Computational Approaches to Analysis of DNA Microarray Data

J. Quackenbush
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and
Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

Summary

Objectives: To review the current state of the art in computational methods for the analysis of DNA microarray data.

Methods: The review considers methods of microarray data collection, transformation and representation, comparisons and predictions of gene expression from the data, their mechanistic analysis, related systems biology, and the application of clustering techniques.

Results: Functional genomics approaches have greatly increased the rate at which data on biological systems is generated, leading to corresponding challenges in analyzing the data through advanced computational techniques . The paper compares and contrasts the application of computational clustering for discovery, comparison, and prediction of gene expression classes, together with their evaluation and relation to mechanistic analyses of biological systems.

Conclusion: Methods for assaying gene expression levels by DNA microarray experiments produce considerably more data than other techniques, and require a wide variety of computational techniques for identifying patterns of expression that may be biologically significant. These will have to be verified and validated by comparison to results from other methods, integrated with other systems data, and provide the feedback for further experimentation for testing mechanistic or other biological hypotheses.

Haux R, Kulikowski C, editors. IMIA Yearbook of Medical Informatics 2006. Methods Inf Med 2006; 45 Suppl 1: S91-103.

Keywords

DNA microarrays, computational clustering methods, classification of gene expression data, verification and validation of microarray results.

Overview

The genome project has fundamentally changed the way in which we approach questions in biology, but not in the manner that many of us had envisioned. While genome sequences and preliminary gene catalogues have been useful, they have not revolutionized our understanding, for example, of the link between genotype and phenotype or the mechanisms governing an organism's growth and development. It has been the technology that the genome project has enabled, rather than the data it has produced, that has the most profound impact on our conduct of biological research. In particular, functional genomics approaches, such as DNA microarrays, proteomics, and metabolomics have greatly increased the rate at which we can generate data on biological systems, allowing us for the first time to begin to observe on a molecular level the holistic response of an organism to a particular stimulus.

As microarrays and other technologies have become more commonplace, the challenges associated with collecting, managing, and analyzing the data from each experiment have increased substantially. Increasingly robust laboratory protocols, falling prices for commercial platforms, and an improved understanding of the intricacies of experimental design all have combined to drive the field to more complex experiments, generating enormous amounts of data. Just a few years ago, microarray studies typically included on the order of 10 hybridization assays; now, studies tend to have 100 or more such assays. The goal of this chapter is to present an overview of some of the issues associated with analyzing such data to extract meaningful biological results from the data.

Microarray Data Collection, Transformation, and Representation

A DNA microarray experiment begins with the choice of an appropriate experimental platform and this dictates a good deal of how an experiment is designed and analyzed. One key element in any microarray analysis is understanding which genes are represented by the individual probes. With "complete" reference genomes, one would imagine that this is a solved problem, but in most instances, the genome sequence and its annotation is still evolving and we do not yet have a comprehensive catalogue of the genes and their variants. However, this and issues of data management are beyond the scope of what we will cover in this chapter. Instead we will assume that we have selected an array platform and extracted expression data from a series of hybridization assays representing a group of biologically interesting samples.

The starting point for understanding how one uses DNA microarrays for

classification is to understand how the data are collected and represented. The underlying technology is relatively straightforward. Gene-specific probes, representing thousands of individual genes, are arrayed on an inert substrate and used to assay levels of gene expression in a target biological sample. RNA is extracted from tissues of interest, labeled with a detectable marker (typically a fluorescent dye), and allowed to hybridize to the arrays with individual messages hybridizing to their complementary gene-specific probes on the array. Stoichiometry dictates that the relative quantity of nucleic acid bound to any probe should be a function of concentration. A more intense signal is caused by a higher degree of hybridization which, in turn, implies higher expression levels. Following hybridization and washing, the arrays are imaged using a confocal laser scanner and the relative fluorescence intensity for each gene-specific probe is extracted as a measure of the expression level for that gene. The actual value reported depends on the microarray technology platform used and the experimental design. For Affymetrix GeneChips™, where each sample is hybridized to an individual array, expression for each gene is measured as an "Average Difference" that represents an estimated expression level, less nonspecific background. For two-color arrays, such as cDNA and Agilent arrays, assays typically compare paired samples and report expression as the logarithm of the ratio of a query sample to a control (the log-ratio). Regardless of the approach or technology, the fundamental data used in all subsequent analyses are the expression measures for each gene in each experiment.

This expression data is typically represented as an "expression matrix" in which each row represents a particular gene and each column represents a specific biological sample. In this representation, each row is a "gene expression vector" where the individual entries are its expression levels in the samples assayed and each column is a "sample expression vector" that records the expression of all genes in that sample. While we will focus on classification using DNA microarray data, it should be noted that any data which can be placed into this "genes by samples" expression matrix format (for example. "proteins by samples") can be analyzed using exactly the same techniques.

Following collection, the data are generally normalized to facilitate comparison between individual hybridization assays to compensate for differences in labeling, hybridization, and detection efficiencies. There are a number of approaches to data normalization, and again the approaches used depend on platform and the assumptions made regarding the biases in the data [1, 2, 3, 4, 5]. Further, a number of filtering transformations are often applied to the data using a variety of statistical approaches that, for example, eliminate genes that have minimal variance across the collection of samples or those that fail to provide data in a majority of the experiments. The value of these filtering transformations is that they reduce the complexity of the dataset by eliminating those genes that are not likely to contribute to either class discovery or classification.

It is important to note, however, that the choice of normalization and filtering transformations can have a profound effect on the results that are obtained [6]. Normalization adjusts the fluorescence intensities on each array and can therefore change the relative difference observed between samples – the fold change. While some normalization is generally necessary to compensate for systematic errors that are introduced during the measurement process, over-normalizing the data can distort the final results. Similarly, the manner in which the data are filtered can produce very different results. All statistical tests that are applied rely on assumptions regarding the nature of the variance in the measurements. Different statistical tests applied to the very same data set can often produce different (but generally overlapping) sets of significant genes and the appropriate means of dealing with these "high dimensional" datasets in which there are often more measurements (genes) than samples is an area of active research and debate.

DNA microarray experiments can be broadly placed into four primary groups: class discovery, class comparison, mechanistic studies, and class prediction. Each generally has a very different final goal that can affect the experimental design and analysis, although any one experiment may yield novel insights into all four areas and the most useful newly discovered classes are those with a clear mechanism associated with them that can then be used to classify future samples.

Once the data have been collected, normalized, and filtered in some way, the real analysis begins. There are many possible experimental designs and many approaches to data analysis that build on trying to answer the fundamental questions posed in each experiment. Table 1 summarizes the broad classes of experiments that one might perform and some of the software tools that are useful for such analyses; a more detailed discussion of these data mining approaches follows immediately below.

Class Discovery

Class discovery analysis is generally the first step in any genomics experiment

Table 1 A wide range of algorithms have been developed to facilitate analysis of genomic expression datasets. Although most approaches have been applied in the context of gene expression microarray data, the algorithms themselves are generally applicable to any expression-based data.

Application	Algorithm	References
Class Discovery	hierarchical clustering	[7] [8] [9]
	k-means clustering	[10]
	self-organizing maps	[11] [12] [13]
	self-organizing trees	[14]
	Relevance networks	[15]
	force-directed layouts	[16]
	Principal component analysis	[17]
Class Comparison	*t*-test	[65]
	SAM	[22]
	analysis of variance (ANOVA)	[66]
Class Prediction	*k*-nearest neighbors (*k*NN)	[33]
	Weighted voting	[23]
	artificial neural networks (ANNs)	[24] [25]
	discriminant analysis	[26] [27] [28] [29]
	classification and regression trees (CART)	[30]
	support vector machines (SVM)	[31] [32]
Mechanistic Analysis	EASE	[36]
	MAPPFinder	[34]
	GOMiner	[35]
	Cytoscape	[67]
	Boolean Networks	[37] [38] [39]
	Probabilistic Boolean Networks	[40] [41] [42] [43]
	Bayesian Networks	[38] [44] [45] [46] [47]

because it takes an unbiased approach to looking for new groups in the data. For example, one might examine a group of cancer patients to see if their expression profiles allow them to be placed into distinct groups without using any prior knowledge of their disease progression, outcome, or their response to treatment. After finding new groups based on expression profiles, the challenge then becomes finding a link to some clinical or biological factor that can explain the difference.

Class discovery analyses rely on *unsupervised* data analysis or *clustering* methods to explore expression patterns that exist in the data and these are often among the first techniques used in the analysis of any microarray dataset. The question we are asking in a class discovery experiment is "Are there unexpected but biologically interesting patterns that exist in the data?" Unsupervised methods do not use the sample classification as input – they do not take into account, for example, whether the samples come from ALL or AML patients. They simply group samples together based on some measure of similarity between then. Two of the most widely used unsupervised approaches are hierarchical clustering [7, 8, 9] and *k*-means clustering [10].

There are many approaches that have been applied to unsupervised analysis, including self-organizing maps (SOM) [11, 12, 13], self-organizing trees (SOTA) [14], relevance networks [15], force-directed layouts [16], principal component analysis [17], and others. Fundamentally, each of these uses some feature of the data and a rule for determining relationships to group genes (or samples) that share similar patterns of expression. In the context of disease analysis, all of these can be extremely useful for identifying new subclasses in the data – provided that the classes are reproducible and that they can be related to other clinical data. All of these algorithms will divide data into clusters, but whether the clusters are meaningful requires expert input and analysis. Critical assessment of the results is essential. There are anecdotal reports of clusters being found that separate data based on the hospital in which the sample was collected, the technician who ran the microarray assay, or the day of the week on which the array was run. Clearly arrays can be very sensitive; one just has to work to minimize unnecessary variability and then to filter the biological signal from the noise. However, recent reports have suggested that adherence to good, standard laboratory practices and careful analysis of data can lead to high quality, reproducible results where the biology of the system under study drives the expression profiles that are observed [18, 19, 20, 21]. While clustering approaches can be useful in such studies, in general the classification of new samples based on their expression profiles generally relies on the application of class comparison methods followed by the development of robust and reliable classification algorithms.

Class Comparison

Class comparison experiments are focused on comparing different phenotypic groups (treated and control groups, disease tissue versus normal, or two compounds affecting the same cell

type but through different mechanisms) in order to discover the genes and their expression patterns that best distinguish the groups. The starting point in such an experiment is the assumption that one knows the classes that are represented in the data, a logical approach to data analysis is to use the information about the various classes in a *supervised* fashion to identify those genes (or proteins or metabolites) that can be used to distinguish the various groups. One starts by assigning samples to particular biological classes based on some objective criteria. For example, the data may represent samples treated with two different drugs known to elicit different responses or disease and normal tissues. The first question to be asked is: "Which genes best distinguish the various classes in the data?" The goal at this stage is to find those genes that are most informative for distinguishing the samples based on class.

Fortunately, there are a wide variety of statistical tools that can be brought to bear on this question, including t-tests (for two classes) and analysis of variance (ANOVA; for three or more classes) that assign *p*-values to genes based on their ability to distinguish between groups. One concern with these statistical approaches is the problem of multiple testing. Simply put, in an array with 10,000 genes, applying a 95% confidence limit on gene selection ($p = 0.05$) means that, by chance, one would expect to find 500 genes as significant. Clearly, we need to be more stringent in our gene selection to avoid potential problems with these. However, the important thing to remember is that what these methods provide are a means for prioritizing genes for further analysis. It should be noted that there are other widely used approaches, such as Significance Analysis of Microarrays (SAM) [22] which uses an adjusted t-statistic (or F-statistic), modified to correct for overestimates arising from small values in the denominator, along with permutation testing to estimate the False Discovery Rate (FDR) in any selected significant gene set.

Ultimately, the result of such an analysis is a collection of genes that are deemed significant for distinguishing the biological groups being compared in the analysis. The challenge at this point is generally to place these genes into a biological context. This in many ways is the key unsolved problem for functional genomics: if we knew what each of the individual genes, proteins, and metabolites did, as well as how genetic variation and other factors play a role in producing a particular outcome, there would be no need to use functional genomic approaches. Rather, one could simply focus on those key elements that are causally involved in any process and use those to determine whether a particular compound is likely to produce a specific response.

Having selected a set of genes whose expression patterns are useful for distinguishing two or more classes of samples, such as individuals who develop a particular disease and matched controls who do not, one can either use them as a starting point for mechanistic studies or attempt to classify new samples based on their expression profiles.

Class Prediction

Class prediction experiments attempt to go beyond the simple clustering approaches used in class discovery experiments to use catalogued expression profiles as a means of predicting to which group a new sample belongs based on its unique profile. The question we ask in such an experiment is "Can I find a particular pattern of expression and an appropriate mathematical rule that allows me to predict what group my sample belongs to?" Typically one starts out with a well characterized set of samples and their associated phenotypes and through a careful comparison of the expression profiles finds genes whose patterns of expression can be used to distinguish the various phenotypic groups under analysis. Class prediction approaches then attempt to use such a set of "significant" genes to develop a mathematical rule (or computational algorithm) that can use the expression profiling data and take any one sample and assign it to its particular group. The goal, however, is not to merely separate the samples, but to create a rule (or algorithm) that can serve to predict the phenotype based on expression profiling data alone.

When developing a classification approach, the mathematical rules for analyzing new samples are encoded in a *classification algorithm*, and there are a wide range of algorithms that have been used for this purpose, including weighted voting [23], artificial neural networks (ANNs) [24, 25], discriminant analysis [26, 27, 28, 29], classification and regression trees (CART) [30], support vector machines (SVM) [31, 32], and *k*-nearest neighbors (kNN) [33], as well as a host of others. Essentially each of these uses an original set of samples, or training set, to develop a rule that takes a new test sample from a test set and uses its expression vector sample, trimmed to a previously identified set of classification genes, to place this test sample into the context of the original sample set, thus identifying its class.

There is great interest in classification approaches and one example is in its application to toxicology. While current toxicological assays rely on large

exposures and analysis of specific tissues, one could envision developing a large enough database of a broad enough array of compounds that signatures indicative of specific responses could be identified at lower doses and even, potentially in tissue culture models, providing a testable starting hypothesis for target tissues or putative modes of action for a particular toxic compound. This could have wide-ranging implications for toxicological screening identifying potential unwanted side effects at an early stage in drug development, suggesting new potential uses for failed compounds (as some toxic effects can have therapeutic benefits – for example, in cancer treatment), and identifying environmental compounds that may have some toxic properties so that these can be further evaluated and verified.

Mechanistic Analysis

While class prediction analysis may tell us what group a particular sample belongs to, it does not necessarily shed light on the mechanism underlying a particular response. Moving from predictive signatures to mechanistic understanding often relies on additional work using standard methods to translate functional genomics-based hypotheses to validated findings. Bioinformatics often plays a key role in developing those hypotheses with additional data that can be used in its interpretation, including gene ontology terms (which assign gene products – proteins – to one or more molecular functions, biological processes, and cellular locations), databases of known pathways, genetic mapping data, structure activity relationships, dose response curves, phenotypic or clinical information, the genome sequence and its annotation, and the published literature, among others. A number of software tools have been developed to facilitate this analysis, including MAPPFinder [34], GOMiner [35] and EASE [36], although these only provide hints as to possible mechanisms that might be involved in producing any particular expression profile. At present, reconstructing putative networks requires a good deal of user interaction as there is no universal way to connect the expression of genes, proteins, or metabolites to functionally relevant pathways leading to selected outcomes.

Recently, there have been attempts to predict networks based on observed expression profiles and using a range of techniques, including Boolean Networks [37, 38, 39], Probabilistic Boolean Networks [40, 41, 42, 43], and Bayesian Networks [38, 44, 45, 46, 47], with variations on all of these approaches. These models treat individual objects, such as genes or proteins, as "nodes" in a graph, with edges connecting them representing their interactions and a set of rules for each edge that determines the strength of the interaction and whether a particular response will be induced. To date, these approaches have met with some success, but a great deal of work is necessary to convert these models from descriptive to predictive. In metabolic profiling, techniques that use monitoring of metabolic flux and its modeling [48, 49] also hold hope of providing predictive models.

From Genes to Systems

The advent of global functional genomics technologies, and the data they provide, has opened the possibility of creating quantitative, predictive models of biological systems. This approach, dubbed "Systems Biology," attempts to bring together data from many different domains, such as DNA microarray gene expression data and metabolic flux analysis, and to synthesize these to produce a more complete understanding of the biological response of a cell, organ, or individual to a particular stimulus. Ultimately, this systems-level understanding of organismal response and its relationship with the development of a particular phenotype is the goal of functional genomics. However, the best efforts to date have allowed the prediction of "networks" of potentially interacting genes that have little relation to the biochemical or signal transduction pathways we understand mediate cellular response. Attempts to model metabolic flux, even in simpler organisms like yeast and E. coli, can, at best, provide only rough approximations of the real responses and then only under carefully controlled conditions. However, progress in these areas is promising and additional research will continue to advance the field and its applications.

Mechanistic Versus Non-Mechanistic Studies

The most common question in genomic expression studies is whether a technology will allow us to predict potential outcomes from exposure to a particular substance. When addressing this question, classification methods are most commonly applied. However, the genes, proteins, or metabolites that are identified as the most significant for distinguishing classes of treatment are often not easily interpreted causally or mechanistically with respect to the underlying phenotype or mode of action.

Ultimately, finding predictive elements that can be functionally linked to outcome may provide insight into possible therapeutic interventions. However, the failure to provide a biological interpretation does not diminish the potential predictive utility of well-established biomarkers. It should be noted that there are many clinical examples of biomarkers of unknown function, such as PSA or CEA, that are extremely useful as diagnostic or prognostic markers for various diseases. It may be more useful to consider gene lists emerging from class prediction experiments as nothing more that sets of biomarkers with clinical applications; if they have a biological interpretation, this is simply a bonus.

Having laid out the basics of microarray experiments, it is worthwhile to look at some examples of techniques used to analyzing gene expression data.

Expression Vectors

The ultimate goal of a microarray experiment is to compare patterns of expression across multiple samples hybridized to a particular array. One is typically looking for patterns of gene expression that correlate with the biological states of the system being analyzed or searching for genes that have "similar" patterns of expression across multiple samples. For each gene, the process begins by defining an "expression vector" that represents its location in "expression space." In this view of gene expression, each hybridization represents a separate distinct axis in space, and the expression value measured for that gene in that particular hybridization represents its geometric coordinate. For example, for three hybridizations on a two-color array, the $\log_2$(ratio) for a given gene in hybridization 1 is its *x* coordinate, the $\log_2$(ratio) in hybridization 2 is its *y* coordinate, and the $\log_2$(ratio) in hybridization 3 its *z* coordinate. In this way, all of the information about this gene can be represented by a point in *xyz* expression space. A second gene, with nearly the same $\log_2$(ratio) values for each hybridization will be represented by a (spatially) nearby point in expression space; a gene with a very different pattern of expression will be far from our original gene. The generalization to a greater number of hybridizations is straightforward, although harder to draw; the dimensionality of expression space grows to be equal to the number of hybridizations. In this way, expression data can be represented in *m*-dimensional expression space, where *m* is the number of hybridizations, and where each gene expression vector is represented as a single point in that space. It should be noted that one can use a similar approach to representing each hybridization assay using a "sample vector" consisting of the expression values for each gene; these define a "sample space" whose dimension is equal to the number of genes assayed in each array.

Identifying Differentially Expressed Genes – the t-test

A common goal in DNA microarray experiments is to search for genes that distinguish the various biological classes in any experiment. Even if data mining analysis is going to be performed using one or more of the widely-used clustering methods [7, 9, 11], it is still extremely useful to reduce the dataset to those genes that are best distinguish between the sample classes. The earliest microarray papers used a simple "fold change" approach to find differences, using the assumption that changes above some threshold, typically two-fold, were biologically significant. A simple yet more sophisticated and widely-used approach to the two class experiment is to use Student's *t* test to assess whether a gene is differentially expressed between biological conditions. The basis of this test is the *t* statistic, which is an assessment of signal-to-noise ratio for the particular gene in question, comparing its expression measure for the two conditions under study. Consider two conditions, *A* and *B*. If we use X_i^A to denote the $\log_2$(ratio) that we measure in assay *i* in condition *A*, then its average value across N_A measurements is simply

$$\langle X_A \rangle = \frac{1}{N_A} \sum_{i-1}^{N_A} X_{Ai}$$

and the standard deviation of the mean is

$$s_A = \sqrt{\frac{\sum_{i-1}^{N_A} \left[X_{Ai} - \langle X_A \rangle \right]^2}{N_A}}$$

With these definitions, one can define the *t* statistic as

$$t = \frac{\text{signal}}{\text{noise}} = \frac{\text{difference between groups}}{\text{variability of groups}} = \frac{\langle X_A \rangle - \langle X_B \rangle}{\sqrt{s_A^2 + s_B^2}}$$

Clearly, a large value for the *t* statistic indicates that the populations representing measurements of a gene for conditions *A* and *B* are well separated and, consequently, that the gene is differentially expressed between those conditions. The converse is true for a small value of *t*. But what actually is meant by "large" and "small," and can *t* be used to estimate how likely that a gene is differentially expressed between conditions?

There are a number of approaches to addressing these questions. The first is

to use well-established properties of the *t* distribution for normally distributed random variables. This allows for the calculation of a probability, or *p* value that the two distributions of the expression measures for a gene under conditions *A* and *B* overlap for a given value of *t*. Although the *t* statistic is based on the assumption that variability in these measurements follows a normal distribution and there is evidence to suggest that this is not necessarily the case for gene expression, the *t* test is well known to be quite robust to violations of the assumption of normality. The second approach is to use the properties of the expression measures themselves to estimate the significance of a given value of *t* by performing a permutation test. Permutation testing randomly swaps expression level measurements between groups *A* and *B*, up to the total number of unique permutations that can be made. Each time, a value for *t* is calculated. One then asks how often, by chance, a value for *t* occurs that is equal to or more than that measured for the real data, which allows an estimate to be made of the probability that our dataset shows a significant separation between the classes. In biological studies, it is common to use a *p* value cutoff of $p \leq 0.05$, which means that there is a 95% or more chance that a gene's expression levels can distinguish between groups. Although this seems like a reasonable approach, there can be problems with using strict *p* value cutoffs. In most biological experiments, one measures a small number of parameters across a relatively large number of samples. However, in most microarray experiments, one is measuring thousands of parameters (the expression levels of the genes) across a relatively small number of samples, and this can lead to the misidentification of genes being differentially expressed even when they are not—the problem of false positives. As an example of this phenomenon, known as the *multiple testing problem*, consider the case where 10,000 distinct gene-specific probes are on a given array. If a cutoff for differential expression of $p \leq 0.05$ (95% confidence) is used, one would expect, by chance, that 5% of the genes represented on the array, or 500 of them, would be identified as being significant. There are a number of approaches to dealing with this problem, but it remains an area of active research. For more complex experiments with multiple classes, the use of Analysis of Variance (ANOVA) techniques is now standard.

Having identified a significant set of genes that correlate with biological phenomena, there are a variety of approaches that can be used to mine the data, including hierarchical and k-means clustering, Self Organizing Maps, Self Organizing Trees, and others. There area also a range of applications, including the classification of biological samples using a range of computational tools such as Artificial Neural Networks, *k* nearest neighbors, regression analysis, decision trees, support vector machines, and others. Very often the greatest challenge in any of these approaches is the biological interpretation of the data and the validation of the method. Many classification approaches face the problem of overfitting brought on by the relatively large number of genes, the small number of samples, and the unknown biological noise that must be dealt with. Although many questions remain to be answered, this remains an area of active research and one that continues to be exciting and challenging. Finally, although this presentation focuses on microarrays, the same techniques can be used in a wide array of applications in proteomics, metabolomics, and other fields.

Clustering Approaches

A useful first approach to the analysis of microarray data is to use an unsupervised method to explore expression patterns of that exist in the data. Three of the most widely used methods are hierarchical clustering, *k*-means clustering, and self-organizing maps. Although each of these approaches will work with any dataset, in practice they often do not work well for large datasets where many of the genes do not vary between samples. Consequently, it is useful to first apply a statistical filter to the data to exclude genes which simply are not varying between experimental classes. If there are no pre-determined classes in the data, a useful alternative is simply to eliminate those genes that have minimal variance across the collection of samples as those genes are not changing significantly in the dataset and are therefore the least likely to shed any light on subclasses that exist in the sample collection.

Hierarchical Clustering

Hierarchical clustering has become one of the most widely-used techniques for the analysis of gene expression data; it has the advantage that it is simple and the result can be easily visualized [7, 9, 50]. Initially, one starts with *N* clusters, where *N* is the number of genes (or samples) to be in the target dataset. Hierarchical clustering is an agglomerative approach in which single expression profiles are joined to form nodes, which are further joined until the process has been carried to completion, forming a single hierarchical tree. The algorithm proceeds in a straightforward manner:

1. Calculate the pairwise distance matrix for all of the genes to be clustered.

2. Search the distance matrix for the two most similar genes or clusters; initially each cluster consists of a single gene. This is the true first stage in the "clustering" process. If several pairs share the same degree of similarity, a predetermined rule is used to decide between alternatives.
3. The two selected clusters are merged to produce a new cluster that now contains at two or more objects.
4. The distances are calculated between this new cluster and all other clusters. There is no need to calculate *all* distances since only those involving the new cluster have changed.
5. Steps 2-4 are repeated until all objects are in one cluster.

There are a number of variants of hierarchical clustering that reflect different approaches to calculating distances between the newly defined clusters and the other genes or clusters:

- *Single linkage* clustering uses the shortest distance between one cluster and any other,
- *complete linkage* clustering takes the largest distance between any two clusters, and
- *average linkage* clustering uses the average distance between two clusters.

Typically, the relationship between samples is represented using a dendrogram, where branches in the tree are built based on the connections determined between clusters as the algorithm progresses. In order to visualize the relationships between samples, the dendrogram is used to rearrange the rows (or columns as appropriate) in the expression matrix to visualize patterns in the dataset.

Hierarchical clustering is often misused to partition data into some number of clusters without the application of any objective criterion. Fortunately, there are a number of approaches that can be used to identify subgroups in the clustering dendrograms. One method is to simply use the distances calculated in building the clusters as a measure of the connectivity of the individual clusters. As one moves up the dendrogram from the individual elements, the distance between clusters increases. Consequently, as one increases the distance threshold, the effective number of clusters decreases. An alternative approach is to use bootstrapping or jack-knifing techniques to measure the stability of relationships in the dendrogram, using this stability as a measure of the number of clusters represented. In bootstrapping, there are a number of approaches that can be used, but the simplest is to use sampling of the dataset with replacement, each time calculating a new hierarchical clustering dendrogram and simply counting how often each branch in the dendrogram is recovered; a percentage cutoff on the dendrogram sets the number of clusters. In making a bootstrap estimate for gene cluster stability, it is appropriate to resample the collection of biological samples while in estimating the number of clusters in the biological samples, one bootstraps the gene expression vectors. Jack-knifing is similar, but instead of resampling, the appropriate vectors are sequentially left out as new dendrograms are calculated until all vectors have been considered. Once again, the stability of each cluster is estimated based on how often a given relationship in the dendrogram is recovered.

One potential problem with many hierarchical clustering methods is that, as clusters grow in size, the expression vector that represents the cluster when calculating distance may no longer represent any of the genes within the cluster. Consequently, as clustering progresses, the actual expression patterns of the genes themselves become less relevant. Furthermore, if a bad assignment is made early in the process, it cannot be corrected. An alternative, which can avoid these artifacts, is to use a divisive clustering approach, such as *k*-means, to partition data (either genes or samples) into groups having similar expression patterns.

k-means Clustering

If there is advance knowledge regarding the number of clusters that should be represented in the data, *k*-means clustering is a good alternative to hierarchical methods [10, 51]. In *k*-means, objects are partitioned into a fixed number (*k*) of clusters such that the clusters are internally similar but externally dissimilar. No dendrograms are produced, but one could use hierarchical techniques on each of the data partitions after they are constructed. The process involved in *k*-means clustering is conceptually simple, but can be computationally intensive:

1. All initial objects are randomly assigned to one of *k* clusters (where *k* is specified by the user).
2. An average expression vector is then calculated for each cluster and this is used to compute the distances between clusters.
3. Using an iterative method, objects are moved between clusters and intra- and inter-cluster distances are measured with each move. Objects are allowed to remain in the new cluster only if they are closer to it than to their previous cluster.
4. Following each move, the expression vectors for each cluster are recalculated.
5. The shuffling proceeds until moving any more objects would make the clusters more variable, increasing intra-cluster distances and decreasing inter-cluster dissimilarity.

Some implementations of *k*-means clustering allow not only the number of

clusters to be specified, but also seed cases for each cluster. This has the potential to allow one to use prior knowledge of the system to help define the cluster output, such as a typical profile for a few key genes known to distinguish classes of patients. Of course, the "means" in *k*-means refers to the use of a mean expression vector for each emerging cluster. As one might imagine, there are variations that also use other measures, such as *k*-medians clustering.

Self Organizing Maps

A self-organizing map (SOM) is a neural network-based divisive clustering approach [11, 12, 13]. A SOM assigns genes to a series of partitions based on the similarity of their expression vectors to reference vectors that are defined for each partition. It is the process of defining these reference vectors that distinguishes SOMs from *k*-means clustering. Prior to initiating the analysis, the user defines a geometric configuration for the partitions, typically a two-dimensional rectangular or hexagonal grid. Random vectors are generated for each partition, but before genes can be assigned to partitions, the vectors are first "trained" using an iterative process that continues until convergence so that the data are most effectively separated:

1. Random vectors are constructed and assigned to each partition.
2. A gene is picked at random and, using a selected distance metric, the reference vector that is closest to the gene is identified.
3. The reference vector is then adjusted so that it is more similar to the randomly picked gene. The reference vectors that are nearby on the two dimensional grid are also adjusted so that they too are more similar to the randomly selected gene.
4. Steps 2 and 3 are iterated several thousand times, decreasing the amount by which the reference vectors are adjusted and increasing the stringency used to define closeness in each step. As the process continues, the reference vectors converge to fixed values.
5. Finally, the genes are mapped to the relevant partitions depending on the reference vector to which they are most similar.

In choosing the geometric configuration for the clusters, the user is, effectively, specifying the number of partitions into which the data are to be divided. As with *k*-means clustering, the user has to rely on some other sources of information, such as principal component analysis (PCA), to determine the number of clusters that best represents the available data. There are many other approaches to partitioning the data as noted in the class discovery section above.

Beyond Statistical Significance and Clustering

Many analyses of microarray data reach the stage where some collection of genes that share similar patterns of expression has been identified. The challenge at this stage is to attach some biological meaning to the gene sets identified through this process. Some approaches use relationships identified by linking genes to PubMed abstracts or associated MeSH terms [34, 35, 52, 53, 54]. Others use constraints from the biological system under analysis, such as using genetic linkage or quantitative trait locus (QTL) maps to narrow down the set of significant genes to those mapping to regions of the genome associated with appropriate trait [55, 56, 57, 58]. In solid tumor studies, one might look for correlations with genome deletions or amplifications as determined by comparative genomic hybridization on arrays (array CGH; [59, 60]. Finally, in developmental imprinting studies, gene expression may be compared to patterns of methylation [61, 62].

Another very attractive approach is to use the properties of the data and the construction of the array to look for significant functional associations. Recall that one of the key elements in establishing an array platform is the annotation of the arrayed probe elements. For example, imagine that 20% of the genes on the array are annotated as belonging to gene ontology (GO) categories representing energy metabolism. If this is the case, randomly selecting a collection of "significant" genes would most likely yield about 20% of its elements as belonging to the same energy metabolism class. In fact, it would not be surprising to find that 30% of the genes in the "significant" set were energy metabolism genes; however, if the fraction were 80%, it might suggest that the experiment affected energy metabolism with a much higher frequency than would be expected by chance. Such insight may indeed provide clues as to the mechanisms at work in the biological system under study.

An obvious question is whether the probability that a given functional class is over-represented in our significant gene set can be estimated. This can be done using the Fisher Exact Test, and the mathematics behind the approach are described in Box 16.3.

The Classification Problem

As mentioned earlier, some microarray experiments do not focus on identifying function, but rather on finding genes that can be used to group samples into biologically or clinically relevant classes and supervised approaches to data analysis are particularly useful for these

studies. One typically begins with *a priori* knowledge of the groups represented in the data, although any hypothesis along these lines can be further explored using clustering techniques and other information. With those groups, one then asks whether there are genes that can be used to separate the relevant classes. For two groups of samples, a t-test or unpaired two-class SAM are useful tools while, for a larger number of classes, ANOVA or multi-class SAM are appropriate. Having identified a set of genes that show significant differences, one then builds a classification algorithm that can be used to assign a new sample to one of the classes.

There are a wide range of algorithms that have been used for classification, including weighted voting [11], artificial neural networks [24, 25], discriminant analysis [26, 27, 28, 29], classification and regression trees [30], support vector machines [32, 63], *k*-nearest neighbors [33], and a host of others. Essentially, each of these uses an original set of samples – a *training set* – to develop a rule that takes a new test sample from a test set and uses its expression vector sample, trimmed to a previously identified set of classification genes, to place this test sample into the context of the original sample set, thus identifying its class.

In many ways, kNN is the simplest approach to doing classification. First, one must assemble a collection of expression vectors for our samples and assign the samples to various experimental classes. We will refer to these samples, about which we have prior knowledge, as our training set. Next, genes are selected that separate the various classes using an appropriate statistical test to identify good classification candidate genes, thus reducing the size of the sample classification vectors. This represents a first-pass collection of classification genes. The next step is to identify and eliminate samples that appear to be outliers. These may be important because they possibly represent new subclasses in our original sample classification set; alternatively, they may just represent poor-quality data. The outlying samples are identified by applying a correlation filter to the reduced sample expression vectors, as follows:

1. The Pearson correlation coefficient (r) is computed between a given vector and each member of the training set; the maximum r identified is called the r_{max} for that vector.
   The vector is randomized a user-specified number of times. Each time, an r_{max} is calculated using the randomized vector (called r^*_{max}), just as in Step 1.
2. The fraction of times r^*_{max} exceeds r_{max} over all randomizations is used to calculate a p-value for that vector.
3. If the p-value for a vector is less than a user-specified threshold (meaning it is well-correlated with other samples), that vector is retained for further analysis. Otherwise, it is discarded.
   Steps 1-4 are repeated for every sample vector in the set.

At this point, the training set has led to the generation of a collection of sample vectors that represent prior knowledge regarding the biological classes represented in the data. The next step in the analysis involves assigning new samples from the *test set* to classes, based on their expression vectors.

For each sample in the test set, its expression vector is reduced to include only those genes previously identified as being significant for classification. The distance between this reduced expression vector and the reduced expression vectors is then computed for each and every sample in the training set. As the name kNN implies, some number k of nearest neighbors is chosen from the training set – those k vectors that have the smallest distances from the test sample. The new test vector is then assigned to the class most highly represented in its k nearest neighbors. If there is a tie, the new sample remains unclassified.

A Few Closing Thoughts

During the past few years, there have been many discussions in the literature on "noise" in microarray assays: disparate results arising from the use of different platforms, questions regarding the validity of microarray results, and the need to validate the findings. If one closely examines the underlying issues, it is clear that microarrays are no different than any other approach to assaying levels of gene expression – each method has its own biases and limitations. Microarrays simply provide much more data than do techniques such as quantitative RT-PCR or Northern blots, and the likelihood of false positives and false negatives increases as the number of genes assayed increases – a manifestation of the multiple testing problem. What is underlying all of these issues is trying to understand what can be done with the data that emerges. Although there are no absolute answers, there are some overarching generalizations that can be made that will help guide the follow-on experiments.

First, whether one is doing a mechanistic study or trying to identify genes that can be used for sample classification, what microarray assays generally give us are lists of genes that can be significantly correlated with some classes in our experiments. These should be treated not as truths, but as hypotheses that can be tested.

Second, statistical significance is fine, but biological significance is better. Statistics provides very powerful tools for identifying candidate genes, for prioritizing them in the lack of any other evidence, and in helping to re-

solve features in the data. For example, if 30 of the top 50 genes in a list are "energy metabolism" genes, a likely working hypothesis is that the experimental system under study involves changes in energy metabolism, regardless where these 30 fall in the list.

Third, it is very important to note that any change in how one does the analysis is likely to change what is identified as significant in any experiment. This obviously includes changing microarray platform, but even subtle changes in the analysis performed on a single platform can change what is identified as significant. Starting from the laboratory protocols to parameters for slide scanning, image processing, data normalization, and the choice of the analysis tools all contribute significantly to outcome of any analysis. One way to approach this problem is to apply multiple approaches and then look for a common set of "high confidence" genes. Another way that may be more useful is to look at pathways and functional classes in order to identify common biological themes that overlie all of the analyses. It is these functional classes and pathways that we are ultimately interested in.

Fourth, in confirming any microarray result is often useful to use an alternative technique to assay gene expression levels. At this stage, it is useful to define two approaches – verification and validation. Verification involves using another technique with the same RNA samples used for array profiling. What verification does is to confirm the observed patterns of expression in the sample set under analysis and addresses questions related to the bias in the technique. Validation, on the other hand, uses independent RNA sources to assay the individual genes and their patterns of expression. This can confirm the results independent of biases in the sample selection and in the choice of a particular technique. Validation also includes LKOCV and LOOCV approaches to classification algorithms. Validation is a much more powerful statement than verification.

Finally, all microarrays provide are correlations between a particular pattern of expression and some biological class. The real biology is not on the array, but back in the laboratory. Microarray experiments can be powerful tools for developing testable hypotheses, and can even play a significant role in conducting such tests. The value of functional genomics experiments is that they provide unbiased surveys of large numbers of genes and this can be extremely powerful for discovering potential new mechanisms and new subgroups within classes. Ultimately, microarrays remain a tool for discovery, and bioinformatics is simply a filter that can increase the power of any array experiment.

The one relevant question we might ask is when, if ever, microarrays will have an impact on the practice of medicine. The truth is that their applications, particularly in the realm of disease classification, are already starting to be seen. One important example comes from the Netherlands breast cancer study [64], which sought to distinguish between patients with the same stage of disease but different response to treatment and overall outcome. The study was motivated by the observation that the best clinical predictors for metastasis, including lymph node status and histological grade, did not provide adequate prediction of clinical outcome. As a result many patients receive chemotherapy or hormonal therapy regardless of whether they are likely to benefit from this additional treatment. The goal of their analysis was to identify signatures that would allow for individually tailored therapeutic strategies. By profiling tumors from 117 young patients and looking for correlations with clinical outcome, they were able to identify a "poor prognosis" signature comprised of 70 genes that was predictive of a short interval to distant metastasis in lymph node negative patients. Their analysis demonstrated that microarray-based signatures could outperform any clinically-based predictions of outcome in identifying those patients who would benefit most from adjuvant therapy. The success of this initial study motivated a more extensive independent follow-up study involving 295 patients, 12 of whom showed that the 70-gene classification profile was a more powerful predictor of the outcome of disease in young patients with breast cancer than standard systems based on clinical and histological criteria. The success of these two studies has led to a nation-wide clinical trial in the Netherlands in which gene expression profiles for these 70 classifier genes are being collected on all breast cancer patients and used as an adjunct to classical clinical staging. Although we are still eagerly awaiting the outcome of this study, it is clear that the use of expression profiles as biomarkers to predict disease prognosis and outcome is coming of age and other applications are sure to follow.

References

1. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003 31(4): e15.
2. Quackenbush J. Microarray data normalization and transformation. Nat Genet 2002; 32 Suppl: 496-501.
3. Schadt EE, Li C, Ellis B, Wong WH. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J Cell Biochem 2001; Suppl 37: 120-5.
4. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, et al. Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol 2002; 3(11): research0062.
5. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalizatoin for cDNA microarray data: a

robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002; 30(4): e15.
6. Hoffmann R, Seidl T, Dugas M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biol 2002; 3(7), RESEARCH0033.
7. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998; 95(25): 14863-8.
8. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ, Jr., Kohn KW, et al. An information-intensive approach to the molecular pharmacology of cancer. Science 1997; 275: 343-9.
9. Wen X, Fuhrman S, Michaels GS, Carr DB, Smit, S, Barker JL, et al. Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci U S A 1998; 95: 334-9.
10. Soukas A, Cohen P, Socci ND, Friedman JM. Leptin-specific patterns of gene expression in white adipose tissue. Genes Dev 2000; 14(8): 963-80.
11. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999; 96(6): 2907-12.
12. Toronen P, Kolehmainen M, Wong G, Castren E. Analysis of gene expression data using self-organizing maps. FEBS Lett 1999; 451(2): 142-6.
13. Wang J, Delabie J, Aasheim H, Smeland E, Myklebost O. Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. BMC Bioinformatics 2002; 3: 36.
14. Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 2001; 17(2): 126-36.
15. Butte AJ, Kohane IS. Unsupervised knowledge discovery in medical databases using relevance networks. Proc AMIA Symp 1999;: 711-5.
16. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, et al. A gene expression map for Caenorhabditis elegans. Science 2001; 293: 2087-92.
17. Raychaudhuri S, Stuart JM, Altman RB. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 2000;: 455-66.
18. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, et al. Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods 2005; 2(6): 351-6.
19. Dobbin KK, Beer DG, Meyerson M, Yeatman TJ, Gerald WL, Jacobson JW, et al. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. Clin Cancer Res 2005; 11: 565-72.
20. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, et al. Multiple-laboratory comparison of microarray platforms. Nat Methods 2005; 2(5): 345-50.
21. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. Independence and reproducibility across microarray platforms. Nat Methods 2005; 2(5): 337-44.
22. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001; 98(9): 5116-21.
23. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286(5439): 531-7.
24. Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, et al. Multi-platform, multi-site, microarray-based human tumor classification. Am J Pathol 2004; 164(1): 9-16.
25. Ellis M, Davis N, Coop A, Liu M, Schumaker L, Lee RY, et al. Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses. Clin Cancer Res 2002; 8(5): 1155-66.
26. Antoniadis A, Lambert-Lacroix S, Leblanc F. Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics 2003; 19(5): 563-70.
27. Le QT, Sutphin PD, Raychaudhuri S, Yu SC, Terris DJ, Lin HS, et al. Identification of osteopontin as a prognostic plasma marker for head and neck squamous cell carcinomas. Clin Cancer Res 2003; 9(1): 59-67.
28. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 2002; 18(1): 39-50.
29. Orr MS, Scherf U. Large-scale gene expression analysis in molecular target discovery. Leukemia 2002; 16(4): 473-7.
30. Boulesteix AL, Tutz G, Strimmer K. A CART-based approach to discover emerging patterns in microarray data. Bioinformatics 2003; 19(18): 2465-72.
31. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 2000; 97(1): 262-7.
32. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001; 98(26): 15149-54.
33. Theilhaber J, Connolly T, Roman-Roman S, Bushnell S, Jackson A, Call K, et al. Finding genes in the C2C12 osteogenic pathway by k-nearest-neighbor classification of expression data. Genome Res 2002; 12(1): 165-76.
34. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conkli, BR. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol 2003; 4(1): R7.
35. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 2003; 4(4): R28.
36. Hosack DA, Denni, G, Jr., Sherma, BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol 2003; 4(10): R70.
37. Akutsu T, Miyano S, Kuhara S. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. J Comput Biol 2000; 7(3-4): 331-43.
38. Savoie CJ, Aburatani S, Watanabe S, Eguchi Y, Muta S, Imoto S, Miyano S, et al. Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades. DNA Res 2003; 10(1): 19-25.
39. Soinov LA. Supervised classification for gene network reconstruction. Biochem Soc Trans 2003; 31(Pt 6): 1497-502.
40. Datta A, Choudhary A, Bittner ML, Dougherty ER. External control in Markovian genetic regulatory networks: the imperfect information case. Bioinformatics 2004; 20(6): 924-30.
41. Hashimoto RF, Kim S, Shmulevich I, Zhang W, Bittner ML, Dougherty ER. Growing genetic regulatory networks from seed genes. Bioinformatics 2004; 20(8): 1241-7.
42. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 2002; 18(2): 261-74.
43. Shmulevich I, Dougherty ER, Zhang W. Gene perturbation and intervention in probabilistic Boolean networks. Bioinformatics 2002; 18(10): 1319-31.
44. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol 2000; 7(3-4): 601-20.
45. Imoto S, Kim S, Goto T, Miyano S, Aburatani S, Tashiro K, Kuhara S. Bayesian network and nonparametric heteroscedastic regression for non-linear modeling of genetic network. J Bioinform Comput Biol 2003; 1(2): 231-52.
46. Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, et al. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. Bioinformatics 2003; 19 Suppl 2: II227-36.
47. Zou M, Conzen SD. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 2005; 21(1): 71-9.
48. Famili I, Mahadevan R, Palsson BO. k-Cone analysis: determining all candidate values for kinetic parameters on a network scale. Biophys J 2005; 88(3): 1616-25.
49. Wiback SJ, Mahadevan R, Palsson BO. Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the Escherichia coli spectrum. Biotechnol Bioeng 2004; 86(3): 317-31.
50. Michaels GS, Carr DB, Askenazi M, Fuhrman S, Wen X, Somogy, R. Cluster analysis and data visualization of large-scale gene expression data. Pac Symp Biocomput 1998;: 42-53.
51. Aronow BJ, Toyokawa T, Canning A, Haghighi K, Delling U, Kranias E, et al. Divergent transcriptional responses to independent genetic causes of cardiac hypertrophy. Physiol Genomics 2001; 6(1): 19-28.
52. Fink JL, Drewes S, Patel H, Welsh JB, Masys DR, Corbeil J et al. 2HAPI: a microarray data analysis system. Bioinformatics 2003; 19(11): 1443-5.

53. Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 2001; 28(1): 21-28.
54. Masys DR, Welsh JB, Lynn Fink J, Gribskov M, Klacansky I, Corbeil J. Use of keyword hierarchies to interpret gene expression patterns. Bioinformatics 2001; 17(4): 319-26.
55. Cook DN, Wang S, Howles GP, Speer M, Churchill G, Quackenbush J et al. The genetics of innate immunity in the lung. Chest 2003; 123(3 Suppl): 369S.
56. Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 2002; 3(1): 43-52.
57. Kwitek-Black AE, Jacob HJ. The use of designer rats in the genetic dissection of hypertension. Curr Hypertens Rep 2001; 3(1): 12-8.
58. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature 2003; 422(6929): 297-302.
59. Cheung ST, Chen X, Guan XY, Wong SY, Tai LS, Ng IO, et al. Identify metastasis-associated genes in hepatocellular carcinoma through clonality delineation for multinodular tumor. Cancer Res 2002; 62(16): 4711-21.
60. Gray JW, Collins C. Genome changes and gene expression in human solid tumors. Carcinogenesis 2000; 21(3): 443-52.
61. Chen H, Liu J, Zhao CQ, Diwan BA, Merrick BA, Waalkes MP. Association of c-myc overexpression and hyperproliferation with arsenite-induced malignant transformation. Toxicol Appl Pharmacol 2001; 175(3): 260-8.
62. Ehrlich M. DNA hypomethylation, cancer, the immunodeficiency, centromeric region instability, facial anomalies syndrome and chromosomal rearrangements. J Nutr 2002; 132(8 Suppl): 2424S-2429S.
63. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000; 16(10): 906-14.
64. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415(6871): 530-6.
65. Baggerly KA, Coombes KR, Hess KR, Stivers DN, Abruzzo LV, Zhang W. Identifying differentially expressed genes in cDNA microarray experiments. J Comput Biol 2001; 8(6): 639-59.
66. Long, AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, Baldi P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. J Biol Chem 2001; 276(23): 19937-44.
67. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13(11): 2498-504.

Correspondence to:
John Quackenbush
Department of Biostatistics
Dana-Farber Cancer Institute
Mayer 232
44 Binney Street
Boston, MA 02115
USA
E-mail: johnq@jimmy.harvard.edu