DOI: 10.1534/genetics.104.031500

Adding Confidence to Gene Expression Clustering

B. Munneke,*,[1] K. A. Schlauch,[†,1] K. L. Simonsen,* W. D. Beavis[‡] and R. W. Doerge*,[§,2]

**Department of Statistics, Purdue University, West Lafayette, Indiana 47907, [†]Center for Biomedical Genomics and Informatics, George Mason University, Manassas, Virginia 20110, [‡]National Center for Genome Resources, Santa Fe, New Mexico 87505 and [§]Department of Agronomy, Purdue University, West Lafayette, Indiana 47907*

Manuscript received May 21, 2004
Accepted for publication April 25, 2005

ABSTRACT

It has been well established that gene expression data contain large amounts of random variation that affects both the analysis and the results of microarray experiments. Typically, microarray data are either tested for differential expression between conditions or grouped on the basis of profiles that are assessed temporally or across genetic or environmental conditions. While testing differential expression relies on levels of certainty to evaluate the relative worth of various analyses, cluster analysis is exploratory in nature and has not had the benefit of any judgment of statistical inference. By using a novel dissimilarity function to ascertain gene expression clusters and conditional randomization of the data space to illuminate distinctions between statistically significant clusters of gene expression patterns, we aim to provide a level of confidence to inferred clusters of gene expression data. We apply both permutation and convex hull approaches for randomization of the data space and show that both methods can provide an effective assessment of gene expression profiles whose coregulation is statistically different from that expected by random chance alone.

MICROARRAY technology has been applied experimentally across many biological disciplines; some of the earliest examples in agriculture used expressed sequence tags (ESTs) in studies of plant gene expression (Ewing *et al.* 1999). Similarly, in human studies, cDNA microarray data have been used as the vehicle for the investigation of biologic variation in mammary epithelial cells among breast tumor samples (Perou *et al.* 1999). The evolutionary biology of yeast was studied using microarray techniques to compare genome-wide expression patterns in evolved *vs.* parental strains after 250 generations of growth (Ferea *et al.* 1999). Since its initial impact the use of microarray technology now includes much broader investigations of regulatory science (Doerge 2002; Cheung *et al.* 2003; Schadt *et al.* 2003; Brem and Kruglyak 2005; Kim *et al.* 2005; Ronald *et al.* 2005) and reconstruction of genetic networks (Knudsen 2002).

Microarray technologies allow researchers to simultaneously monitor cellular activity of many gene transcripts. These experiments produce mRNA expression data in great abundance and provide useful information to pursue the conjectures of functional genomics. Two different approaches for analyzing these data, both of which rely heavily on statistics, include testing each gene for differential expression and/or summarizing gene expression profiles by assessing similarities in their pattern of behavior across treatments or conditions. The statistical issues (*e.g.*, statistical models, hypotheses, multiple comparisons, etc.) involved in testing differential expression have been investigated more thoroughly (see review by Craig *et al.* (2003)) than those statistical issues surrounding the statistical assessment of gene expression profile patterns and their clustering (Zhang and Zhao 2000; Kerr and Churchill 2001; Munneke 2001; Tibshirani *et al.* 2001; McShane *et al.* 2002).

In efforts to understand the underlying data structure resulting from expression studies, data are typically summarized by grouping the expression intensities via similarity of response to various experimental conditions. Statistical methodologies that supply these groups, or clusterings, of gene expression profiles are known as clustering techniques and originate from a well-established area of statistics referred to as multivariate statistics (Johnson and Wichern 1998). The clustering methods involved in the previously mentioned studies tend to be either partition based, such as *K*-means clustering (Tavazioe *et al.* 1999) and self-organizing maps (Tamayo *et al.* 1999), or nested, as with hierarchical clustering (Eisen *et al.* 1998). Regardless of the clustering method used, an assignment of statistical significance for the resulting partitions is necessary to interpret cluster reliability and biological meaning (Zhang and Zhao 2000; Kerr and Churchill 2001; McShane

[1]These authors contributed equally to this work.

[2]*Corresponding author:* Department of Statistics, 1399 Mathematical Sciences Bldg., 150 N. University St., Purdue University, West Lafayette, IN 47907-1399. E-mail: doerge@purdue.edu

et al. 2002). Several such approaches have been developed. Kerr and Churchill (2001) present a bootstrapping technique to assess the stability of profile clusters. Genes are assigned to a set of fixed profiles via their correlation; a correlation coefficient >0.90 is enough to assign two profiles to the same cluster. Bootstrapping is then employed to assess stability of a gene by counting how many times it is assigned to the same fixed profile. Zhang and Zhao (2000) also use bootstrapping to generate perturbed data sets. Their reliability measure depends on the number of times genes occur in the set of perturbed clusters. McShane *et al.* (2002) present a principal components analysis to assess the overall clustering of expression patterns and then test whether gene expression profiles arise from a single multivariate normal distribution. New data sets with artificial experimental error are generated by adding Gaussian white noise to the original expression levels. Two reproducibility indexes are associated to each cluster by computing the number of pairs existing in both the original and the perturbed cluster. A perturbed cluster is matched to the original cluster if it contains a majority of elements in common with the original cluster.

CLUSTERING

Agglomerative hierarchical clustering (Hartigan 1975; Gordon 1999) is a technique that is both familiar to biologists and widely used in reporting groups of coregulated genes. It produces an easy-to-interpret clustering tree (dendrogram), provided the number of genes or treatments is small. A dendrogram is a visual representation of data points or gene expression profiles as merged together from isolated points toward increasingly larger subgroups. As a general rule, in the process of agglomerating the subgroups, once two items are associated to a subgroup, they remain in the same subgroup as the number of subgroups grows larger. The continued agglomeration of subgroups defines the nesting property of hierarchical clustering, and it is this property that aids interpretation of the results. The mechanics of hierarchical clustering require two important and sensitive specifications: a dissimilarity, or distance, measure and a joining, or agglomerative, method.

All clustering methods when applied to gene expression data depend on a dissimilarity measure, d_{ij}, in such a way that for two genes g_i and g_j the measure obeys three properties (1–3), while a true distance measure also satisfies a fourth (see, *e.g.*, Johnson and Wichern 1998):

1. $d_{ij} \geq 0$.
2. $d_{ii} = 0$.
3. $d_{ij} = d_{ji}$.
4. $d_{ik} \leq d_{ij} + d_{jk}$.

As the mRNA expression information is the result of an experiment comparing genes from an organism under various developmental time points or treatments, for a t treatment experiment a single gene will have a t-dimensional observation vector known as its gene expression profile. A dissimilarity measure in this setting is a real-valued function that, for any two expression vectors, assigns a nonnegative real number denoting the dissimilarity between them. In this work, we introduce a dissimilarity measure that takes into account both magnitude and pattern of gene expression. Typically, expression studies employ measures that depend on either the pattern or the relative magnitude of profile pairs. Our dissimilarity measure, also known as the Munneke metric,

$$d_{ij} = \left(1 - \frac{g_i \cdot g_j}{\|g_i\| \, \|g_j\|}\right) + 2\left(\frac{(1/t)\sum_{k=1}^{t}|g_{i_k} - g_{j_k}|}{1 + (1/t)\sum_{k=1}^{t}|g_{i_k} - g_{j_k}|}\right), \quad (1)$$

takes into account both notions. Its first term provides a notion of dissimilarity via the uncentered correlation coefficient, which measures similarity of profile patterns. The second term adds a measure of difference in magnitude between pairs, based on the average distance between paired expression levels across all experimental states. Defined as such, the dissimilarity measure is able to differentiate profiles of the same pattern, but at different intensity levels, which we believe is a biologically important feature of a metric.

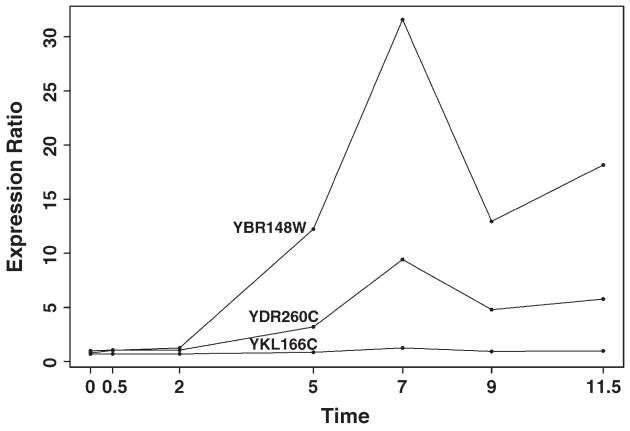

Figure 1.—Three expression profiles (YBR148W, YDR 260C, and YKL166C) from the yeast sporulation (Eisen *et al.* 1998) data set. Denoting Pearson's correlation coefficient as cor, $\text{cor}_{(\text{YBR148W,YDR260C})} = 0.991$, $\text{cor}_{(\text{YDR260C,YKL166C})} = 0.998$, and $\text{cor}_{(\text{YBR148W,YKL166C})} = 0.993$, rendering all three profiles essentially identical. The Munneke metric (1) provides the dissimilarity measure $d_{(\text{YBR148W,YDR260C})} = 1.77$, $d_{(\text{YDR260C,YKL166C})} = 1.58$, and $d_{(\text{YBR148W,YKL166C})} = 1.97$.

As an example, consider the three gene expression profiles in Figure 1, taken from the Eisen *et al.* (1998) yeast sporulation data set. Letting cor denote Pearson's correlation coefficient, we have $\text{cor}_{(\text{YBR148W,YDR260C})} = 0.991$, $\text{cor}_{(\text{YDR260C,YKL166C})} = 0.998$, and $\text{cor}_{(\text{YBR148W,YKL166C})} = 0.993$, rendering all three pairs almost identical. If we let euc denote the Euclidean distance, then $\text{euc}_{(\text{YBR148W,YDR260C})} = 28.14$, $\text{euc}_{(\text{YDR260C,YKL166C})} = 10.48$, and $\text{euc}_{(\text{YBR148W,YKL166C})} =$

38.55, implying that gene YBR148W is more similar to YDR260C than to YKL166C, a result that does not agree with the correlation coefficient result. Employing the Munneke metric (1) on these same data yields $d_{(YBR148W,YDR260C)} = 1.77$, $d_{(YDR260C,YKL166C)} = 1.58$, and $d_{(YBR148W,YKL166C)} = 1.97$. While the correlation coefficient between the two pairs is almost identical, the Munneke metric (1) is able to differentiate the two pairs with respect to relative magnitude, which we believe to be a more intuitive measure of dissimilarity. We realize that many penalty terms can be used (Gordon 1999), but we restrict our attention to the described penalized dissimilarity measure (1) as it distinguishes between induced and repressed gene expression of the same magnitude.

The joining method complements the dissimilarity measure and is chosen from three of many possible methods; nearest neighbor (single linkage), farthest neighbor (complete linkage), and average linkage (Gordon 1999). Using any measure, the dissimilarity between two groups is assessed by considering all pairs of genes formed by taking one member of each group. Throughout this work we concentrate on agglomerative hierarchical clustering and employ the Munneke dissimilarity measure (1) with these joining methods.

The amount of random variation found in gene expression studies is known to have numerous sources (Knudsen 2002; Craig *et al.* 2003). Given the lack of ability to discriminate how this variation affects clustering results, users of clustering algorithms are growing increasingly careful in interpreting their results. Furthermore, because the technology is expensive, there is also a desire to compare the results across laboratories and experiments. Ideally, this comparison should be possible at the level of cluster analysis, independent of the algorithm, dissimilarity function, or joining method. Our purpose in this work is to rely on the penalized dissimilarity measure (1) and to place a level of confidence on the groupings suggested by hierarchical clustering (or any other clustering algorithm). Our focus is not on the clustering mechanism itself, but rather on deriving a statistical assessment of significance as one traverses down the branches of a hierarchical clustering tree. In essence, we propose a comparison of the original data with a randomization of the data space (*i.e.*, no association between the expression profiles) for the purpose of assessing how probable any given cluster is to have occurred by chance, thus providing researchers with a baseline for comparing results independent of the clustering algorithm used.

METHODS

Clustering methodologies operate on objects in a binary fashion such that each operation separates a group into two distinct subgroups. Here, we utilize randomization methods to determine when such splits are spurious or real (Good 2000). At some branch point it is expected that no further distinct subgroup structure will exist, and that the existing structure will not differ from a random distribution of gene expression profiles. As such, subsequent branches or subclusters will have no statistically significant meaning.

Two main features are involved in our proposed method of placing a level of certainty on branches of a clustering tree. The first is the creation of a random representation of the actual expression profiles. The second is a test statistic that summarizes the existing cluster tree structure, whether from the actual data or from the randomized data. Randomization of gene expression information in this context is accomplished using two different methods: permutation tests and convex hulls. Permutation tests (Fisher 1935) have been successfully employed in the quantitative trait loci (QTL) literature (Churchill and Doerge 1994; Doerge and Churchill 1996; Nettleton and Doerge 2000) to establish experimental thresholds for declaration of statistically significant QTL. The application of permutation methods to gene expression data is analogous to the QTL application in that individual expression measures that are related (associated) through coregulation will lie in the same cluster. However, in certain circumstances (discussed later) permutation methods are less effective in randomizing the data space, and a convex hull (Barber *et al.* 1997) approach is employed as an alternative. Finally, a test statistic is required for the purpose of measuring the amount of structure present in a dendrogram via a single number. This statistic distinguishes between the data sets without structure (*i.e.*, reflecting a randomized data space) and the original data space. The underlying goal is to calculate a statistic that complements the dissimilarity measure used to create the dendrogram while clearly distinguishing between random gene expression profiles and clusters and gene expression data that demonstrate true associations.

Randomization: *Permutation method:* Consider the gene expression data to be real-valued t-dimensional vectors, one dimension for each treatment in the experiment. The data vectors represent specific gene expression levels recorded under each of the t treatment conditions. The random gene expression profiles (or the permuted data) are constructed by focusing on one treatment at a time and sampling without replacement from this treatment for every expression observation (gene). Via the permutation, the value of each expression profile is equally likely to be any one of the set of all values observed for that variable over the course of the original experiment. Thus, under the null hypothesis, a random permutation represents an outcome that is as likely to have been observed as the original data, without parametric assumptions of data structure. Furthermore, the permutation method also allows for any inherent association between individual profiles to be broken. While one permutation of the data creates a single randomized

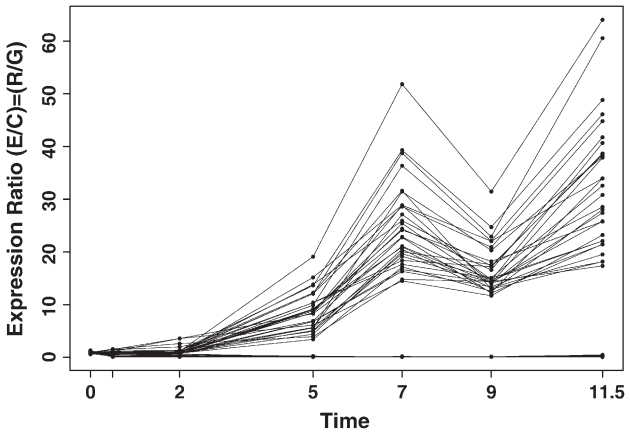

FIGURE 2.—Original data set: gene expression profiles for 58 yeast genes across seven temporal states. These data are part of a larger publicly available gene expression experiment of 6118 genes. (http://cmgm.stanford.edu/pbrown/sporulation/additional/spospread.txt).

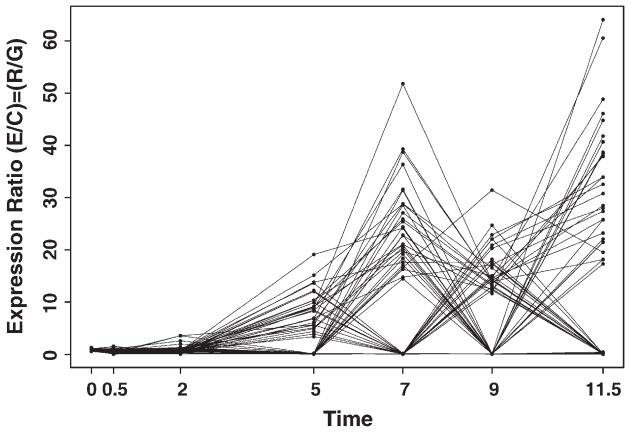

FIGURE 3.—Permuted data set using permutations only: a single randomization of the 58 distinct gene profiles using permutation.

data set, repeating this process n times creates a collection of n random data sets, each potentially representing an experiment having no inherent structure among its data points.

Permutation techniques for null model generation are easy to implement and aim to remove much of the association between gene expression profiles (*i.e.*, coregulation). Not all data sets can be associated with permutations without cluster structure: even after permuting the original data up to 10,000 times it is possible, and has been seen, that the randomized data retain a looser cluster structure. We have found that some gene expression data are aligned in the data space in a manner that prohibits permutation from randomizing the data space efficiently, *e.g.*, a group of expression vectors forming a long thin cloud (*i.e.*, cigar shaped) in two-dimensional space. We are motivated to provide a null data set that has little or no cluster structure, but is still contained within the boundaries of the original data. For this, we introduce an alternate approach of randomly mixing gene expression vectors within their convex hull.

Convex hull: The convex hull of a set G of m gene expression vectors in R^t is the smallest convex set in R^t containing all m vectors and is defined as

$$C_G = \left\{ \sum_{i=1}^{m} a_i g_i \Big| \sum_{i=1}^{m} a_i = 1,\ a_i \geq 0 \right\}$$

(BRØNDSTED 1983). Randomly generating elements within C_G provides a random distribution of points within the hull. Therefore, the null distribution of a test statistic (described later) that is used to assess the resulting dendrogram can be generated by performing the clustering procedure on the n convex hull generated data sets. Observe that elements defined by

$$\{x_{\text{new}} = \lambda x_i + (1 - \lambda) x_j;\ \lambda \in [0, 1],\ i \neq j\},$$

where x_i and x_j are expression vectors in the original data, lie within the convex hull of the original data set G. To generate data sets with boundaries identical to those of the original data, the null model data sets are generated to be within the convex hull of the original data values, by representing new elements as random linear combinations of original elements.

Permutation *vs.* convex hull: As mentioned previously, it is sometimes the case that the permutation approach alone (without the convex hull) generates random data sets with subcluster structure. Consider the gene expression profiles of 58 yeast genes across seven temporal states as shown in Figure 2. These data are part of a large public gene expression experiment of 6118 yeast genes undergoing sporulation (http://cmgm.stanford.edu/pbrown/sporulation/additional/spospread.txt). A single random data set generated by permutations alone is shown in Figure 3, and one generated using the convex hull approach is shown in Figure 4. We believe that the convex hull approach is more reliable than the permutation approach in generating data sets without subcluster structure for original data in which strong cluster structure exists. In such cases, artificial/permuted data generated by permutations will tend to retain some, if not all, of the original subcluster structure. Consider the example of Figure 2, in which each expression vector belongs to one of two clusters. For each experiment, except the first at time 0, the set of possible values is divided into two distinct sets. Thus, the permuted value for each experiment will belong to one of these two distinct sets, generating expression vectors that lie within one of $64 = 2^6$ possible distinct clusters, as is demonstrated in Figure 3. Alternatively, the convex hull approach first generates two random numbers in [0, 1] and then constructs new profiles via linear combinations of original profiles using the two random numbers, making it unlikely that the new profile will lie exactly within one of the original two clusters.

Test statistic: As a means of assessing the subgroup

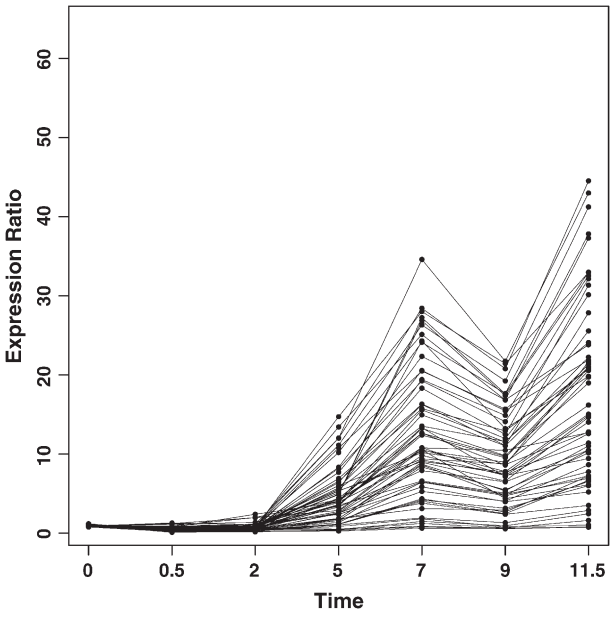

FIGURE 4.—Convex hull permutation: a single randomization of the 58 distinct gene profiles using the convex hull approach.

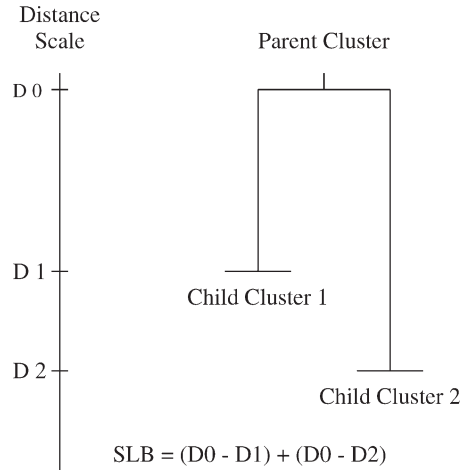

FIGURE 5.—The test statistic SLB is calculated as the *s*um of the branch *l*engths *b*elow the parent branch point and is a function of the penalized distance measure and the agglomeration method.

architecture within any given gene expression experiment or randomized data set, we construct a test statistic based upon the sum of the branch lengths of a cluster dendrogram. The distribution of the test statistic under the null hypothesis is estimated and used to establish statistically significant results in the original gene expression data when compared to the null distribution. The clusters being tested at each stage are provided by the (hierarchical) clustering of the original data, based upon our penalized dissimilarity measure (Munneke metric) (1) and one of the joining methods mentioned previously. Consider Figure 5, noting that for each branch in the dendrogram a left child subgroup and a right child subgroup exist. The sum of branch lengths is calculated as a function of the sum of the differences between the last join, D_0, and each of the last joins for the child clusters, D_1 and D_2 (Figure 5). The distribution of the *sum of the* branch *lengths below* (SLB) the parent node under the null hypothesis is used to assess the statistical significance of the original data dendrogram. SLB focuses on the dissimilarities provided in the clustering of the data for the purposes of assessing the subgroup structure and avoiding criteria such as sum-of-squares.

We define the SLB statistic using both a dissimilarity measure d and an agglomerative method D as required by hierarchical clustering. Three standard agglomerative methods are defined as

$$D_{\max} = \max\{d_{ij} | g_i \in G_0,\ g_j \in G_1\}$$

$$D_{\min} = \min\{d_{ij} | g_i \in G_0,\ g_j \in G_1\}$$

$$D_{\text{avg}} = \frac{1}{n_i n_j} \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} d_{ij}, \quad \text{for } g_i \in G_0,\ g_j \in G_1,$$

where $D_{\max}$, $D_{\min}$, and D_{ave} correspond to the nearest neighbor (single linkage), farthest neighbor (complete linkage), and average linkage agglomerative methods, respectively. Note that the SLB statistic can be used with any agglomerative method.

The cluster of genes G is partitioned into the child subgroups G_0 and G_1, *i.e.*, $G = G_0 \cup G_1$. To calculate the test statistic we denote the partitions of G_0 and G_1 by extending our notation,

$$G = (G_{00} \cup G_{01}) \cup (G_{10} \cup G_{11}).$$

The test statistic is calculated as

$$\text{SLB} = (D_*(G_0, G_1) - D_*(G_{00}, G_{01})) + (D_*(G_0, G_1) - D_*(G_{10}, G_{11})),$$

where D_* corresponds to the user-selected agglomerative method.

Implementation: Using the SLB test statistic, we approach a clustering tree one branch point at a time. Starting from the top, SLB is calculated for the first branch point, which divides the total set of gene expression profiles G into two subgroups G_0 and G_1. Recall that the original data set has a representation in the form of a dendrogram as does each of the randomized data sets. Therefore, the statistic (SLB) is calculated for the original data, as well as for each of the random data sets. If the statistic calculated for the original data is large (*i.e.*, exceeds the $1 - \alpha$ percentile) relative to the distribution of SLB under the null hypothesis, this suggests that the original data space has a stronger subgroup structure than would be expected from a random association of gene expression profiles. The statistically significant subgroup structure is then accepted for this branch point, a probability (P-value) is assigned, and the algorithm continues by operating on each of the two statistically significant subgroups independently. Note

TABLE 1

Number of clusters found

S/N	1	% miscl.[a]	2	% miscl.	3	% miscl.	4	% miscl.	Avg. % miscl.
3.0	0.9600	0.50	0.040	0.4940	0.000	0.0000	0.000	0.00	0.499
4.0	0.8900	0.50	0.104	0.0900	0.005	0.0600	0.001	0.02	0.455
5.0	0.2870	0.50	0.645	0.0047	0.065	0.0102	0.003	0.02	0.147
6.0	0.0080	0.50	0.886	0.0000	0.104	0.0100	0.001	0.02	0.005

Results of signal/noise (*S/N*) simulation. $S/N = |\mu_1 - \mu_2|/\sigma$; $|\mu_1 - \mu_2| = 3.0, 4.0, 5.0, 6.0$; and $\sigma = 1.0$. Data are simulated for two gene clusters of an increasing signal-to-noise ratio. These results display one (1), two (2), three (3), and four (4) clusters, where two clusters is the correct resolution.

[a] Percentage of misclassification.

that each additional partitioning of the data is conditional on acceptance of the partition for the preceding branch point. The algorithm self-terminates when all branch points below valid partitions in the original data set are found not to be statistically significant. If further investigation is desired below a specific subgroup, then the algorithm can be reinitiated beginning at the parent node of this subgroup.

DATA ANALYSIS

Simulation: As a means of assessing the power of both randomization methodologies, we first relied on data simulation. Hierarchical clustering routines HC and HCASS2 (Fionn Murtagh at Université Louis Pasteur, Strasbourg, http://newb6.u-strasbg.fr/fmurtagh//mda-sw/) were employed along with a standard uniform random number generator for permutation. Normal random variates were produced by the Box-Mueller transformation of uniform variates, and two distinct multivariate groups with 50 observations (genes) in each group, at three treatments (or dimensions), were simulated. The distance between the two groups was measured by a signal- (expression intensity) to-noise (variation) ratio. For example, in one dimension the signal-to-noise comparison gives an indication of the level of separation for the means of the two distributions defining the gene clusters. As such, in one dimension, the signal-to-noise ratio can be calculated as $|\mu_1 - \mu_2|/\sigma$ for two normal distributions with means μ_1 and μ_2, respectively, and a common standard deviation σ. The distance between the two multivariate groups was measured using the average signal-to-noise ratio for each of the t (treatment) dimensions, calculated as

$$\frac{1}{t}\sum_{k=1}^{t}\frac{|\mu_{1_k} - \mu_{2_k}|}{\sigma}$$

and tested at four different ($|\mu_1 - \mu_2|$) distances (3.0, 4.0, 5.0, and 6.0). For each original simulated data set 1000 randomizations of the data set were performed using both permutation and convex hull methods. The (SLB) test statistic was calculated for each randomization and each method, and then 1000 random test statistics were used to estimate the null distribution for the permutation method and the convex hull method. Each decision in the cluster process was based on a significance level $\alpha = 0.05$. The number of significant clusters and the misclassification rate were based on 1000 repetitions of this process (Table 1).

Simulation results: Each simulation resulted in a number of significant clusters (ranging from 1 to 100) for each data set. As the data were simulated under the assumption of two known clusters, any number of significant clusters different from two is incorrect. The degree of incorrectness of a cluster analysis, in this context, is the percentage of genes misclassified (*i.e.*, a gene profile is assigned to a cluster different from the cluster from which it was simulated). The percentage of misclassification is the maximum overlap between two of the discovered clusters and the known true generating clusters. For example, if the two true clusters are $G_1 = \{1, \ldots, 50\}$ and $G_2 = \{51, \ldots, 100\}$, and the discovered cluster is $G_{\text{disc.1}} = \{1, \ldots, 100\}$ then 50% of the genes are misclassified. If the discovered clusters are $G_{\text{disc.1}} = \{1, \ldots, 48\}$, $G_{\text{disc.2}} = \{49, \ldots, 94\}$, and $G_{\text{disc.3}} = \{95, \ldots, 100\}$, then 8% of the genes are misclassified. A perfect scenario (for this simulation), as the signal-to-noise ratio increases, maintains 100% of the simulations as having two distinct clusters with 0% misclassification. In fact, we find that the average misclassification tends to 0.0, indicating no misclassification, as the true separation between the two generating clusters tends to infinity. Because our simulation study was based upon two three-dimensional gene expression clusters, we allowed the signal-to-noise ratio to increase in a stepwise manner from 3.0 to 6.0, while the variance remained at 1.0. Only the results for the permutation randomization are shown in Table 1. For both randomization methods, as the signal-to-noise ratio increases the ability to identify statistically significant gene clusters increases. The average percentage of the observations misclassified reflects an increase in the power of the test as the dissimilarity between the clusters becomes larger.

Experimental data: We consider the 517-gene subset of the 8613 genes from human fibroblast cells treated

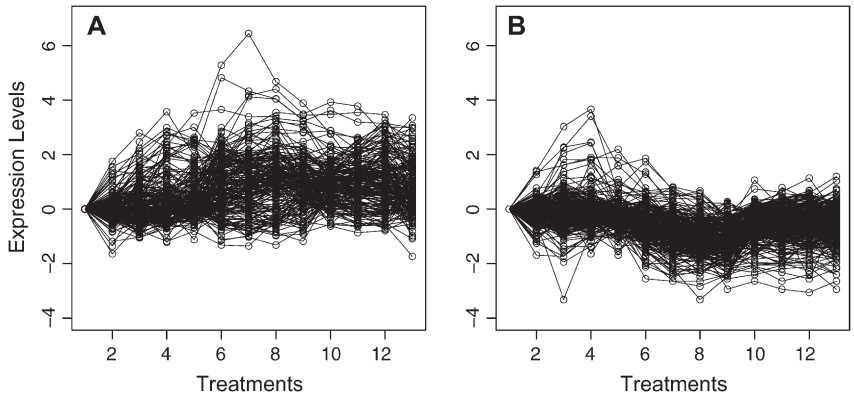

FIGURE 6.—The two gene cluster subgroups resulting from the analysis of the original serum data (IYER *et al.* 1999), using average linkage, the cosine of the angle distance measure, and $\alpha = 0.05$. (A) Group 1 expression profiles, average link. (B) Group 2 expression profiles, average link.

with serum (IYER *et al.* 1999). These data were used to create the cluster dendrogram in IYER *et al.*'s (1999) Figure 2. This figure is available at http://genome-www.stanford.edu/clustering/Figure1.jpg and the data are available at http://genome-www.stanford.edu/serum/data.html. The data consist of measurements of mRNA present at 13 time points following the treatment with serum. The clustering in IYER *et al.*'s (1999) Figure 2 is the result of a hierarchical method, utilizing the cosine of the angle between the gene vectors as the dissimilarity measure, and the agglomerative method. First, we use a permutation-based hierarchical clustering routine with a cosine dissimilarity and average agglomeration method to compare our results directly with those of Iyer. Second, we repeat the test using the Munneke metric (1) and the complete agglomeration method. Third, we repeat the analysis using a significance threshold, $1 - \alpha$. Finally, the analysis is repeated using the convex hull method with the Munneke metric and the average agglomerative method.

The results of the permutation-based randomization (based on 1000 permutations) with hierarchical clustering (using the cosine metric) reveal two subgroups. Using a significance level of $\alpha = 0.05$, the permutation subgroups share common clusters with IYER *et al.*'s (1999) Figure 2. The first branch point is the most obvious and gives the only statistically significant subgroup structure. The interpretation of the groupings (Figure 6) indicates one group of induced genes (group 1) and one group of repressed genes (group 2). Even when the significance level α is increased to $\alpha = 0.40$, this grouping remains intact.

We modified the dissimilarity measure by adding a penalty term (Munneke metric) to the cosine dissimilarity measure (1) and repeated the permutation-based analysis. Any change in the number of significant subgroups between the two permutation-based analyses should be a result of the penalization term in the dissimilarity measure. The permutation-based results present essentially the same results for subgroup membership as the initial analysis (using the cosine metric), indicating that the previously described scenario of equally induced and repressed gene expressions is not of concern for these data.

The final analysis was performed using the Munneke metric (1) and the complete linkage agglomeration method. We chose the complete linkage method because it gives the highest resolution of 517 serum genes into clusters when the significance is increased to $\alpha = 0.15$. Nine cluster subgroups result (Figure 7), highlighting the use of the permutation methodology as an exploratory data analysis tool. All three analyses thus far are based on the empirical distribution of the SLB test statistic, so the results can be compared directly. The penalized dissimilarity measure coupled with the complete joining method discerns the most distinct expression clusterings for the serum data of IYER *et al.* (1999). Results based on the convex hull method differ from those based on permutations: using the convex hull approach, the first branch point is not detected with $\alpha < 0.025$, at which point the subsequent second branch point is also accepted. This may indicate that the convex hull approach provides a stricter criterion for detecting valid cluster structure. All the branches that follow have P-values >0.50, indicating no statistical significance. The relative difference in the results between the permutation randomization and the convex hull randomization is most likely due to the inability of the permutation randomization to remove all valid subcluster structure in the randomized null model data sets. The convex hull approach avoids these issues by guaranteeing, on average, that the randomized data sets have no valid subcluster structure. As a focus of future research, we are considering different criteria to guide the determination of which randomization method to use.

DISCUSSION

An analysis tool that implements permutation or convex hull coupled with any clustering algorithm can be incorporated easily with standard computer code or into any database analysis package so that investigators may derive their own conclusions or compare their results with those of other experiments. We are aware that the

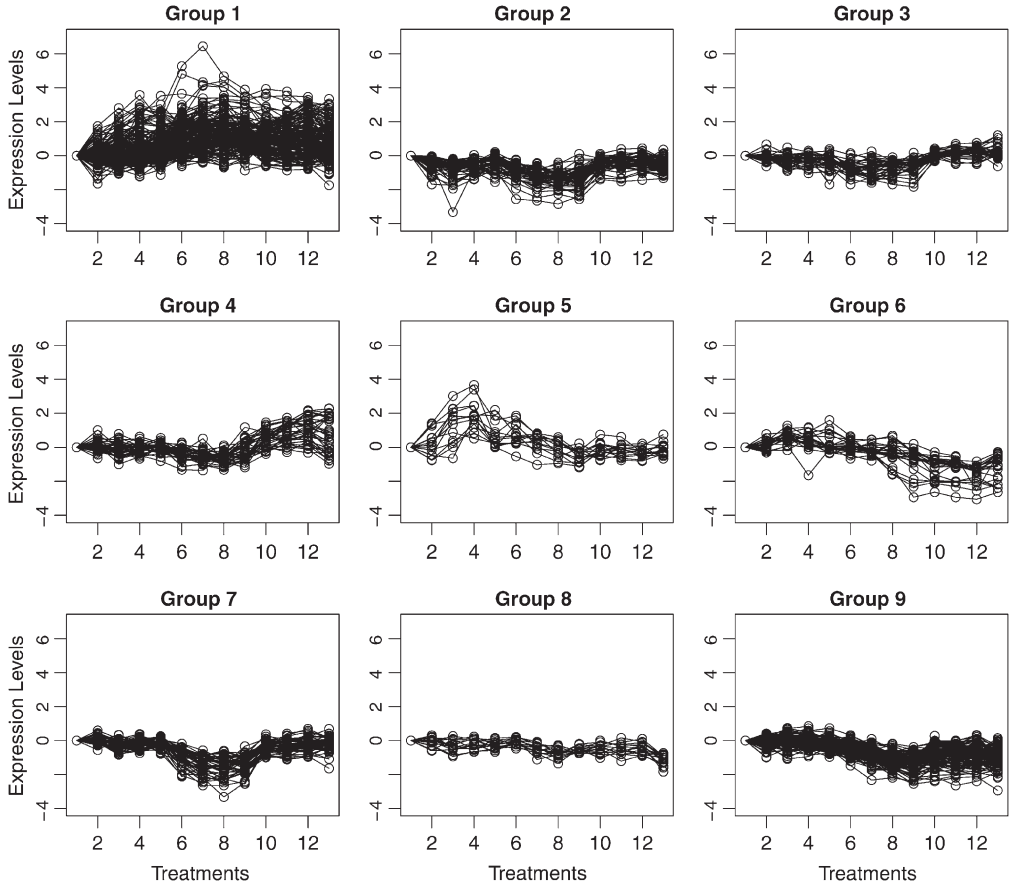

FIGURE 7.—Profile plots of groups 1–9 discovered using complete linkage, penalized distance (1), and $\alpha = 0.15$ for the serum data (IYER *et al.* 1999).

repetitive nature of randomization calculations requires intensive CPU time; however, mainframes with multiple processors are now able to reduce the computer time by a factor proportional to the number of available processors. Thus, the implementation of permutation testing is straightforward. For example, the permutation-based clustering technique has been implemented in the R language and made available by the GeneX system (versions 1.05 and 2.0), available at http://sourceforge.net/projects/genex. Version 2.0 also includes the convex hull clustering approach, as described previously. GeneX is a publicly available gene expression database with an analysis toolkit (MANGALAM *et al.* 2001). Both permutation and convex hull approaches are also available as stand alone programs written in R at http://mason.gmu.edu/~kschlauc/.

The broad utility of randomization techniques as applied to cluster analysis is not limited to genomics or gene expression applications. The randomization methods presented here in the context of assessing gene expression profiles can be implemented in combination with any clustering technique, thus allowing researchers to identify statistically significant subgroup structures in a group of genes selected for study in a given organism. Furthermore, these gene clusters can then be compared across clustering methods and criteria via their attached level of confidence (or *P*-value). The resulting subgroups have great potential to suggest genes that may be coregulated under the conditions studied in the experiment. In work by LAN *et al.* (2003) hierarchical clustering with oblique principal components was used to reduce the dimension of the data space when mapping mRNA abundance as quantitative traits. While no levels of confidence were attached to the clusters, it is easily understood that randomization methods not only fit the application, but also in fact benefit the results.

Several resampling techniques that associate reliability metrics to clustering results have been established. Those mentioned in the Introduction are based on bootstrapping approaches and rely upon the examination and counting of individual elements in the original and perturbed clusters. Through this work we introduced the use of permutation or randomization methods that do not rely upon cluster size and individual elements, but assign one value to each cluster via a test statistic, to assess its "tightness" relative to its subclusters. Additionally, our proposed methods are independent of clustering technique and offer a dissimilarity metric capable of assessing both magnitude and pattern of expression. We believe that our approach is complementary to those mentioned previously, and that it offers a reasonable alternative to assessing reliability of clusters generated by any clustering procedure.

Clustering techniques have been used extensively to

explore the results of gene expression experiments. This is due in part to the intuitive visual appeal that clusters provide for discerning patterns in very large and complex data sets. Perhaps more importantly, biologists use the results from clustering to filter and select sets of (candidate) genes for further hypothesis-driven research. If used for such decision making, it is important to determine whether subsets of gene expression patterns are distinguishable or merely random artifacts. In the absence of independent biological replications we have demonstrated how randomization can aid the exploration process that often accompanies microarray experiments.

We thank the many people who have discussed this work over the years; most significant are two anonymous reviewers, Joel Bader, Michael (Mik) Black, Bruce Craig, Rob Martienssen, George McCabe, Richard Michelemore, Bruno Sobral, and Bruce Trumbo. This work was supported by a Pioneer Hi-Bred research grant and by a U.S. Department of Agriculture-Initiative for Future Agriculture and Food Systems (IFAFS) grant to R.W.D. Work based upon this material was supported by National Science Foundation (NSF) grant 0078307 (K.A.S. and W.D.B.) and NSF grant BDI-0244167 (GeneX; K.A.S.).

LITERATURE CITED

Barber, C., D. Dobkin and H. Huhdanpaa, 1997 The quickhull algorithm for convex hulls. ACM Trans. Math. Sci. **22:** 469–483.

Brem, R. B., and L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. USA **102** (5): 1572–1577.

Brøndsted, A., 1983 *An Introduction to Convex Polytopes* (Graduate Texts in Mathematics, No. 90). Springer-Verlag, New York.

Cheung, K. J., V. Badarinarayana, D. W. Selinger, D. Janse and G. M. Church, 2003 A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. Genome Res. **13** (2): 206–215.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Craig, B. A., M. A. Black and R. W. Doerge, 2003 Microarrays: the technology and analysis. J. Agric. Biol. Environ. Stat. **8** (1): 1–28.

Doerge, R. W., 2002 Mapping and analysis of quantitative trait loci in experimental populations. Nat. Rev. Genet. **3:** 43–52.

Doerge, R. W., and G. A. Churchill, 1996 Permutation tests for multiple loci affecting quantitative character. Genetics **142:** 285–294.

Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein, 1998 Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95:** 14863–14868.

Ewing, R. M., A. B. Kahla, O. Poirot, F. Lopez, S. Audic *et al.*, 1999 Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res. **9:** 950–959.

Ferea, T. L., D. Botstein, P. O. Brown and R. F. Rosenzweig, 1999 Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc. Natl. Acad. Sci. USA **96:** 9721–9726.

Fisher, R. A., 1935 *The Design of Experiments*, Ed. 3. Oliver & Boyd, London.

Good, I. P., 2000 *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypothesis*. Springer, New York.

Gordon, A. D., 1999 *Classification*. Chapman & Hall, London.

Hartigan, J. A., 1975 *Clustering Algorithms*. Wiley, New York.

Iyer, V. R., M. B. Eisen, D. T. Ross, G. Schuler, T. Moore *et al.*, 1999 The transcriptional program in the response of human fibroblasts to serum. Science **283:** 83–87.

Johnson, R. A., and D. W. Wichern, 1998 *Applied Multivariate Statistical Analysis*, Ed. 4. Prentice-Hall, Englewood Cliffs, NJ.

Kerr, M. K., and G. A. Churchill, 2001 Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc. Natl. Acad. Sci. USA **98** (16): 8961–8965.

Kim, K., M. A. L. West, R. W. Michelmore, D. A. St.Clair and R. W. Doerge, 2005 Old methods for new ideas: dissection of the determinants of gene expression levels. Proceedings of the Stadler Genetics Symposium, Columbia, MO.

Knudsen, S., 2002 *A Biologist's Guide to Analysis of DNA Microarray Data*. Wiley, New York.

Lan, H., J. P. Stoehr, S. T. Nadler, K. L. Schueler, B. S. Yandell *et al.*, 2003 Dimension reduction for mapping mRNA abundance as quantitative traits. Genetics **164:** 1607–1614.

Mangalam, H., J. E. Stewart, J. Zhou, M. Waugh, K. Schlauch *et al.*, 2001 GeneX: an open source gene expression database and integrated toolset. IBM Syst. J. **40:** (2): 552–569.

McShane, L., R. Radmacher, B. Freidlin, R. Yu, M.-C. Li *et al.*, 2002 Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics **18** (11): 1462–1469.

Munneke, B., 2001 Null model methods for cluster analysis of gene expression data. Ph.D. Thesis, Department of Statistics, Purdue University, West Lafayette, IN.

Nettleton, D., and R. W. Doerge, 2000 Accounting for variability in the use of permutation testing to detect quantitative trait loci. Biometrics **56:** 285–291.

Perou, C. M., S. S. Jeffrey, M. van de Rijn, C. A. Rees, M. B. Eisen *et al.*, 1999 Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc. Natl. Acad. Sci. USA **96:** 9212–9217.

Ronald, J., J. M. Akey, J. Whittle, E. N. Smith, G. Yvert *et al.*, 2005 Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. Genome Res. **15** (2): 284–291.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse, and man. Nature **422:** 297–302.

Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan *et al.*, 1999 Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA **96:** 2907–2912.

Tavazioe, S., J. D. Hughes, M. J. Campbell, R. J. Cho and G. M. Church, 1999 Systematic determination of genetic network architecture. Nat. Genet. **22:** 281–285.

Tibshirani, R., G. Walther and T. Hastie, 2001 Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B Stat. Methodol. **63:** 411–423.

Zhang, K., and H. Zhao, 2000 Assessing reliability of gene clusters from gene expression data. Funct. Integr. Genomics **1:** 156–173.

Communicating editor: J. B. Walsh