

New methods for identifying significant clusters in gene expression data

Katherine S. Pollard and Mark J. van der Laan

Div. of Biostatistics, Univ. of California, Earl Warren Hall #7360, Berkeley, CA 94720

Key Words: clustering; silhouette; homogeneity; significance; gene expression.

1. Motivation

An important goal with large-scale gene expression studies is to find biologically important subsets of genes and samples. Clustering algorithms have been widely applied to this problem. These can be classified into partitioning and hierarchical clustering algorithms. Examples of hierarchical algorithms include agglomerative clustering (as implemented in the Cluster program by Eisen et al.) and HOPACH (van der Laan and Pollard [2001]). Partitioning algorithms include K-Means, Self-Organizing Maps (Törönen et al. [1999]), and Partitioning Around Medoids (Kaufman and Rousseeuw [1990]). With both types of algorithms, we are interested in the number of clusters. In a hierarchical tree, this corresponds with the lowest level at which the clusters are still significant. Methods for determining the number of clusters are reviewed and compared by Milligan and Cooper and Fridlyand and Dudoit, who claim that none of the existing methods are satisfactory for gene expression data analysis. We have also noted the need for a better method in the gene expression context, where small, biologically meaningful clusters can be difficult to identify. In particular, existing methods tend to identify only the global structure in the data, for example, the over and under expressed genes.

2. Clustering Gene Expression Data

A gene expression experiment results in an observed data matrix X whose columns are n copies of a p -dimensional vector of gene expression measurements, where n is the number of observations and p is the number of genes. The genes are a set of p elements \mathbf{x}_j , $j \in \{1, \dots, p\}$, where each element \mathbf{x}_j is an n dimensional vector $(x_{1j}, \dots, x_{nj})^T$. Typically, X_{ij} measures the abundance of mRNA for gene j in sample i relative to a control. The data is usually preprocessed and then screened to eliminate certain genes, such as those showing no difference in expression, from the subset. Then, the genes and/or the samples are clustered.

Clustering algorithms are (either implicitly or explicitly) functions of a dissimilarity matrix which measures the distance between every pair of elements. Let $d(\mathbf{x}_j, \mathbf{x}_{j'})$ denote the dissimilarity between elements j and j' and let \mathbf{D} be the $p \times p$ symmetric matrix of dissimilarities. Typical choices of dissimilarity include Euclidean distance, 1 minus correlation, 1 minus absolute correlation and 1 minus cosine-angle. Partitioning methods generally require that the user specify the number of clusters, whereas hierarchical methods produce a tree of clusters with sequentially more clusters as one moves from top to bottom. With both types of methods, identifying a main clustering result corresponds with choosing the number of clusters.

Methods for selecting the number of significant clusters include direct methods and testing methods. Direct methods consist of optimizing a criteria, such as functions of the within and between cluster sums of squares (Milligan and Cooper [1985]), occurrences of phase transitions in simulated annealing (Rose et al. [1990]), likelihood ratios (Scott and Simmons [1971]), or average silhouette (Kaufman and Rousseeuw [1990]). The method of maximizing average silhouette has the advantage of being able to be used with any clustering routine and any distance metric. A disadvantage of average silhouette is that, like many criteria functions for selecting the number of clusters, it measures the global structure only. We discuss this problem in more detail in Section 3.. Testing methods take a different approach, assessing evidence against a specific null hypothesis. Examples of testing methods that have been used with gene expression data are the gap statistic (Tibshirani et al. [2000]), the weighted average discrepant pairs (WADP) method (Bittner et al. [2000]), a variety of permutation methods (eg: Bittner et al. [2000], Hughes et al. [2000]), and Clest (Fridlyand and Dudoit [2001]). Testing methods involve permutations or resampling so that they are computationally much more difficult than direct methods.

In this paper, we present new methods for selecting the significant clusters. In Section 3., we present a general method for identifying optimally homogeneous clusters and then illustrate the method with a specific criteria function, Mean Split Silhou-

ette (MSS). In Section 4., we propose an alternative method based on testing for statistical significance and then demonstrate the method with a specific test statistic, average distance. We discuss the relative merits of each approach and the power of both approaches compared to existing methods.

3. A New Direct Method

We have found many cases, in both real and simulated gene expression data, where existing direct methods for selecting the number of clusters fail to find the main clusters. The problem of finding relatively small clusters in the presence of one or more larger clusters is particularly hard. Another challenging problem arises when the clusters are not equally distant from each other, but rather form nested clusters within clusters. This type of data structure arises frequently in gene expression data. It is frequently this finer structure that is of interest biologically, but current methods find only the global structure. Inspired by this lack of performance, we present a new direct method for selecting the number of clusters which can be applied with both partitioning and hierarchical clustering algorithms.

3.1 General Method

Consider a series of proposed clustering results and a global criteria function. With a partitioning algorithm, these may consist of applying the clustering routine with $k = 2, 3, \dots, K$ clusters where K is a user-specified upper bound on the number of clusters. With a hierarchical algorithm, the series may correspond to levels of the tree. In either case, evaluate each proposed result separately using the following method. Apply the clustering routine independently to the elements in each of the clusters and then evaluate the criteria function to obtain a measure of cluster heterogeneity for each cluster. Average this measure over clusters. Repeat the procedure for each of the proposed clustering results in the series. The minimum indicates the clustering result with most homogeneous clusters. The key idea behind the method is to evaluate how well the elements in a cluster belong together by diving into each cluster and applying the clustering algorithm and criteria function to the elements in that cluster alone, ignoring the other clusters.

3.2 Mean Split Silhouette (MSS)

We present a particular application of this method called Mean Split Silhouette (MSS), which uses silhouette as the criteria. Since silhouettes can be calculated with any clustering algorithm and any distance metric, MSS can be used to determine the

number of clusters with all partitioning and hierarchical clustering algorithms. Suppose we are clustering genes.

3.2.1 Silhouette

The silhouette for a given gene is calculated as follows (Kaufman and Rousseeuw [1990]). For each gene j , calculate a_j which is the average dissimilarity of gene j with other elements of its cluster:

$$a_j = \text{avg } d(\mathbf{x}_j, \mathbf{x}_{j'}), j' \in \{i : l_1(\mathbf{x}_i, M) = l_1(\mathbf{x}_j, M)\}.$$

For each gene j and each cluster k to which it does not belong (that is, $k \neq l_1(\mathbf{x}_j, M)$), calculate b_{jk} , which is the average dissimilarity of gene j with the members of cluster k :

$$b_{jk} = \text{avg } d(\mathbf{x}_j, \mathbf{x}_{j'}), j' \in \{i : l_1(\mathbf{x}_i, M) = k\}.$$

Let $b_j = \min_k b_{jk}$. The silhouette of gene j is defined by the formula:

$$S_j(\mathbf{M}) = \frac{b_j - a_j}{\max(a_j, b_j)}. \quad (1)$$

Note that the largest possible silhouette is 1, which occurs only if there is no dissimilarity within gene j 's cluster (*i.e.*: $a_j = 0$). The other extreme is -1. Heuristically, the silhouette measures how well matched an object is to the other objects in its own cluster versus how well matched it would be if it were moved to the next closest cluster.

3.2.2 Average Silhouette

The average silhouette over all elements has been used to evaluate and compare clustering results, including selecting the number of clusters k by maximizing average silhouette over a range of possible values for k (Kaufman and Rousseeuw [1990]). It has been our experience that average silhouette is actually a very good global measure of the strength of clustering results: see also [Fridlyand, 2001] for a favorable performance of average silhouette relative to other validation functionals. As we have argued, however, it is important to go beyond global structure in the analysis of gene expression data. Average silhouette alone is not able to identify this finer structure.

3.2.3 MSS

Given a clustering result with k clusters, consider splitting each cluster into two or more clusters (the number of which can be determined, for example, by maximizing average silhouette). In the hierarchical tree context, this corresponds with computing the child clusters in the next level of the tree, while in

the partitioning context it corresponds with treating the elements in each cluster as a new data set and partitioning them. In both cases, each element has a new silhouette after the split, which is computed relative to only those elements with which it shares a parent. We call the average of these for each parent cluster the split silhouette $SS_i, i = 1, 2, \dots, k$. The split silhouette is a measure of that cluster’s homogeneity (i.e.: it is low if the cluster was homogeneous and should not have been split). We define MSS as the mean of the split silhouettes over the k clusters:

$$MSS(k) = \frac{1}{k} \sum_{i=1}^k SS_i. \quad (2)$$

Then, MSS is a measure of the average homogeneity of the clusters in the clustering result. All of the means can be replaced with medians for a more robust criteria. In fact, we frequently prefer the results found using Median Split Silhouette.

3.2.4 Choosing the Number of Clusters

Given a series of clustering results, we propose to choose the proposed clustering result which minimizes MSS. In this way, we choose the number of significant clusters that produces (on average) the most homogeneous groups. One nice benefit of this approach is that it is possible to select one cluster (i.e.: no groups) without using a testing approach and defining a null distribution. Unlike most global criteria, MSS is defined for $k = 1$; it is in fact the usual average silhouette. If the data is homogeneous, the minimum MSS will occur at $k = 1$, as illustrated in Figure 1.

3.3 Performance of MSS

We have previously reported simulation results for MSS on different data sets and relative to other direct methods (Pollard and van der Laan [2002a]). Here, we summarize a few results. First, MSS can be used to identify both clusters of genes and clusters of samples (using all genes or gene cluster profiles). Second, as the noise level in the data increases for a fixed sample size, MSS identifies fewer clusters, and the clusters correspond well with groups that are visible in the distance matrix. Third, unlike most direct criteria, MSS is defined for one cluster and therefore can be used to assess whether or not there is evidence of groups in a data set. Figure 1 shows that the minimum of MSS is indeed at one cluster when we can not visually distinguish any groups in the distance matrix. Finally, in a comparison with other direct methods on simulated data with nested clusters, MSS is best able to identify the smaller

clusters and is also less variable than other methods.

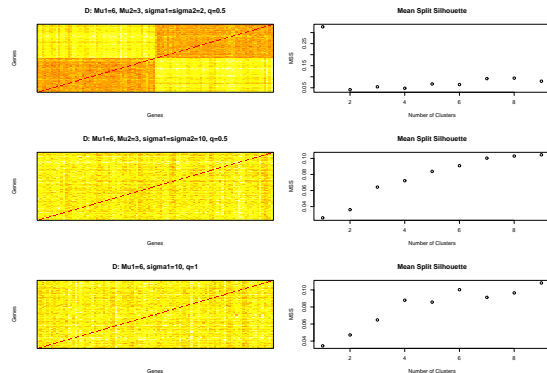


Figure 1: Distance matrices (left panels) and plots of MSS for a range of number of clusters k (right panels). Data simulated from normal mixture distributions with $n = 30$ samples, $p = 100$ genes, mixing parameter q , means (μ_1, μ_2) and diagonal covariance matrices with standard errors (σ_1, σ_2). Distances plotted on a red-white scale with bright red corresponding to the smallest distance. The elements are ordered in the distance matrix according to which component of the mixture they came from. The top row is a well separated mixture in which we can see two clusters in the distance matrix. The middle row is an overlapping mixture in which we can not visually distinguish the clusters in the distance matrix. The bottom row is a single normal distribution. The plots of MSS show that the criteria chooses two clusters in the top row and one in the other two cases.

4. New Testing Methods

Direct methods have the advantage of being computationally easy since they simply select the number of clusters optimizing the value of a criteria function on the observed data. They do not, however, have a statistical interpretation. In order to identify clusters that are *statistically* significant, one can take a testing approach. Testing methods involve defining a specific null hypothesis, selecting a test statistic, and evaluating the observed value of the statistic relative to an appropriate null distribution using a precisely defined type of error control. In order to make accurate inferences, the null distribution needs to respect the sampling framework of the data. Furthermore, the method should identify more clusters as the number of samples n increases (for a fixed noise level). Finally, if many comparisons are being made, appropriate adjustment should be made for

the multiplicity. Most testing methods in the literature, however, do not possess these three important properties.

4.1 Testing Method for Gene Clustering

Our goal is to test the null hypothesis that a group of genes is homogeneous. We first present a general method for assigning a p-value to any group of genes. Then, we discuss specific choices of test statistic and null distribution in more detail. Next, we comment on how to appropriately adjust the procedure for multiple testing. Finally, we present some preliminary simulation results illustrating how the procedure performs.

4.1.1 General Method to Assign a P-value to a Cluster

Consider a group of genes, either one of k partitions or a node in a hierarchical tree. First, we select a test statistic of interest, such as the average of all pair-wise distances, which is a measure of cluster homogeneity. We evaluate the value of the statistic on the observed data and obtain an estimate T_n . Next, we must compute the distribution of this test statistic under an appropriate null distribution. For example, we might simulate a large number B of data sets from a multivariate normal distribution with all genes having mean zero and the observed gene-by-gene covariance matrix. By calculating the test statistic T^b for each simulated data set $b = 1, \dots, B$, we obtain a null distribution for the statistic, which we can use to compute a p-value for T_n :

$$p = \frac{1}{B} \sum_{b=1}^B I\{T_n > T^b\}, \quad (3)$$

where $I\{\cdot\}$ is the indicator function which equals one if the expression is true. If we repeat this procedure for many clusters, then we need to adjust the p-values for multiple comparisons by applying a procedure to these marginal p-values which guarantees a particular multiple testing Type-I error control, as described in Section 4.1.4.

We propose to increase the number of clusters until none of the adjusted p-values are significant at a pre-specified level α . For a partitioning algorithm, this means increasing the number of partitions. In a hierarchical tree, any node with a significant p-value should be split further (ie: examine its children in the next level of the tree).

4.1.2 Test Statistic

This general method can be applied with any choice of test statistic. The idea is to select a statistic that measures cluster homogeneity and that has a

null distribution which can be easily estimated for the purpose of testing. Possible test statistics include the average distance and MSS. Average distance has the advantage of being much easier to compute. Many other test statistics are possible and it is an area of future research to determine those that are best suited for gene cluster testing.

4.1.3 Null Distribution

Computing a p-value for a cluster requires a good estimate of the null distribution of the test statistic. Since it is important to use the joint distribution in the gene expression context (where many genes are expected to be correlated), we do not recommend using tabled distributions for each gene's marginal distribution. Instead, we simulate the joint null distribution. One simulation method is to assume a parametric distribution and estimate the parameters from the observed data. Another method is to use the empirical distribution, perhaps with a smoothing technique (eg: resampling $2n$ out of n observations) to avoid problems associated with the discreteness of the distribution. In both cases, we set the mean of the null distribution equal to zero (or the empirical mean) so that all genes have the same mean. The null of equal means works well with Euclidean or cosine-angle distance, since these distance metrics are mostly responding to differences in mean. If we use correlation as our distance metric, then we might choose as null distribution a normal with diagonal covariance matrix. Thus, the particular choice of null distribution is application specific.

We have conducted some simulations comparing different choices of null distributions with the average distance statistic and Euclidean distance. We favor the multivariate normal with zero means and the empirical covariance matrix, since the empirical distribution tends to over estimate the covariance in the data and hence slightly under estimates the average distance. Because the variance of average distance is quite small, this small downward bias can have a significant affect on p-values computed from the empirical null distribution. Smoothing the empirical distribution does not remove this bias.

4.1.4 Multiplicity Adjustment and Error Control

This method can be applied to any group of genes. Usually, we will apply it to all the clusters from a partitioning algorithm or all the nodes in one level of a tree. Suppose we are applying the method to k clusters. Then, we need to adjust the marginal p-values for multiple comparisons. An easy to compute method is the Bonferroni adjustment which divides each p-value by k . When the

clusters are not independent, this method is conservative. Many other methods have been proposed in the multiple testing literature (eg: Hochberg and Tamhane [1987], Westfall and Young [1993], Benjamini and Hochberg [2002]). In addition to selecting a multiplicity adjustment method, it is also important to precisely define what type I error rate you wish to control. For example, one might choose the False Discovery Rate (expected proportion of false rejections among all rejections) or the Per Family Error Rate (expected number of false rejections), rather than the usual Family-Wise Error Rate (probability of at least one false rejection).

4.1.5 Performance of the Testing Method

We have applied this testing method to gene clustering of simulated data using both hierarchical (HOPACH, van der Laan and Pollard [2001]) and partitioning (Partitioning Around Medoids (PAM), Kaufman and Rousseeuw [1990]) methods. In our implementation, we use average distance as the test statistic and a multivariate normal null distribution, which is particularly appropriate since we simulate the data from a multivariate normal ($n = 20$). The $p = 512$ genes are independent and have one of eight different means $\mu = (1, 2, 5, 6, 14, 15, 18, 19)$, which form a nested series of eight clusters. We use Euclidean distance, since this is the appropriate metric for detecting differences in means. With PAM, as we increase the number of clusters from $k = 1, \dots, 10$, $k = 8$ is the first clustering result for which all of the clusters are not significantly more heterogeneous than expected under the null. In the HOPACH tree, the adjusted p-values for each node also suggest that we should keep splitting until we have $k = 8$ clusters.

4.2 Testing Method for Sample Clustering

Clustering samples corresponds with dividing the observed samples into subpopulations. The independent sampling unit is still a sample (experiment or patient) on which we observe a p -dimensional gene expression vector. Hence, we can not simply invert the data matrix, treat genes as observations, and use a method that would be appropriate for gene clustering. The null distribution of the data is the same distribution described for gene clustering (eg: multivariate normal with equal means and empirical covariance matrix or the empirical distribution centered at zero). The testing procedure is, however, different.

4.2.1 Multiple Testing Approach

For a patient clustering result with k clusters, we propose to view testing the significance of the clusters as a k sample problem or as a collection of two

sample problems (corresponding to all pairs of subpopulations). Consider, for example, two clusters of samples ($k = 2$). For each gene, we compute a test statistic (eg: t-statistic). If we use family-wise error control then we would reject the null hypothesis that the two subpopulations come from the same distribution if for at least one gene the adjusted p-value exceeds a pre-specified error level α . In other words, we use the maximum test statistic over all genes as the statistic to compare the subpopulations. Alternatively, we could use the mean test statistic over all genes or require that a pre-specified proportion of all test statistics be rejected. A p-value for the observed statistic can be calculated using the appropriate null distribution (eg: empirical distribution with all subpopulation means set equal). We declare the two clusters significantly different if the p-value for the split is sufficiently small. Appropriate adjustments must be made for multiple comparisons when more than one split is examined. This method can be easily generalized to $k > 2$.

To apply this approach in a hierarchical tree, consider each node of the tree one at a time (beginning at the top) and continue splitting any node with a significant adjusted p-value. With a general partitioning algorithm, treat the whole sample in the same way as one node in a tree. In both cases, the number of significant clusters is defined as the maximum number for which the defined subpopulations are *all* statistically different (for a specific null hypothesis, test statistic and error rate control).

5. Conclusions

We have proposed several new approaches to identifying significant clusters in gene expression data. Both our direct and our testing methods can be used with partitioning and hierarchical algorithms. Our direct method, MSS, finds the set of maximally homogeneous clusters. This criteria function is better able to identify small clusters than other direct methods in the literature. Small clusters, often nested within larger clusters (eg: the over and under expressed genes), are of particular interest to biologists, since these some times represent distinct causal mechanisms. Although we are quite satisfied with the performance of MSS on real and simulated data, we feel that it is some times important that clusters be significant statistically. In this case, it is necessary to use a testing approach. Here we propose testing methods for gene and sample clustering that (i) respect the sampling framework of the data, (ii) identify more clusters as the number of samples n increases (or the noise level decreases), and (iii) account for multiple comparisons. Addressing

these three issues improves upon currently proposed methods. For example, using permutation methods for evaluating sample clustering breaks the within subpopulation covariance structures, which can have serious implications for null distribution estimation when the subpopulation sample sizes and/or covariance structures differ (Pollard and van der Laan [2002b]). By resampling from a null distribution that resembles the observed distribution except that the genes have equal means, we are able to perform tests for each gene without making any assumptions about the equality of the covariances.

We have presented direct and testing methods as general approaches, offering specific implementations for illustration. Additional simulations and data analyses are needed to fully understand both methods and to assess different implementations (eg: choices of test statistics and error rates for testing).

References

- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statistical Society*, 12: 7–29, 2002.
- M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536–540, 2000.
- M. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95: 14863–14868, 1998.
- J. Fridlyand. Ph.d thesis. Department of Statistics, UC Berkeley, 2001.
- J. Fridlyand and S. Dudoit. Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. Technical Report 600, Statistics Department, University of California, 2001.
- Y. Hochberg and A.C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, 1987.
- T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, Lum P.Y., S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburttty, J. Simon, M. Bard, and S.H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.
- L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, 1990.
- G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985.
- K.S. Pollard and M.J. van der Laan. A method to identify significant clusters in gene expression data. In *SCI2002 Proceedings*, volume II, pages 318–325. International Institute of Informatics and Systemics, 2002a.
- K.S. Pollard and M.J. van der Laan. Resampling-based methods for identification of significant subsets of genes in expression data. Technical report, Group in Biostatistics, University of California, 2002b. Submitted.
- K.S. Pollard and M.J. van der Laan. Statistical inference for simultaneous clustering of gene expression data. *Mathematical Biosciences*, 176(1):99–121, 2002c.
- K. Rose, E. Gurewitz, and G.C. Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65:945–948, 1990.
- A.J. Scott and M.J. Simmons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27: 387–397, 1971.
- R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. Technical report, Department of Statistics, Stanford University, March 2000.
- P. Törönen, M. Kolehainen, G. Wong, and E. Castrén. Analysis of gene expression data using self-organizing maps. *FEBS Letters*, 451:142–146, 1999.
- M.J. van der Laan and K.S. Pollard. Hybrid clustering of gene expression data with visualization and the bootstrap. Technical Report 93, Group in Biostatistics, University of California, May 2001. To appear in JSPI.
- P.H. Westfall and S.S. Young. *Resampling-based Multiple Testing: Examples and methods for p-value adjustment*. John Wiley & Sons, 1993.