Modern microarray-based functional genomics holds great promise for revealing novel molecular and cellular mechanisms of disease. First introduced commercially in 1996, microarrays have been used widely to monitor the expression of thousands of genes in biological samples, as described in the following paragraphs. Other microarray-based genomic applications are also in development, including comparative genomic hybridization, on-chip sequencing, and novel drug discovery. For example, DNA array-based comparative genomic hybridization identifies chromosomal gains and losses with greatly improved resolution compared with conventional methods that use metaphase chromosomes as hybridization targets. \(^1\) This increase in resolution will continue to improve as the technology advances. Moreover, microarrays provide a better platform for automation than is possible with standard metaphase techniques. Where genetic mutations and aberrations are already well characterized, microarrays can be customized to be effectively used as a diagnostic and prognostic tool. \(^2,3\) In the field of drug discovery, microarrays have the potential to dramatically enhance progress, being used at all stages from target discovery (through validation of new molecular targets and understanding modes of action) to predicting patient response. \(^4\)

These devices are beginning to revolutionize how scientists explore the operation of normal cells in the body and the molecular aberrations that underlie medical disorders. DNA microarrays, which are based on well-established principles of nucleic acid hybridization, simultaneously interrogate thousands of genes. \(^5-7\) The actual mechanics of data capture from raw material are ever-improving and well documented, and it is the analysis and discovery of meaningful gene expression patterns within these data to which we now must turn our attention.

Analytical approaches to gene expression analysis using a cancer classification model are illustrated in the recent article by Pomeroy et al. \(^8\) Several important clinical questions were answered via the application of microarray technology and emerging data analysis techniques to pediatric brain tumors. \(^8\) Using microarrays that monitor the expression of more than 6800 genes, we endeavored to definitively differentiate a group of embryonal tumors whose diagnosis on the basis of morphologic features remains controversial and to predict outcome in the most common of these tumors, medulloblastoma, for which patient response to treatment is unpredictable.

There are 2 general approaches to data analysis: supervised and unsupervised. Unsupervised methods are applied to the entire gene expression data set without any previous knowledge of sample classification, allowing an impartial assessment of the underlying features within a data set. Two examples of unsupervised methods are principal component analysis and self-organizing maps (SOMs). Principal component analysis allowed us to differentiate at a molecular level between the different
brain tumor types and normal cerebellum (Figure 1). The marker genes responsible for this distinction supported the conclusion that medulloblastomas are derived from cerebellar granule cell precursors and that they are molecularly distinct from supratentorial primitive neuroectodermal tumors. This argues against the hypothesis that medulloblastomas are a subset of primitive neuroectodermal tumors, differing only in their location in the cerebellum. Self-organizing maps are ideally suited for exploratory data analysis in the generally large and complex data sets generated in the study of a particular disease, in our case brain tumors. Using SOMs, we identified 2 distinct biological subtypes of medulloblastomas with low and high ribosomal protein expression (Figure 2). Electron microscopy subsequently confirmed that these differences in ribosomal gene expression were reflected at a cellular level by differences in ribosome biogenesis. Although this was not an expected result, it provided us with an interesting therapeutic target. Sirolimus and its analogues are currently under clinical investigation in tumors reliant on the PI3K signaling pathway and ribosome biogenesis.

This approach, although useful in its ability to pull out prominent structure (eg, medulloblastoma vs primitive neuroectodermal tumors) in a data set, may miss more subtle distinctions. We found this to be true for outcome prediction. Neither principal component analysis nor SOMs identified prognostically significant subgroups of medulloblastomas, so we turned to supervised analysis. Expression profiles were obtained from 60 children with medulloblastomas who received similar treatment and whose outcome was known. Supervised methods were used to “learn” the distinction between survivors and patients who failed treatment (Figure 3). Using take-one-out cross-validation, gene expression patterns predict survival with substantially more accuracy than current clinical risk criteria. Several supervised analysis methods showed a similar degree of accuracy, including k-nearest neighbor, support vector machines, and structural pattern localization analysis by sequential histograms.

Supervised methods were also used to successfully classify classic and desmoplastic medulloblastomas (histologically confirmed by a single neuropathologist). These algorithms allowed us not only to classify tumors and predict outcome but also to discover previously unknown relationships between coordinate gene expression and tumor characteristics. For example, we demonstrated that the genes encoding sonic hedgehog (shh)–related proteins are highly expressed in desmoplastic medulloblastomas, suggesting that they arise as a consequence of dysregulated shh signaling. Thus, microarray analysis can identify gene expression profiles that signify an activated regulatory pathway or interacting molecular processes leading to a known cellular response.

There are, of course, limitations to any approach that involves the generation of such a large amount of data for each of a relatively small group of samples. One of the most significant risks is finding statistically significant associations by chance. Consequently, identification of gene expression patterns that may underlie the pathogenesis of brain tumors requires validation. Validation of the expression of single genes can be done using well-established techniques such as Northern or Western blotting, as well as immunohistochemistry or in situ hybridization. Hypotheses that arise from the interpretation of significant patterns of gene expression can be tested in a variety of ways. For example, we used electron microscopy to demonstrate that tumors with increased coordinate expression of ribosomal proteins have high numbers of free ribosomes. Our gene expression–based outcome predictions must be validated in an independent, prospective cohort of

![Figure 1](http://archneur.jamanetwork.com/pdfaccess.ashx?url=/data/journals/neur/13519/)
Figure 2. Self-organizing maps were used to discover 2 predominant classes of medulloblastoma: class 0 (high ribosome content) and class 1 (low ribosome content). The top 50 genes for each class are shown. Each column represents an individual sample and each row represents a single gene. Relative gene expression is depicted by red when high and blue when low. mRNA indicates messenger RNA.
Figure 3. Top 50 genes associated with survival (A) and treatment failure (B). Each column represents an individual sample and each row represents a single gene. Relative gene expression is depicted by red when high and blue when low. mRNA indicates messenger RNA.
Principal Component Analysis

A simple multidimensional scaling of the data set was obtained by plotting the top principal components (combinations of genes) that account for a significant fraction of the variance in scatterplots. To study the natural clustering of the brain tumor samples, we initially considered the subset of genes with the highest variation across samples (Figure 1A). In this case, the top 3 principal components account for approximately 43% of the variance of the marker genes. We then plotted principal components based on the top 10 marker genes associated with each tumor, selected by the signal-to-noise statistic. The top 3 principal components of this data set accounted for approximately 61% of the variance, and the degree of separation and clustering of tumor types was significantly improved over that obtained by the analysis of genes with highest variation across samples (Figure 1B). These calculations were performed using Mathsoft software available on the Internet at http://www.mathsoft.com.

Self-organizing Maps

We performed SOMs using the Gene-Cluster software package available on the Internet at http://www.genome.wi.mit.edu. As an exploratory data analysis method, SOMs identify groups of samples with common gene expression patterns within a large heterogeneous sample set. To calculate SOMs, one initially randomly selects and maps a grid of nodes onto the tumor sample set. Through an iterative series of calculations testing the similarity of gene expression between samples, the geometry of the nodes is adjusted to reflect the data structure. If the number of nodes exceeds the number of “natural” clusters in the sample set, then the nodes will converge to reflect the natural clustering. In our case, applying this unsupervised approach to the medulloblastoma data set led to the discovery that 2 is the optimum number of groups identifiable by SOMs (Figure 2).

SUPERVISED

To build supervised classifiers, we defined target classes based on morphologic features, tumor class, or treatment outcome. The method is illustrated by our analysis of treatment outcome (Figure 3). In this case, we created 2 classes of patients based on clinical outcome. Gene expression profiles from patients who died of progressive disease due to treatment failure were compared with expression profiles of patients who were still alive at the end of the study and who had been followed for at least 1 year after cessation of therapy. Genes correlated with the 2 outcome classes were identified by sorting all of the genes on the array according to the signal-to-noise statistic. We built a sample classifier in cross-validation by removing 1 sample and then using the rest as a training set and then repeating this procedure until all samples were tested as “unknowns.” Several models were built using different numbers of marker genes, and the final chosen model was the one that minimized the total error (number of samples that were misclassified) in cross-validation. For this, k-nearest neighbor, weighted voting, and support vector machine algorithms were used (Figure 4).
known sample by calculating the distance of that sample from those surrounding it in gene expression space (class-specific marker genes, ie, those associated with survival or treatment failure). The unknown sample was predicted to be in one of the other outcome class based on the similarity of gene expression with that of most of the k-nearest (neighbor) samples (Figure 4A). Marker genes were chosen from those highly correlating with the predetermined classes using the signal-to-noise statistic.

Weighted Voting

The weighted voting algorithm\(^{10}\) makes a weighted linear combination of relevant “marker” or “informative” genes obtained in the training set to provide a classification scheme for new samples. The selection of marker genes for each outcome class was determined by computing the signal-to-noise statistic of each gene for the predefined classes. The algorithm determined the decision boundary (halfway) between the outcome class means for each gene (Figure 4B). To predict the class of the unknown sample, each gene in the sample expression profile casts a vote, and the unknown sample is assigned to the class with the most positive voting genes. The distance of that sample from the decision boundary determines the weight that each gene carries in this voting process. The closer each gene of the unknown sample is to the decision boundary for a particular gene, the less weight that gene carries toward the assignment of the sample to an outcome class. Confidence in the class prediction of the unknown sample was determined by the size of the voting margin responsible for putting the sample in one class vs the other.

Support Vector Machines

The basic idea behind support vector machines is to construct an optimal class-separating hyperplane (decision boundary) by mapping the gene expression data to a high-dimensional space.\(^{12,13}\) Linear separation in this higher dimensional space corresponds to a non-linear decision boundary separating the 2 outcome classes (Figure 4C). This allowed us to more optimally separate outcome classes than with weighted voting, where the decision boundary is linear. As with weighted voting, samples are assigned to a class by their position in relation to the decision boundary, and, again, confidence of classification is dependent on the relative distance of the sample from that boundary into a particular class.

Relevance to the Study of Neuroscience and the Practice of Neurology

The use of microarrays provides a springboard from which we can start to examine the cellular pathogenesis underlying neurological disease and perhaps narrow down the search to a manageable group of therapeutic targets. Examples of this can be seen in neurological disorders such as multiple sclerosis, Huntington disease, Parkinson disease, and Alzheimer disease.\(^{14-17}\) Applying microarray technology to a transgenic animal model of Huntington disease resulted in the finding that genes encoding certain neurotransmitters, calcium and retinoid signaling pathway components, were down-regulated, whereas those encoding inflammatory components were up-regulated.\(^{16}\) These findings were unexpected consequences of the mutant huntington protein and could only have been discovered on this scale by microarray analysis.

Multiple sclerosis is a complex disorder with multiple clinical subtypes that cannot be diagnosed by clinical criteria at initial presentation.\(^{18}\) Immunomodulatory treatment has proved to be relatively successful in relapsing-remitting disease, but it is not as useful in primary or secondary progressive disease.\(^{18,19}\) To date, there are no in vivo markers that allow specific direction of treatment. Microarray analysis has begun to dissect the molecular heterogeneity of multiple sclerosis, identifying genes related to cell metabolism, structure, cytokines, and cell adhesion molecules. In addition, a gene not previously associated with multiple sclerosis, encoding the Duffy chemokine receptor, was identified using this technology.\(^{14,17}\) Although these results are preliminary, eventually microarray expression analysis may lead to the identification of markers that are detectable in living patients, allowing prognosis to be accurately predicted at the time of initial diagnosis and treatment to be tailored accordingly. An even greater future challenge is offered by the investigation of psychiatric disorders, which seem to result from the interplay of polygenic and epigenetic factors on multiple brain circuits.\(^{20}\)

Conclusions

The development of array-based DNA mutation screening may, in the future, prove to be beneficial for the identification of an individual’s genetic propensity to acquire disorders such as Alzheimer or Parkinson disease. Consequently, at-risk candidates can be selected for close monitoring, intensive preventive care, and early clinical intervention. Other applications may include screening of patients for gene variants that affect the individual’s response to certain medications, allowing the physician to tailor the best treatment regimen for a given disease in an individual patient.

Useful Web Sites

The following Web sites are useful in the study of microarray-based functional genomics:

- Center for Genome Research: http://www-genome.wi.mit.edu/

Accepted for publication December 11, 2002.

Author contributions: Study concept and design (Dr Pomeroy); acquisition of data (Drs Sturla, Fernandez-Teijeiro, and Pomeroy); analysis and interpretation of data (Drs Sturla and Pomeroy); drafting of the manuscript (Drs Sturla, Fernandez-Teijeiro, and Pomeroy); critical revision of the manuscript for important intellectual content (Dr Pomeroy); statistical expertise (Dr Pomeroy); obtained funding (Dr Pomeroy); administrative, technical, and material support (Dr Sturla); study supervision (Dr Pomeroy).

Corresponding author and reprints: Scott L. Pomeroy, MD, PhD, Division of Neuroscience, Department of Neurology, Children’s Hospi
REFERENCES


Call for Papers

The ARCHIVES Express section in the September 2000 issue. This section will enable the editors to publish highly selected papers within approximately 2 months of acceptance. We will consider only the most significant research, the top 1% of accepted papers, on new important insights into the pathogenesis of disease, brain function, and therapy. We encourage authors to send their most exceptional clinical or basic research, designating in the cover letter a request for expedited ARCHIVES Express review. We look forward to publishing your important new research in this accelerated manner.

Roger N. Rosenberg, MD
Editor