthia A. Afshari, and Hisham K. Hamadeh

Microarray technology has been applied to a variety of different fields to address fundamental research questions. The use of microarrays, or DNA chips, to study the gene expression profiles of biologic samples began in 1995. Since that time, the fundamental concepts behind the chip, the technology required for making and using these chips, and the multitude of statistical tools for analyzing the data have been extensively reviewed. For this reason, the focus of this review will be not on the technology itself but on the application of microarrays as a research tool and the future challenges of the field. Key words: DNA chips, gene expression profiling, microarray, toxicogenomics. Environ Health Perspect 109:881-891 (2001). [Online 16 August 2001] http://ehpnet1.niehs.nih.gov/docs/2001/109p881-891lobenhofer/abstract.html

The Herculean effort to sequence every nucleotide in a genome has generated complete sequences for many of the common experimental organisms, including Saccharomyces cerevisiae, Drosophila melanogaster, and Caenorhabditis elegans. Significant progress has also been made sequencing entire genomes for more complex species, such as mice and humans. The analysis and application of this information have the potential to advance rapidly our understanding of the molecular mechanisms underlying normal cellular processes as well as the molecular basis for disease. A limitation of this information is that complete sequence data merely enable the identification of all theoretical genes within a genome. Sequence information alone does not indicate either the expression patterns for a given gene or the cell type in which it is expressed.

The functional expression of genes may be assessed by measuring the levels of mRNA and their corresponding proteins. Previously, one could assess the expression of mRNA for known genes or unidentified genes or expressed sequence tags (ESTs) using somewhat labor-intensive, low-throughput techniques, including Northern blot analysis and RNase protection assays. The advent of newer technologies such as quantitative reverse transcription-polymerase chain reaction (RT-PCR) and DNA chips, or microarrays, has enabled rapid and simultaneous comparison of mRNA levels for hundreds to thousands of genes in virtually any biologic sample.

The number of publications employing microarrays has undergone exponential growth since its inception in 1995. The original paper by Schena et al. (1) detailing this technique to monitor gene expression changes has been referenced over 700 times. Furthermore, a search of published literature using the keyword “microarray” or “DNA chip” yields over 800 articles (Figure 1). Most of these articles have focused on studies related to cancer research. Microarrays have also been used to study questions in pathology, cell biology, pharmacology, and toxicology. The technology and methods involved in microarray experiments have been the focus of several reviews (2-9). Therefore, we will summarize briefly the principle of microarray, and then compile a summary of studies in which the application of DNA microarrays has addressed critical questions in different scientific fields as well as discuss several future challenges.

Oncology
In the past five years, microarray technology has been used as a tool in many studies in cancer research. Many of these reports are listed in Table 1, which categorizes the studies by tumor location. Microarrays have been used in cancer research to address three main objectives: determining the molecular differences between normal and malignant cells; improving the classification of tumors to increase the effectiveness of therapeutics; and identifying mutations in genes that are implicated in tumor formation or progression.

Microarray technology has revolutionized the manner in which cancer researchers identify genes that are differentially expressed during tumor progression. Using breast tissue that was obtained immediately after a modified radical mastectomy, Sgroi et al. (10) examined differences in gene expression in the tumor and normal tissue from a single patient. The tissue was frozen, sectioned, and stained using hematoxylin and eosin to distinguish normal breast epithelial cells from invasive carcinoma cells as well as metastatic cells (from an axillary lymph node). Each cell population (normal, primary tumor, metastatic tumor) was isolated using laser capture
microarray technology to avoid cross-contamination, inclusion of adipose tissue, stromal components, or lymphocytes that may have been associated with an immune response to the disease. Isolation of different cell populations from a single patient meant that any observed gene expression changes could potentially be attributed directly to disease progression, a task that would be more difficult if the study had been performed using tissue from multiple women. Sufficient numbers of each cell type were harvested using laser capture microscopy to yield enough RNA for microarray analysis without using RNA amplification procedures (11, 12). (Appropriate cell numbers will vary depending upon the cell type and tissue of origin, and the amount of RNA depends on the microarray platform and the labeling protocol.) RNA amplification should be used with caution to avoid the risk of misrepresentation of gene expression biases (12-14).

This study (10) identified genes that were differentially expressed in invasive or metastatic breast cancer cells compared with levels present in normal breast epithelial cells. To increase the statistical confidence in the data, the analysis included data from replicate hybridizations. Many of the differentially expressed genes were verified by quantitative RT-PCR and immunohistochemical assays. A significant proportion (41%, 83 of 202) of the differentially expressed genes were unknown; however, many of the known genes were previously implicated in tumor biology or specifically in breast cancer.

For example, the increased expression of tissue factor (also known as thromboplastin or coagulation factor III) has been described previously during angiogenesis and metastasis (reviewed by Ruf and Mueller (15)). In a more recent study (16), increased levels of tissue factor protein have been detailed in the tumor, stroma, and plasma of breast cancer patients. The microarray data demonstrated a 5-fold increase in tissue factor transcripts in primary tumor cells relative to normal breast epithelial cells. This finding corroborated the association of increased levels of tissue factor with tumor progression, and also indicated that tissue factor expression is under transcriptional control.

Cumulatively, this study demonstrated that the use of carefully controlled experimental parameters (unamplified RNA from distinct cell populations obtained from one individual patient) in the preparation of samples for microarray analysis can yield a descriptive examination of the transcriptional changes associated with tumor progression.

Microarrays have also been used to study tumor progression in cell cultures. It has been reported (17) that the human melanoma cell line UACC-903 can be reverted to a normal growth rate after introduction of a normal human chromosome 6 (UACC-903(-6)), in a more recent study (18), fluorescently labeled cDNA probes were generated from mRNA extracted from UACC-903 or UACC-903(-6) cells and hybridized to a DNA chip to identify changes in gene expression associated with the reduced rate of proliferation and tumor suppressed phenotype. The criteria used to determine whether a gene was differentially expressed identified 63 upregulated genes (7.3% of the 870 genes present on the chip) on introduction of normal chromosome 6, whereas 15 (1.7%) genes were significantly downregulated. The levels of 15 of these differentially expressed genes were verified independently by Northern blot analysis. One of the validated genes demonstrating increased expression in the UACC-903(-6) line was p21/WAF1, a mediator of p53-induced growth suppression, which correlated with the reduced growth rate observed in transfected cells. This report exemplifies the quality and usefulness of microarray data and provides further validation of the potential utility of in vitro models of cancer.

Microarray technology can facilitate accurate classifications of cancer. Historically, tumors have been classified by pathologic examination supplemented by special stains and antibodies. Similar to the way in which DNA sequencing is revolutionizing the field of taxonomy, gene expression profiling is increasing the number of markers useful in tumor classification. For example, because of standard immunohistochemical techniques that showed positive staining for neural-specific markers, small cell lung cancers were believed to originate from neuroendocrine tissue (19). Pulmonary carcinoid tumors, which are benign lung tumors, also stain positively for neuroendocrine markers and are therefore hypothesized to arise from neuroendocrine cells. Using high-density cDNA arrays, Anbazhagan et al. (20) supported this theory by demonstrating that pulmonary carcinoid tumors exhibit a gene expression profile similar to that of two neural-derived tumors, oligodendroglioma and high-grade astrocytoma. Surprisingly, the expression profile of small cell lung cancer was more similar to that of bronchial epithelial cells than to those of carcinoid tumors, oligodendrogliomas, or high-grade astrocytomas. This finding suggests that small cell lung cancer may actually be an epithelial tumor rather than neuroendocrine origin. In studies of melanomas, diffuse large B-cell lymphomas,

Figure 1. Exponential growth of microarray technology use in scientific reports. Articles published per year containing the keyword “microarray” or “DNA chip” were identified by a Medline search.

| Year | Number of publications containing “microarray” or “DNA chip” |
|------|---------------------------------------------------------|
| 1994 | 50                                                      |
| 1995 | 100                                                     |
| 1996 | 200                                                     |
| 1997 | 300                                                     |
| 1998 | 400                                                     |
| 1999 | 500                                                     |
| 2000 | 600                                                     |

Table 1. Use of microarray analysis in cancer research.

| Location of primary tumor/ tumor type | Reference |
|--------------------------------------|-----------|
| Breast                               | (10, 26, 28, 29, 31, 81-96) |
| Ovary/cervix/endometrium             | (32, 97-100) |
| Prostate                             | (201, 109) |
| Colon                                | (29, 110-115) |
| Pancreas                             | (216, 117) |
| Kidney                               | (218) |
| Blood/lymph                          | (212, 23, 119-129) |
| Liver                                | (213, 130) |
| Bladder                              | (213, 131) |
| Head/neck/oral                       | (214, 132) |
| Brain                                | (215, 133) |
| Lung                                 | (216, 140) |
| Skin                                 | (217, 150-152) |
| Rhabdomyosarcoma                     | (153-156) |
| Ewing’s sarcoma                      | (157) |
| Multiple sites/tumor types           | (63, 64, 158-161) |
and leukemias, previously uncharacterized tumor subtypes were identified based on differential gene expression patterns observed using microarrays (21–23). Ultimately, more accurate classification of cancerous lesions based on gene expression profiles may improve the therapeutic strategies employed to treat patients by allowing selection of chemotherapeutic agents that target tumor subtype-specific molecules.

Breast cancers are currently classified based on several parameters, including tumor size and stage as well as the level of estrogen receptor (ER) expression (24,25). Using several breast cancer cell lines as well as primary tumor samples, Martin et al. (26) identified clusters on a hierarchical tree (groups of genes that demonstrate similar expression patterns) that are associated with each of these clinical features. Grouping—statistical clustering of tumors according to their respective gene expression profiles—enabled discrimination of tumors based on ER status. However, several tumors were not grouped in accordance with the original immunohistochemical data. Not a shortcoming of the gene expression profiling procedure, this discrepancy represents a strength of this technique by providing additional insight into tumor biology. Some tumors are classified as ER-positive based on immunohistochemical analysis, but actually express a nonfunctional protein and therefore should be typed ER-negative. This would indicate that tumors expressing ER (determined by classic assays) but not grouped with other ER-positive tumors by clustering of gene expression data might represent this unique group of tumors. The accurate classification of these tumors using a combination of classic pathologic as well as gene expression techniques would enable physicians to predict whether these tumors would be refractive to hormonal therapy, and allow the consideration of alternative treatment regimes.

Advances in cancer research can be facilitated by studies that use microarrays to explore the function of genes implicated in tumor progression or genetic susceptibility. For example, women with a germline mutation in BRCA1 are predisposed to breast and ovarian cancer (27). BRCA1 has been implicated in the response to DNA damage, homologous recombination, and transcription. To isolate target genes of BRCA1 action, two different laboratories completed high-density array hybridizations using RNA from cells ectopically expressing BRCA1 (28,29). Consistent with the proposed action of BRCA1 in DNA repair, both groups identified genes implicated in the DNA damage response (such as GADD45) as being significantly upregulated in cells overexpressing BRCA1. Further exploration into the molecular mechanisms of BRCA1 action and the effect of gene mutations on its activity will undoubtedly provide more insight into the basis for genetic predisposition to cancer.

To detect gene mutations, a variation on array-based technology has been developed. N-mer arrays contain oligonucleotides of specific length, which correspond to a portion of the coding region of the gene of interest. Wild-type sequences as well as single-base substitutions at every position are synthesized on a chip. After hybridization, the intensity of each oligonucleotide is assessed; the oligonucleotide with the highest signal corresponds to the actual sequence for the gene being analyzed (30). This technique has identified single base insertions, deletions, and substitutions in three different tumor suppressor genes, BRCA1, BRCA2, and p53 (31,32). This modified application of microarray technology has the promise to deliver high-throughput data with the same accuracy as more popular, gel-based methods of mutation screening while being less time- and labor-intensive (32).

Infectious Disease

The underlying basis of infectious disease stems from the complex interaction between a pathogen and a host. Understanding the molecular details of this interaction will aid in the identification of virulence-associated microbial genes and host-defense strategies. This would enable characterization of the genes involved in pathogenesis and the regulatory mechanisms controlling these genes. Ultimately, this knowledge could guide the generation of new treatment regimes to combat infections. Microarray technology promises to expedite microbial research by aiding in annotation of the microbial genome, examination of microbial physiology, and identification of candidate virulence factors.

Most of the time, the function of bacterial genes is inferred by finding homologs that have been previously characterized in other organisms. The study of genomewide expression patterns will greatly enhance the annotation of gene function: Similar expression may imply related function. Thus, when a gene with an unknown function is clustered with known genes based on similar expression patterns, it is implied that the function of the unknown gene is related to that of other genes in that cluster. This hypothesis was first tested in a study examining the gene expression profile of Saccharomyces cerevisiae, and it demonstrated the coregulated expression of genes that performed similar cellular functions (33). Theoretically, expression profiling will enable the function of unknown genes to be extrapolated, even when sequence similarity has not been established with characterized genes. This is especially true for pathogens because the whole genome has been sequenced for a significant number of these organisms using rapid techniques such as the “shotgun” approach (34–36).

A series of studies demonstrated that the evaluation of mRNA expression profiles by DNA microarray analysis is a powerful approach for characterizing and understanding host–pathogen interactions (37–43). The ability to identify virulence factors using DNA microarrays has been shown by comparing gene expression of pathogen-infected and -uninfected cells. Cohen et al. (39) observed changes in the gene expression profile of the human promyelocytic THP-1 cell line after infection with Listeria monocytogenes, a pathogenic intracellular organism responsible for meningitis, meningoencephalitis, and in some cases gastroenteritis. The underlying premise of their study was that this gram-positive bacterium is able to parasitize and grow in both phagocytes and nonphagocytic target cells. The infection stimulated transient activation of signal transduction pathways in the host cell, leading to modulation of gene expression. The authors’ findings were consistent with previously published data, and also identified several novel genes associated with the infection. L. monocytogenes-regulated genes belonged to different functional categories, including inflammation, chemotaxis, transcription, apoptosis, transduction, metabolism, and cell cycle. This information will enhance the understanding of the molecular physiology of L. monocytogenes and guide the development of new therapeutic drugs.

Cellular Biology

Microarray technology has influenced cell biology by enabling characterization of different cellular events at the molecular level (Table 2). This information has increased the understanding of major biologic events such as organ development, homeostasis, as well as inter- and intraorgan communication.

Studies using microarrays have not only enhanced the understanding of known cellular processes, but also aided the discovery of previously unknown properties or functions of different cells. Iyer et al. (44) demonstrated this through analysis of the response of human fibroblast cells to serum. Genes were clustered into groups on the basis of their temporal patterns of expression. M any features of the transcriptional program appeared to be related to the physiology of wound repair, suggesting that fibroblasts may play a larger role in this complex, multicellular response than had previously been assumed.

When studying human responses, the goal is to use human-derived tissue that is
Table 2. Application of microarray technology to different scientific disciplines.

| Discipline, investigation topic | Reference |
|---------------------------------|-----------|
| Pathology                        |           |
| Pseudomonas aeruginosa           | (38)      |
| Human papilloma virus            | (42)      |
| Human immunodeficiency virus     | (39)      |
| Listeria monocytogenes           | (41)      |
| Salmonella                       | (40)      |
| Tuberculosis                     | (43)      |
| Malarial parasite                | (162)     |
| Bordetella pertussis             | (37)      |
| Toxicology/pharmacology          |           |
| Carbon tetrachloride (HepG2)     | (163)     |
| Beta-naphthflavone (mouse liver) | (164)     |
| 2,3,7,8-Tetrachlorodibenzo-p-dioxin (HepG2) | (165) |
| Copper regulation in S. cerevisiae | (166)    |
| Multidrug resistance in S. cerevisiae | (167) |
| Isoniazid                        | (168)     |
| Cell biology/cellular signaling  |           |
| Yeast cell cycle                 | (169)     |
| Yeast adaptation to growth conditions | (170-172) |
| Transcription control            | (173-179) |
| Functional diversity among kinase A subunits | (180) |
| Role of histones in telomeric silencing | (181) |
| Kinase inhibitors                | (182)     |
| Effect of clotting factor on fibroblasts | (183) |
| Differentiation of embryonic stem cells | (184) |
| erythroid maturation             | (119)     |
| M aturation of human dendritic cells | (185) |
| Response of pancreatic cells to glucose | (186) |
| Mechanically induced genes in human muscle cells | (187) |
| Mid-gestation placenta and embryo associated genes | (188) |
| Metamorphosis                    | (189)     |
| Functional diversity of T cells  | (190)     |
| Transcript profile of hematopoietic stem cells | (191) |

The use of genetic manipulations, in addition to genome-wide transcript profiling in yeast, revealed changes in gene expression underlying pheromone signaling, cell cycle control, and polarized morphogenesis (54). These experiments showed cross talk and overlap of multiple mitogen-activated protein kinase pathways such as filamentous growth and mating responses. More globally, this study demonstrated the potential to correlate gene expression changes with alterations in physiologic or developmental processes.
Numerous other studies have focused on elucidating components of signaling cascades in biologic models (Table 2). Cumulatively, these studies contribute to a greater understanding of the molecular communication within cells and how those signals mediate cellular processes.

**Toxicology and Pharmacology**

Identification of environmental carcinogens and other hazards constitutes a major challenge in the field of toxicology. The assessment of risk associated with different chemical exposures is limited partially by difficulties relating to cross-species extrapolation and dose response as well as estimations of exposure levels. Thus, one must identify trans-species, chemical-specific biomarkers that result from quantified exposures to drugs or toxic chemicals. This would improve the extrapolation of data from model systems to potential effects in humans. Many disciplines employ in vitro models to study cell processes and signals, however, extrapolating findings to humans, toxicology and pharmacology require the use of relatively large, complex organisms such as rodents, rabbits, canines, and monkeys. The process of gene discovery and deciphering these genomes is challenging, but current efforts to sequence the genomes of these organisms will greatly enhance the DNA microarray field by offering more defined sequences to monitor in the response of a biologic model to a chemical exposure.

The potential of DNA microarray technology to identify gene expression changes associated with toxic or pharmacologic processes has been the focus of several studies. For example, the biochemical and molecular mechanisms of lead neurotoxicity have not been fully elucidated; however, lead has been shown to interfere with several signal transduction pathways (55–58). Changes in gene expression in target cells is hypothesized to be a mechanism by which this chemical interferes with normal brain development (59). Hossain et al. (60) studied the mechanisms underlying lead neurotoxicity using cDNA microarray gene expression analysis and identified lead-sensitive genes in immortalized human fetal astrocytes (SV-FHA). Their findings indicated that lead induces vascular endothelial growth factor (VEGF) expression in SV-FHA as via a pathway involving protein kinase C and the AP-1 transcription factor, yet independent of hypoxia-inducible factor-1. This report highlights the potential of DNA microarrays for the discovery of novel toxicant-induced gene expression alterations and the ability to dissect the second messenger pathways and transcription factors mediating these changes. By allowing determination of the mechanisms of toxicant action, such studies will ultimately contribute to models developed for risk assessment in humans.

Microarray experiments using pharmacologic agents have focused on identification of mechanisms for drug action as well as isolation of potential drug targets. A landmark study (61) described a method for drug target validation and identification of secondary drug target effects based on genome-wide gene expression patterns. These researchers generated mutations in yeast genes encoding putative targets of the immunosuppressant drugs cyclosporin A or FK506. Expression profiles were then identified for both mutant and wild-type cells after drug treatment to verify essential components of the drug’s molecular mechanism. This research revealed pathways that FK506 affected in a calcineurin- and immunoproteasome-independent manner by induction of drug-dependent effects in “targetless” cells. The described method permitted direct confirmation of drug targets and recognition of drug-dependent changes in gene expression, including those mediated through pathways distinct from the drug’s intended target. This approach in comparison of wild-type and mutant cells may help improve the efficiency of the drug development process.

Currently, there is a limited inventory of expression profiles reflecting the response to chemical treatment in biologic models. This type of information would enhance our understanding of toxicology and pharmacology. The principal hypothesis underlying toxicogenomic studies is that chemical-specific patterns of altered gene expression can be revealed using high-density microarray analysis on the tissues from treated organisms (6). As detailed earlier, gene expression-based pattern discrimination has been effective in tumor classification (21,23,62–64); however, this problem is more complex in toxicology-based studies that are compounded by target organ, dose, and time point variables. This challenge was demonstrated in a recently published effort to distinguish classes of toxicants based solely upon cluster-type analysis of differentially expressed genes in HepG2 cells (65). Initial comparison of gene expression profiles for cytotoxic anti-inflammatory drugs and DNA-damaging agents to a database populated with gene expression profiles from 100 toxicants failed to differentiate between the two classes of compounds. The authors suggest that a lack of replicate hybridizations resulted in low reproducibility of the gene responses. To surmount this challenge, they generated a single gene-expression profile from 13 separate hybridizations using RNA isolated independently from cisplatin-treated HepG2 cultures, producing a set of genes that demonstrated altered expression with relatively high reproducibility. This highlights the need for multiple biologic and hybridization replicates to provide statistical confidence in the gene expression profiles.

We have also verified the hypothesis that chemical-specific patterns of expression can be revealed using microarray analysis (66). We analyzed patterns of gene expression in liver tissue isolated from chemically exposed Sprague-Dawley rats using cDNA microarrays. Biologic and hybridization replicates generated statistically significant information pertaining to the expression pattern of all 1,700 genes on the DNA chip. Clustering analysis and statistical correlation revealed that gene expression profiles produced in animals treated with different agents from a common class of compounds (peroxisome proliferators) were similar. As expected, a distinct gene expression profile was produced using each compound from a completely different class (barbiturates). Not only did our study discriminate among different classes of toxicants on the basis of analysis of validated, statistically significant gene expression profiles, but also corroborated past findings regarding the molecular mechanisms of action for peroxisome proliferators and phenobarbital.

DNA microarrays will be instrumental in revolutionizing the fields of toxicologic and pharmacologic research. However, data generated from microarray analysis cannot be appreciated fully without computational tools that can efficiently handle and analyze large volumes of data. Thus, the field of bioinformatics has emerged and is actively addressing issues such as detecting microarray targets, calculating intensity from raw image scans, transforming data sets into more manageable forms, performing analyses that extract associations from gene expression levels, and facilitating the emergence of new testable hypotheses.

**Data Analysis: Bioinformatics**

The applications of database systems, computer programs, and information technology have revolutionized the way in which biologic scientists manage, analyze, and share information and data. The National Center for Biotechnology Information (www.ncbi.nlm.nih.gov) (67), a division of the National Library of Medicine at the National Institutes of Health, maintains many databases and biologic resources that contain information and data for the scientific community engaged in biomedical research. In addition, several investigative programs in computational biology have been established to leverage theoretical, analytical, and applied computer science research toward addressing complex mathematical and statistical problems inherent in modern day molecular biology. In essence,
the Human Genome Sequencing Project has ascended as a quintessential large-scale, multidisciplinary effort that has benefited immensely from the analysis, interchange, and comparison of DNA sequencing data and results (68). With the advent of microarray technology to monitor genomewide expression of thousands of genes, scientists have adopted an operations research approach to analyze and manage gene expression data in a production-like setting. The essential component of
Gene expression data is the nonlinear nature of data processing. Inherent in microarray environmental health perspectives • progression in the application of DNA microarrays

The microarray data analysis schema comprises three main processes: image analysis and data acquisition, data processing, and multivariate analysis (Figure 2).

The initial stage of microarray gene expression analysis is image analysis and data acquisition. Typically a microarray is scanned in a chip reader using lasers to excite two different fluorophores that have been incorporated into the DNA; simultaneously the chip reader collects the separate emission signals with dual photomultiplier tubes (PMT). The output current of each PMT is then digitized to form a raw 16-bit gray-scale image (8,70). The most current scanning technology performs imaging in the range of 5-10 μm resolution; however, new developments in scanning technology should improve imaging resolution and ultimately enhance the signal-to-noise and dynamic range of detection.

Changes in gene expression are quantified from raw images using image analysis algorithms and data acquisition procedures to detect target regions, compute feature pixel intensity values, subtract background intensity, and compare the pixel intensity values of DNA from a treated sample to that of a control or untreated sample. Normalization or calibration steps are regularly performed in microarray data analysis to accurately adjust and correct the variability in intensity values extracted from two separate samples according to a standard set of control genes or against all the targets represented on the DNA chip. A final mean ratio pixel intensity value is calculated for each gene target and processed into a standard 24-bit pseudo-color composite image for visualization of the gene expression profile.

Given the high “gene feature” density of microarray chips, one may store pixel intensity values, calculated ratio, and statistics for each gene as records in a tab-delimited text file for subsequent data management and analysis. For instance, the microarray database ArrayDB was designed as a relational database management system to store, analyze, and associate gene expression data with information from remote biologic resources (71). More recently, a comprehensive database has been developed to correlate gene expression patterns with information about pharmacologic compounds (63).

The next stage in microarray data analysis is data processing. Inherent in microarray gene expression data is the nonlinear nature of the data. Low-quality or unreliable data points originate from fluorescent signals below threshold detection levels or anomalies on the microarray chip, or are the result of missing DNA features. Typical microarray data analysis software includes functions to log-transform gene expression data, which essentially stabilizes the variance in the data. In addition, cutoff values may be applied to the pixel size of DNA features as well as the intensity range of the values to refine the data set for further analysis. These processing features filter microarray gene expression data and generally improve the reliability of results.

Determination of differentially expressed genes is the staple of this processing stage, which adds another level of complexity to the analysis of microarray gene expression data, but promises to play a pivotal role in fundamental microarray gene expression pattern recognition exercises. At the onset of the microarray era, simplistic statistical methods were used to facilitate the interpretation of complex and large-scale microarray gene expression results. For example, a set number of standard deviations above and below the mean intensity values typically identified transcriptional changes in biologically relevant genes in most published microarray-related studies. In newer, more sophisticated methods, calculated confidence intervals are used to determine significantly changed genes from a ratio distribution of gene expression data (72). In this application, differentially expressed genes are identified on the basis of confidence intervals determined from a probability distribution of the ratios of intensity values, with significantly changed genes being those that fall outside of the confidence limits. Ratio values close to 1.0 represent genes that are expressed similarly in the two samples being compared. More importantly, differentially expressed genes can be validated by comparing replicate measurements of microarrays and performing subsequent biologic assays to confirm the biologic significance of altered gene expression (73,74). Using this method, the probability that a “validated” gene occurred by chance at a given confidence level can be determined using a binomial distribution model (75).

At this time, these methods do not control for false positives and false negatives, low fold changes, or variation of data across independent replicate experiments, nor do they account for genes exhibiting a bias when detected from microarray fluor-reverse experiments. Fortunately, surveying changes in microarray gene expression data has recently attracted significant attention and gained interest from practitioners in statistics, mathematics, and applied physics communities as well as expert bioinformatic companies committed to applying rigorous and innovative computational error models to measurements of microarray gene expression data. For instance, pioneering analysis of microarray gene expression data is currently underway using various mixed linear models and general analysis of variance models to assess gene statistical significance from microarray gene expression data (76,77).

Similarly, a combination of robust additive and multiplicative error models is used to account for the uncertainty and variation in measuring gene expression data as well as assign confidence and probability values, weight, and error bars to define statistical significance of individual genes (78). Though these new and practical approaches emphatically meet the computational needs of the microarray community, it remains to be seen how mainstream and conventional these procedures will be in the next generation of commonplace microarray gene expression analysis tools.

The final stage in microarray gene expression analysis is multivariate analysis. In this step, higher-order computational tools and procedures assist with interpretation and visualization of complex multivariate microarray gene expression data. Although there is no definitive rule about which form of multivariate analysis to use on microarray gene expression data, it is critical to understand fully the implicit characteristics of each algorithm so as to exploit inherent advantages of each to gain the most intriguing and interpretive value of analyzed results.

For example, hierarchical clustering methods are essentially iterative processes to organize objects into groups with similar attributes based on resemblance, proximity, and similarity or dissimilarity measurements. The four basic steps in conducting a cluster analysis on microarray gene expression data are data collection and selection of variables (genes and experiments) for analysis; generation of a resemblance matrix using a particular mathematical formula to measure the extent of similarity; selection and execution of an iterative clustering method to produce a phylogenetic tree diagram (dendrogram) with the branches representing the amount of similarity between clusters; and determination of where to “cut” the tree into a select number of nodes. Recently, robust computational tools and an interactive graphical interface have been developed to facilitate the compilation and visualization of clustered microarray gene expression data (33). Although cluster analysis is a classic and rather simple partitioning technique, its application is relatively new to microarray data analysis and is not supported by any comprehensive body of statistical literature.

Self-organizing maps (SOMs) are another useful algorithm used to interpret and visualize...
large high-dimensional data sets in two-dimensional space. Basically, 50 M's are an array of interconnected cells that become refined to various input signal patterns or classes of patterns in an orderly manner. These artificial neural networks undergo a competitive and unsupervised iterative learning process. When microarray gene expression data are used to train SOM's, structure is imposed on the data, with neighboring nodes defining related clusters. In essence, the SOM constitutes a new paradigm in artificial intelligence and cognitive learning and has recently been implemented into a computer program as a genetic neural network model for interpreting and displaying patterns of gene expression data (79).

Finally, discriminant analysis is commonly used to determine which variables best distinguish between two or more groups. The general concept underlying the use of discriminant analysis for microarray data is to determine whether classes of gene expression measurements differ significantly with regard to distinct and characteristic gene expression profiles. Microarray data are used to assess and rank the importance of particular genes for discrete sample classification. When identified, hallmark gene markers are used to predict the classifications of test microarray treatments as well as apply robust pattern recognition procedures to large volumes of complex gene expression data. Computationally, discriminant analysis is similar to analysis of variance; however, a novel approach that combines a genetic algorithm and the k-nearest neighbor method has been described to positively identify genes that discriminate between classes of tumor cells (80) and toxicant treatments (66).

DNA microarrays have revolutionized the basic approach to research. This technique has generated collaborations among various scientific groups (pathologists, molecular and cellular biologists, toxicologists, and the like). Additionally, microarrays have brought together experts in engineering, bioinformatics, and statistics. The large and complex nature of these studies will facilitate the partnering of different research institutions in both the public and private sectors.

The utility of DNA microarrays is derived from the promise that this technology will rapidly advance our understanding of fundamental biologic questions. Microarrays have the potential to markedly increase our understanding of not only the process of disease but also the interactions between biologic organisms and their environment.

REFERENCES AND NOTES

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470 (1995).

888

VOLUME 109 | NUMBER 9 | September 2001 • Environmental Health Perspectives
al. The transcriptional program in the response of human fibroblasts to serum. Science 283:383-387 (1999).

45. Walker J, Rigley K. Gene expression profiling in human peripheral blood mononuclear cells using high-density filter-based cDNA microarrays. J Immunol Methods 239:167-179 (2000).

46. Guo X, Liao K. Analysis of gene expression profile during ageing of human fibroblasts. Nature 345:458-460 (1990).

47. Harley CB, Futcher AB, Greider CW. Telomeres shorten during ageing of human fibroblasts. Nature 345:458-460 (1990).

48. Greider CW. Telomere length regulation. Annu Rev Biochem 65:337-365 (1996).

49. Kawasaki DA, DelRisi J, M Cusker J, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 94:13057-13062 (1997).

50. Hoeffler WK. Telomerase expression restores dermal integrity to in vitro-aged fibroblasts in a reconstituted skin model. Exp Cell Res 258:270-278 (2000).

51. Fank W, Wang CK, Shelton DN, Harley CB, Pagon GD, Hoeffer WK. Telomerase expression restores dermal integrity to in vitro-aged fibroblasts in a reconstituted skin model. Exp Cell Res 258:270-278 (2000).

52. Gu X, Liao K. Analysis of gene expression profile during ageing of human fibroblasts. Nature 345:458-460 (1990).

53. Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. Unpublished data.
125. Bishay K, Ory K, Lebeau J, Levalois C, Olivier MF, Amundson SA, Do KT, Shahab S, Bittner M, Meltzer P, Cossman J, Annunziata CM, Barash S, Staudt L, Dillon P, Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack 880
124. acid target gene in acute promyelocytic leukemia. Reviews • Lobenhofer et al.

123. differentiation of human K562 erythroleukemic cells by cultured with the pleiotrophic 5-lipoxygenase inhibitor, danamycin, an inhibitor of the hsp90 molecular chaper- 

122. expression profile of a pancreatic cancer cell line. Genes Chromosomes Cancer 26:70–79 (1999).

121. diacylglycerol, and reg I beta mRNAs are concomitantly up-regulated during human colorectal carcinogenesis. Int J Cancer 81:688–694 (1999).

120. show molecular profiling of tumor specimens. Nat Med 5:910–917 (1999).

119. for genomic imbalances. Genes Chromosomes Cancer 32:110–127 (2000).

118. of gene expression patterns characteristic of human hepatocellular carcinoma. Genomics 50:331–340 (1999).

117. of the PAX3-FKHR fusion oncogene. Proc Natl Acad Sci USA 96:7382–7387 (1999).

116. identification of potential tumor markers and antigens by database mining and rapid expression screening. Genome Res 10:1393–1402 (2000).

115. Tsou AP, Wu KM, Tsen TY, Chi CW, Chiu JH, Lui WY, Hu 81. glioblastoma cell lines. Oncogene 18:2711–2717 (1999).

113. expression of cancer-related genes in human cells irradiation of a lymphoid infiltration cell line. Oncogene 9:1339–1346 (1999).

112. Huang P, Feng L, Oldham EA, Keating MJ, Plunkett W. Expression of cell lineage. Blood 94:411–416 (1999).

111. spike 2:1832–1839 (1999). R. J. C. Cormick DL. Expression of cancer-related genes in human cells exposed to 60 Hz magnetic fields. Radiat Res 135:673–684 (1999).

110. comprehensive analysis of prostate gene expression: convergence of expressed sequence tag databases, transcript profiling and proteomics. Electrophoresis 20:1832–1839 (1999).

90. reviews • Lobenhofer et al.

11. comprehensive analysis of tumor and normal tissue profiles by restriction landmark genomic scanning (RLGS-M)-based comparative genomic hybridization. Genes Chromosomes Cancer 32:110–127 (2000).

10. Rhee CH, Hess KR, Caskey LS, Wang R, Bruner JC, Iovanna JL. Molecular cloning, sequencing and expression of the hsp90 molecular chaperone gene. J Biol Chem 269:3800–3807 (1994).

9. Genes Chromosomes Cancer 26:70–79 (1999).

8. hybridization. Biotechniques 29:530–536 (2000).

7. expression profiling and proteomics. Electrophoresis 21:1039–1045 (2000).

6. Alond G, Knonagel M, Muramatsu M, Plass C, Hayashizaki Y. Comparative genomic hybridization, spectral karyotyping, and microarray analysis in the identification of subtype-spe- cific patterns of genomic alterations in rhabdomyosarco- ma. Neoplasia 2:262–275 (1999).

5. Clark EA, Golub TR, Lander ES, Hynes RO. Genomic analysis of metastasis reveals an essential role for RhoC. Nature 496:532–535 (2000).

4. Khan J, Simon R, Bittner M, Chen, Leighton SB, Pohida T, Smith PD, JIang Y, Gooden GC, Trent J, et al. Gene expression profiling of alveolar rhabdomyosarcoma with CDNA microarrays. Cancer Res 58:5009–5013 (1998).

3. Summers GP, Nachman JB, Shen SS, Fiser F, et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotechnol 17:555–561 (1999).

2. A, et al. Use of representational difference analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745–6750 (1999).

1. of metastasis reveals an essential role for RhoC. Nature 496:532–535 (2000).

It seems like the content you provided is a collection of scientific references and papers. If you need assistance with a specific question or topic related to these references, please let me know! I'm here to help you understand and analyze the material.
