Conference on Applied Statistics in Agriculture 2006 - 18th Annual Conference Proceedings

CLUSTERING A SERIES OF REPLICATED POLYPLOID GENE EXPRESSION EXPERIMENTS IN MAIZE

Lingling An

Nicole C. Riddle

James A. Birchler

R. W. Doerge

See next page for additional authors

Recommended Citation

An, Lingling; Riddle, Nicole C.; Birchler, James A.; and Doerge, R. W. (2006). "CLUSTERING A SERIES OF REPLICATED POLYPLOID GENE EXPRESSION EXPERIMENTS IN MAIZE," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1120

Author Information
Lingling An, Nicole C. Riddle, James A. Birchler, and R. W. Doerge

CLUSTERING A SERIES OF REPLICATED POLYPLOID GENE EXPRESSION EXPERIMENTS IN MAIZE

Lingling An[1], Nicole C. Riddle[2*], James A. Birchler[2] and R.W. Doerge[1,3]

[1]Department of Statistics, Purdue University, West Lafayette, IN, 47907
[2]Division of Biological Sciences, University of Missouri, Columbia, MO 65211
[3]Department of Agronomy, Purdue University, West Lafayette, IN, 47907

* Present address: Department of Biology, Washington University, St. Louis, MO 63130

Abstract

Ploidy level is defined as the number of individual sets of chromosomes contained in a single cell. Many important crop plants, such as potato, soybean and wheat are polyploid. Although it is widely known that polyploidy is a frequent evolutionary event, it is not fully understand why polyploids have been so successful. In this work cluster analysis is employed to study gene expression changes in a maize inbred line (B73) across a range of polyploidy levels. The B73 ploidy series includes monoploid, diploid, triploid and tetraploid plants and consists of biological and technical replicates as measured by microarray technology. An improved version of CORE (iCORE; improved Clustering of Repeat Expression) is presented to differentiate highly negatively correlated genes while taking advantage of the additional information that is provided by replication. The error information from the replicate experiments is utilized to cluster gene expression for both simulated and real ploidy-series data. Simulation results indicate that iCORE leads to an improvement in accuracy over both CORE and hierarchical clustering based on average gene expression only. When applied to the maize ploidy series, the iCORE results provide information that may aid in understanding of the effect of gene dose on gene expression in a ploidy series.

Keywords: clustering, CORE, repeated measurements, ploidy series, microarray data

1. Introduction

Cluster analysis is an exploratory data analysis tool that seeks to partition a sample into homogeneous groups, such that the observations within a cluster are more similar than observations in different clusters. Since microarray technology allows researchers to monitor the expression of tens of thousands of genes simultaneously, applying cluster analysis to gene expression data provides groupings of genes that may share similarity in their regulation. Based on known functions of genes in the same cluster, researchers may be able to predict the function of an unknown gene (i.e., genes residing in the same cluster). Typically, gene expression profiles are clustered across treatments or conditions, or a series of conditions, such as a time series

(Luan et al. 2003; Ma et al. 2006; Peddada et al. 2003). When clustering is applied to data from a series, the premise is that genes sharing similar expression profiles might be functionally related or co-regulated and that these shared expression profiles reflect the existence of natural dependencies among the elements/ time points of the series.

There are a variety of algorithms that have been employed to cluster gene expression data. In most cases, however, cluster analysis is performed without taking advantage of the error information that is sometimes available from experimental replication. When in fact replicated measurements do exist for microarray data, the simplest way to proceed is to compute the average expression of the replicated experiments for each gene and then employ traditional clustering methods (e.g., K-means, hierarchical clustering). However, averaging each gene across replicates does not take into account the variability of the gene's expression measurement. The goal of this work is develop a method for clustering gene expression profiles according to patterns of expression, while incorporating information that is gained from replication. Our expectation is that the inclusion of the additional error information will increase the accuracy of the cluster analysis simply because the approach will be less sensitive to the underlying noise in the data.

There are several clustering methods that incorporate the error information into their analysis. Specifically in model-based clustering, a mixture model approach (Medvedovic et al. 2002; Yeung et al. 2003) assumes that each cluster follows a multivariate normal distribution. The advantage of this method is its generality. However, its disadvantage is that it considers the magnitude of the gene expression profile rather than the pattern. Another clustering approach, the error-weighted similarity (Hughes et al. 2000) method, considers error estimates when calculating the pairwise similarities such that expression values with more error or variation are down-weighted. The advantage of this approach is its ease of implementation since it is based on a modified (error-weighted) similarity measure that can be used in any traditional clustering algorithm. Unfortunately, the error-weighted similarity approach to clustering only captures experiment-specific variation, not gene-specific variation.

An extension of the K-means clustering algorithm (Tavazoie et al. 1999) which includes error information while clustering gene expression data is known as CORE (Clustering of Repeated Expression data; Tjaden 2006). CORE captures much of the noise in the underlying data at both the experiment and gene levels. Its main advantage over other clustering methods is that noisy measurements can be down-weighted to achieve more accurate clusterings. However, the algorithm is limited in the sense that it cannot differentiate between groupings of the highly negatively correlated gene expression patterns (e.g., equal and opposite gene profiles are clustered together, as illustrated in Figure 1). Using this as motivation, an improvement of the CORE (iCORE) algorithm is proposed for the purpose of distinguishing the negatively correlated genes while incorporating error information from microarray series data. Using simulated data the performance of iCORE is compared to the performance of both CORE and hierarchical clustering (Eisen 1998) based on average expression profiles. Finally, iCORE, CORE, and hierarchical clustering are applied to a maize ploid series to assess the effects of including both error information and differentiation of negatively correlated expression profiles.

2. Polyploidy

Ploidy refers to the number of individual sets of chromosomes contained in a single cell. An organism is considered polyploid if it has more than two chromosome sets in a single cell. The phenomenon of polyploidy is very common in plants, yet very little is known about the benefits of being a polyploid, or why polyploids have been so successful evolutionarily (Adams and Wendel 2005). One reason for the success and continuation of polyploids is the novel variation that is created by the polyploidization event itself. Maize is an ancient polyploid that has most likely benefited greatly from being polyploid. The motivation in studying different levels of polyploid maize is to explore the hypothesis that gene expression is additive (i.e., 1+ 1=2). In other words, as the genome undergoes polyploidization gene dosage increases in an additive manner, but the affect of increased gene dosage on gene expression is unknown. For example, in a monoploid there is one copy of every gene. In a diploid individual, there are two copies of every gene, while in a triploid and tetraploid there are three and four copies of every gene, respectively. Since microarray technology allows the simultaneous investigation of every gene in a genome, it can be used to explore gene dosage effect on gene expression levels when comparing ploidy levels. If gene expression like gene dosage is additive, one would expect all genes to behave in an additive manner, with gene expression levels increasing with ploidy level.

3. Methods

Model

Consider *n* genes that are represented on a microarray, and *m* experiments. For each of these *n* genes across *m* experiments both the mean expression value and the associated error from the biological and/or technical replicates can be estimated. As mentioned previously, CORE (Tjaden 2006) is a general case of K-means clustering where the number of clusters (k_0) is assumed known. The CORE approach is explained as follows. Assume that each gene's average expression profile comes from a single cluster profile (i.e., each gene can only belong to one cluster). Letting y_{ij} be the expression value for gene i in experiment j, and σ_{ij} be the associated error, the expression profile for gene i from cluster profile A_k (= (A_{k1}, …, A_{km})) can be expressed as,

$$y_{ij} = B_i A_{kj} + C_i + \delta_{ij}, \qquad (1)$$

where δ_{ij} is the error term that is distributed $N(0, \sigma_{ij}^2)$. The parameters B and C are the scaling factor and translation factor, respectively. The objective of clustering is to group these *n* gene profiles into k_0 clusters while accounting for the error information. Specifically, gene expression profiles with low errors will be given high priority (weight) when clustered. Given the data, the scaling factor B, translation factor C, and cluster profile A_k are estimated for each gene and clustered (further details on clustering follow).

iCORE algorithm

The CORE algorithm (Tjaden 2006) is modified to accommodate situations where a gene, or a group of genes, has/have opposing expression profiles (Figure 1). An implementation of the CORE algorithm to the data that provide the profiles in Figure 1 results in the opposing gene profiles (shown in black) being placed in the same cluster (i.e., one gene is up-regulated while the other gene is down-regulated). The iCORE algorithm acknowledges the direction of the gene expression differences from experiment to experiment by placing such profiles in unique clusters.

The iCORE algorithm (as detailed in Figure 2) has the same foundation as the CORE algorithm in that an iterative process places the expression profiles into clusters based on the total sum of within-cluster distances. Initially, k_0 clusters are assumed (determining k_0 for a real data analysis will be discussed later) and all genes are randomly assigned into k_0 clusters. After the random assignment of genes to clusters, the cluster profile A_k for each cluster is estimated using an iterative process where A_k is initialized as the mean profile of all genes currently belonging to that cluster. Once A_k is estimated, then the scaling factor B and translation factor C (Equation (1)) for the genes in the cluster, are estimated (see Figure 2). Given the estimates of B and C a new cluster profile A_k is estimated. This process is iterated until the cluster profile A_k converges.

Once the A_k estimates are obtained (for the current iteration) for all clusters, each gene is reassigned to its closest cluster using a distance calculation. In the original CORE algorithm, the distance of a gene i to a cluster k is represented as a ratio of how well a gene's expression profile across *m* experiments is described by a cluster profile with linear transformation parameters B and C, and the level of variation observed for that gene's expression at each of the *m* experiments:

$$\sum_{j=1}^{m}\left[\frac{y_{ij}-(B_i A_{kj}+C_i)}{\sigma_{ij}}\right]^2. \qquad (2a)$$

The improved CORE (iCORE) algorithm subtly alters Equation (2a) by taking the absolute value of the scaling factor B. In doing so, a gene is less likely to be assigned to a cluster whose pattern is opposing (Figure 3).

$$\sum_{j=1}^{m}\left[\frac{y_{ij}-(|B_i|A_{kj}+C_i)}{\sigma_{ij}}\right]^2 . \qquad (2b)$$

In either the CORE or iCORE algorithm once the nearest cluster for each gene is obtained, the total sum of within-cluster distance is compared to the previous iteration's total sum of within-cluster distance to determine if the algorithm can terminate, otherwise the next iteration results are obtained and the process continues until the total sum of within-cluster distances converge.

4. Performance of Clustering Methods Using an Adjusted Rand Index

In order to compare clustering results against external criteria, or evaluate the consistency of two clustering results on the same data set, a measure of agreement is needed. The Rand index (Rand 1971) is a statistic that indicates the degree of agreement between two sets of clusterings (U and V) derived from the same data set, and can be used to compare the results from the two clustering methods. An assessment, in a pairwise manner, as to whether the profile of two genes in partition U (may, or may not be in the same group) fall in the same group under the partition V can be made. If the two genes do not belong to the same group, then other possibilities are considered. Table 1 illustrates the numbers of pairs of objects in the four possible categories when there are two paritions.

The Rand index (Rand 1971) ranges from 0 to 1 and is defined as the fraction of pairwise comparisons that are in agreement $\frac{a+d}{a+b+c+d}$. A high Rand index indicates a high level of agreement (e.g., the Rand index is 1 when two sets of partitions are in perfect agreement). One limitation of the Rand index is that its expected value for two random partitions is not a constant value (for example, zero). The adjusted Rand index as proposed by Hubert and Arabie (1985) adjusts the score so that its expected value takes the value 0 in the case of random partitions. The adjusted Rand index is:

$$\frac{a-\dfrac{(a+b)(a+c)}{(a+c+c+d)}}{\dfrac{2a+b+c}{2}-\dfrac{(a+b)(a+c)}{(a+b+c+d)}}=\frac{2a(a+b+c+d)-2(a+b)(a+c)}{(2a+b+c)(a+b+c+d)-2(a+b)(a+c)}.$$

Milligan and Cooper (1986) illustrated that even for two partitions with different numbers of clusters the adjusted Rand index maintains its properties. Furthermore, when compared to other clustering validation measures (Milligan and Cooper 1986; Monti et al 2003) the adjusted Rand index outperforms its competitors. Using simulated data with known clusters, the adjusted Rand index can be employed to compare the results from different clustering methods and to evaluate their performance. Thus, computing a Rand index for each method by comparing the results to the known "true" simulated clusters, allows for the assessment of accuracy and a comparison between methods.

5. Simulated Data and Results

Simulated data are used to test the accuracy of the iCORE algorithm. Gene expression values for 100 genes over 4 experiments residing into 10 clusters are simulated via Equation (1). This set-up has the same number of experiments (ploidy levels) as the experimental data that motivated this work. However, it should be noted that in real applications, the number of clusters is typically unknown and has to be estimated prior to clustering; this will be discussed in the next section.

Ten cluster profiles, each from a 4-dimensional unit hypercube, were generated. The correlation of all pairwise profiles is controlled to be less than 0.95. Factors B and C are random values between -1 and 1. The error term is distributed as a normal distribution with mean 0, and the variance is gained from a Chi-squared distribution with 30 degrees of freedom. The Chi-square distribution is appropriate for normally distributed errors (Tjaden 2006). One hundred different data sets were simulated using the same parameter settings. Three clustering methods, iCORE, CORE, and hierarchical clustering were employed to identify groupings of the 100 genes for each of the 100 simulated data sets. When implementing hierarchical clustering, the average gene profile was used as the data, and Pearson's correlation and the average linkage function used for the clustering algorithm. The average adjusted Rand index was calculated, and the three clustering methods were compared to the known simulation setting across a range of assumed cluster numbers (5-20). Even though the number of clusters is known for the simulated data, a range of cluster number is considered for the purpose of studying cluster number effect. At the true/known cluster number (10), a level of accuracy for each clustering of the three methods is obtained.

As illustrated in Figure 4 when iCORE, CORE, and hierarchical clustering are compared across different possible cluster numbers via the average adjusted Rand index, the iCORE algorithm outperforms both CORE and hierarchical clustering. When the accuracy of each method is compared at the true cluster number (10), the average adjusted Rand index value is highest for iCORE, while the hierarchical clustering method has the lowest value. The poor performance of hierarchical clustering is not surprising since the variability of each gene across repeated experiments is not taken into account. In other words, the additional information from replication is lost.

6. Clustering a Maize Polyploid Series

Maize is an ancient polyploid that has undergone recent diploidization. Understanding gene expression at different ploidy levels (e.g., monoploid, diploid, triploid, and tetraploid) is of interest from an evolutionary standpoint since there appears to be some benefit to being a polyploid. In addition, understanding gene action at different ploidy levels (i.e., 1X, 2X, 3X, 4X) is interesting from a genetic point of view because the effect of gene dosage on gene expression remains unknown. The inbred maize line B73 was used in this study, with monoploid, triploid and tetraploid individuals being derived from a diploid progenitor. Microarray experiments were conducted to facilitate gene expression comparisons among the ploidy levels (i.e., 1X versus 2X, 1X versus 3X, 1X versus 4X, 2X versus 3X, 2X versus 4X, and 3X versus 4X). Slide sets (University of Arizona Oligonucleotide Maize Microarrays), containing more than 55,000 unique maize genes printed across two microarrays in the set, were employed for each of the 6 pairwise polyploidy level comparisons. Within each slide set, Chip A contains 28,735 unique genes, Chip B contains 26,543 unique genes, and there are 786 genes spotted in common across both chips.

Four technical replicates of each ploidy level per dye swap were compared (e.g., 4 technical replicates hybridized to reverse labeled slides yields 8 slide sets per single dye swapped comparison) in a duplicated dye swap experimental design (e.g., 2*8=16 slide sets). The duplication of the dye swap provided the biological replicate.

To reduce the overall dye effect and array effect on the expression values an analysis of variance (ANOVA) (Black and Doerge 2002; Craig et al. 2003) was performed for each gene. The original measurements were adjusted by subtracting the array and dye effect as estimated by the ANOVA model. Similar to the simulated data, the adjusted data (all genes across all experiments or polyploidy levels) were clustered with iCORE, CORE, and hierarchical clustering. Since the number of true clusters is a parameter of the maize experiment, the Gap statistic (Tibshirani et al. 2001) was employed to determine the number of clusters. The ideal numbers of clusters for iCORE, CORE and hierarchical clustering was found to be 9, 10, and 8, respectively. Even though the true cluster number for the maize data is unknown, the adjusted Rand index was employed to investigate the agreement between all pairwise clustering method results. When iCORE and CORE are compared an adjusted Rand index of 0.401 is obtained; the index for the comparison between iCORE and hierarchical clustering is 0.262, while it is 0.181 between CORE and hierarchical clustering. Consistent with the results from the simulated data, these findings indicate that iCORE outperforms the other two methods. Therefore, only the results from the iCORE analysis are presented.

The iCORE clustering results of the maize data set are shown in Figure 5. For each cluster the nature of the expression change across the profile is illustrated via the bold line. For example, the overall gene expression pattern in cluster 3 decreases from monoploid to diploid, increases from diploid to triploid, and increases even further from triploid to tetraploid. Even though the expression patterns in cluster 3 (green), cluster 6 (pink) and cluster 8 (gray) are similar, iCORE is able to discern differences. As was previously indicated from a Northern analysis for many genes (Guo et al., 1996), it is quite clear that the hypothesis of additive gene action across polyploidy levels does not hold for these experimental data. For example, in the cluster 3 the expression profile has a peak at the triploid level, while in the cluster 6 the profile takes on almost the same values (highest) at both the monoploid and triploid levels. However, as seen in the cluster 8, the monoploid level is the highest. If in fact, gene expression is additive, there would likely be one cluster. Clearly, there are nine unique clusters (Figure 5) with each differing from the others in its overall cluster profile (i.e., illustrated within each cluster with an opposing color).

7. Summary

A modified or improved clustering algorithm called iCORE (improved Clustering Of Repeated Expression) is presented for the purpose of incorporating both gene expression and gene variance information. The iCORE approach is different from existing methods in that it not only incorporates the error information as gained from replicated experiments, but is also able to differentiate negatively correlated genes. In terms of accuracy of clustering, iCORE demonstrates an improved statistical power over its predecessor, the CORE algorithm, as well as over the well-known hierarchical clustering approach. When iCORE is applied to the maize ploid series data, nine unique clusters were found to clearly dispute the hypothesis that gene expression in polyploids is additive. Both iCORE and CORE capture the heterogeneous variation among genes and experiments while the hierarchical clustering, which is based on average gene expression, does not take into account the error information from replication. Although iCORE appears to outperform CORE and hierarchical clustering it does not identify all possible expression patterns in the maize polyploid series. For example, in clustering the polyploid series data, a cluster whose pattern is incrementally higher across the polyploidy levels (from low to high) is not obtained. In fact, when the data are explored further, such a cluster indeed exists, but it appears to be a minor cluster (with few genes) that is absorbed (or hidden) by the major clusters that contain many genes. In order to improve this limitation, and as a point of future research, the specification of the dependence or correlation structure across the ploidy levels in series, rather than using the individual experiments/ploidy levels independently, will be explored.

8. Acknowledgements

We thank the RWD research group members for helpful discussions. This research is supported by a grant from the National Science Foundation Plant Genome (0501712 DBI) to JB and RWD. NCR was supported by an NRSA post-doctoral fellowship from the National Institute of Health.

9. References

Adams K. and Wendel J. 2005. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* 8:135-141.

Black M.A. and R.W. Doerge. 2002. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics* 18(12):1609-1616.

B.A. Craig, M.A. Black, and R.W. Doerge. 2003. Gene Expression Data: The technology and statistical analysis. Journal of Agricultural, Biological, and Environmental Statistics (JABES) 8(1):1-28.

Eisen M., Spellman P., Brown P. and Botstein D. 1998. Cluster analysis and display of genome-wide expression pattern. Proceedings of the National Academy of Sciences USA 95(25): 14863-14868.

Guo, M., Davis, D., and Birchler, J.1996. Dosage effects on gene expression in a maize ploidy series. *Genetics* 142:349-1355.

Hughes T., Marton M., Jones A., Roberts C., Stoughton R., Armour C., Bennett H., Coffey E., Dai H., He Y., Kidd M., King A., Meyer M., Slade D., Lum P., Stepaniants S., Shoemaker D., Gachotte D., Chakraburtty K., Simon J., Bard M., and Friend S. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102(1):109-126.

Hubert, L. and Arabie, P. 1985. Comparing partitions. *Journal of Classification* 2:193-218.

Luan, Y. and Li, H. 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics* 19(4):474-482.

Ma, P., Castillo-Davis, C., Zhong, W., and Liu, J. 2006. A data-driven clustering method for time course gene expression data. *Nuccleic Acids Research* 34(4):1261-1269.

Medvedovic, M. and Sivaganesan S. 2002. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* 18(9):1194-1206.

Milligan G. and Cooper M. 1986. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research* 21:441-458.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. 2003. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 52(1-2):91-118.

Peddada S., Lobenhofer E., Li L., Afshari C., Weinberg C., and Umbach D. 2003. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics* 19(7):834-841.

Rand, W., 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66:846–850.

Tavazoie S., Hughes J., Campbell M., Cho R. and Church G. 1999. Systematic determination of genetic network architecture. Nature Genetics 22(3): 281-285

Tibshirani R., Guenther W., and Trevor H. 2001. Estimating the number of clusters in a data set via the gap statistic. Journal *of the Royal Statistical Society* 63(2):411-423.

Tjaden, B. 2006. An approach for clustering gene expression data with error information. *BMC Bioinformatics* 7:17.

Yeung, K., Medvedovic, M., and Bumgarner, R. 2003. Clustering gene expression data with repeated measurements. *Genome Biology* 4(5):R34.

		U Partition	
		Same group	Different group
V Partition	Same group	a	b
	Different group	c	d

Table 1. Summarizing the numbers of pairs of objects across two paritions (U and V) according to whether they are in the same or different group.

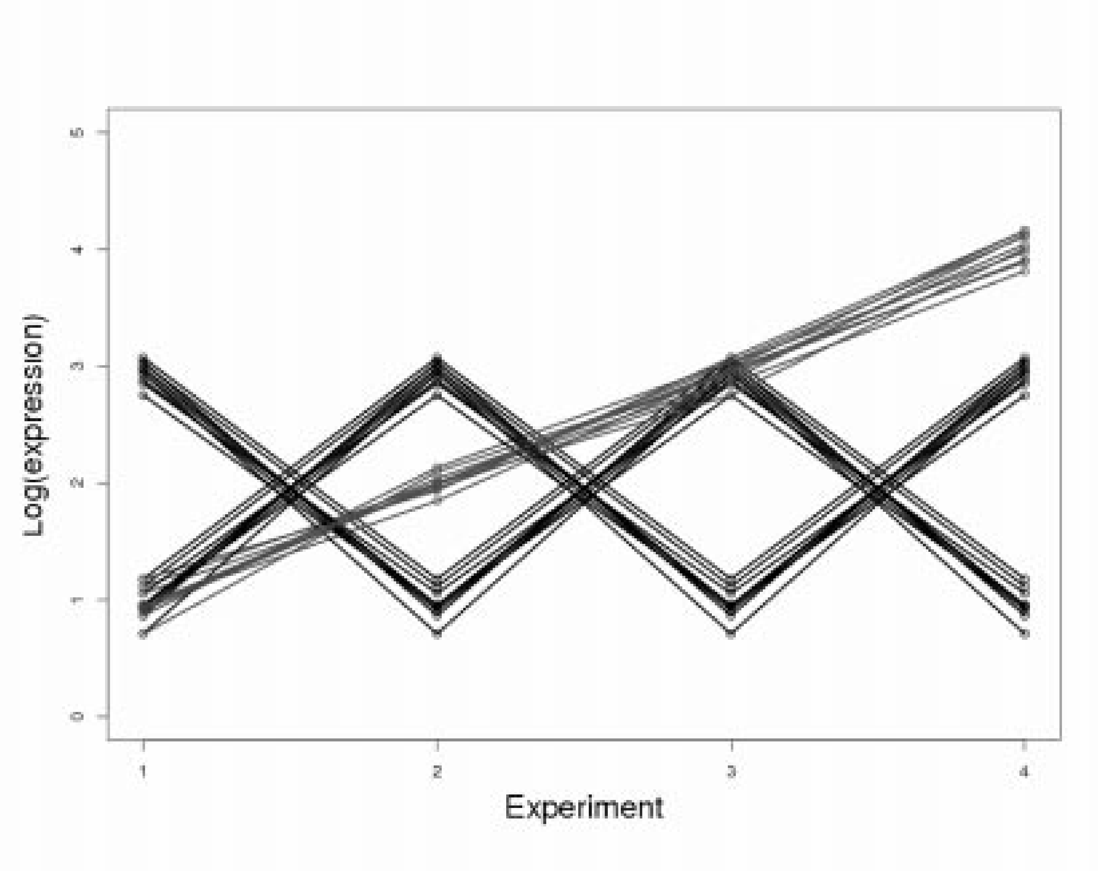

Figure 1. Three gene cluster profiles across four experiments. Using CORE the black (opposing) profiles and the red profile are placed into two unique clusters. iCORE acknowledges the opposing direction of the black profiles and places each in its own unique cluster thus recognizing the three different gene expression profiles.

```
n genes; m experiments, $k_0$ clusters

**Initialization**: randomly assign each gene to one of the $k_0$ clusters uniformly:
    $\Delta(0) = \infty \qquad \Delta(1) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{y_{ij}}{\sigma_{ij}}\right)^2$
**Repeat steps 1- 2** while $(\Delta(x-1) - \Delta(x) >$ threshold) :
  1. In each cluster (k=1, …, $k_0$ ) calculate its cluster profile $A_k$
     • Let $A_k$ = mean of all gene expression profiles in the $k^{th}$ cluster
     • Repeat the following steps a) and b) until vector $A_k$ converges
         a) Given $A_k$, for each gene assigned to that cluster, find $B_i$ and $C_i$ which minimize
              $\sum_{j=1}^{m}\left(\frac{y_{ij} - (B_i A_{kj} + C_i)}{\sigma_{ij}}\right)^2$
         b) Given $B_i$ and $C_i$, find $A_k$ such that it minimizes
              $\sum_{i\in k}\sum_{j=1}^{m}\left(\frac{y_{ij} - (B_i A_{kj} + C_i)}{\sigma_{ij}}\right)^2$
  2. Reassign each gene to the nearest cluster
     • Given the parameter vector $A_k$ obtained in the first step for the $k^{th}$ cluster,
       estimate factors $B_i$ and $C_i$ for gene i in that cluster by using
              $\min_{B_i, C_i}\sum_{j=1}^{m}\left(\frac{y_{ij} - (B_i A_{kj} + C_i)}{\sigma_{ij}}\right)^2$
     • Assign each gene to its closest cluster by defining the distance as:
              $\sum_{j=1}^{m}\left[\frac{y_{ij} - (|B_i| A_{kj} + C_i)}{\sigma_{ij}}\right]^2$
     • Let $\delta_i$ be the distance of gene i to its closest cluster, then the total sum of within-
       cluster distance is :
              $\Delta(x) = \sum_{i=1}^{n}\delta_i$
```

Figure 2: An algorithmic explanation of iCORE.

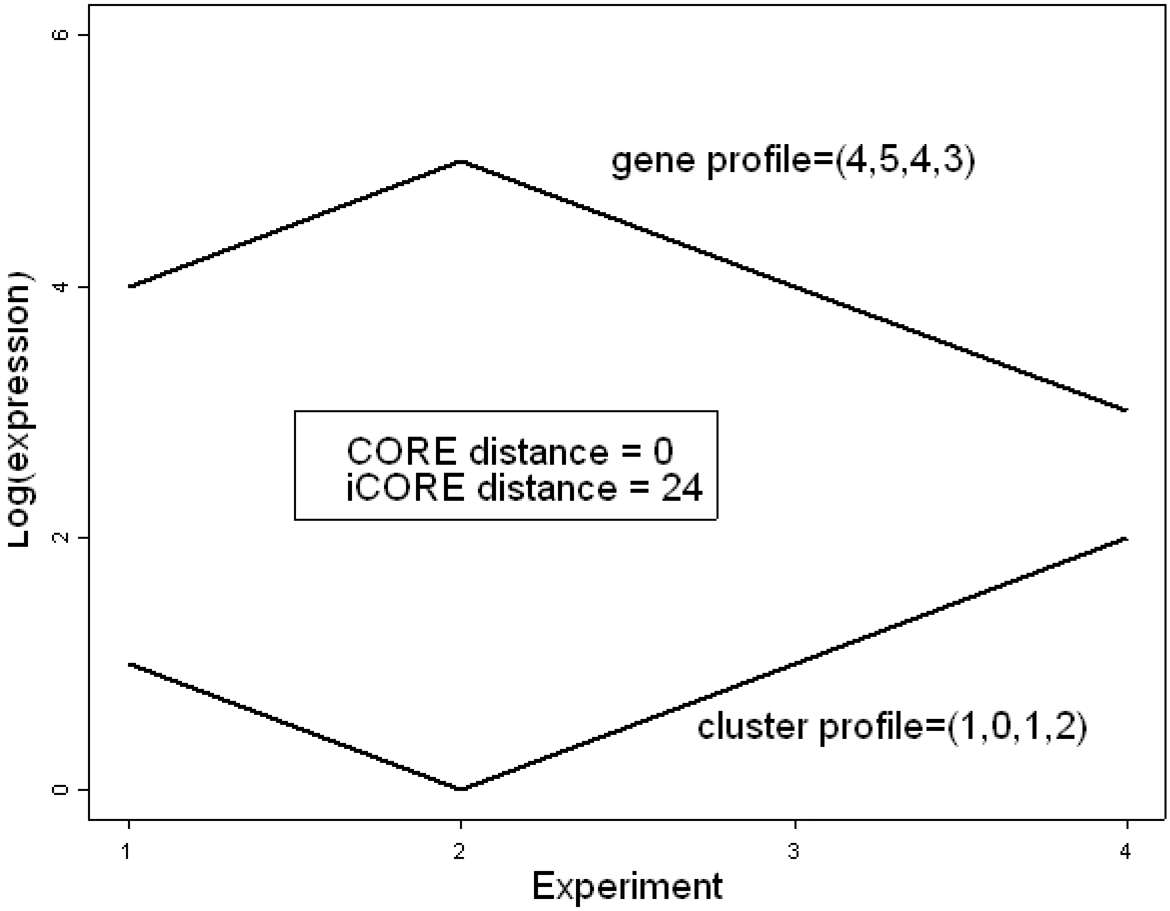

Figure 3. When a gene profile is opposing to a cluster profile the iCORE distance (Equation 2b) is larger than the CORE distance (Equation 2a). Thus, a gene is less likely to be assigned to a cluster whose pattern is opposing to the gene's expression profile.

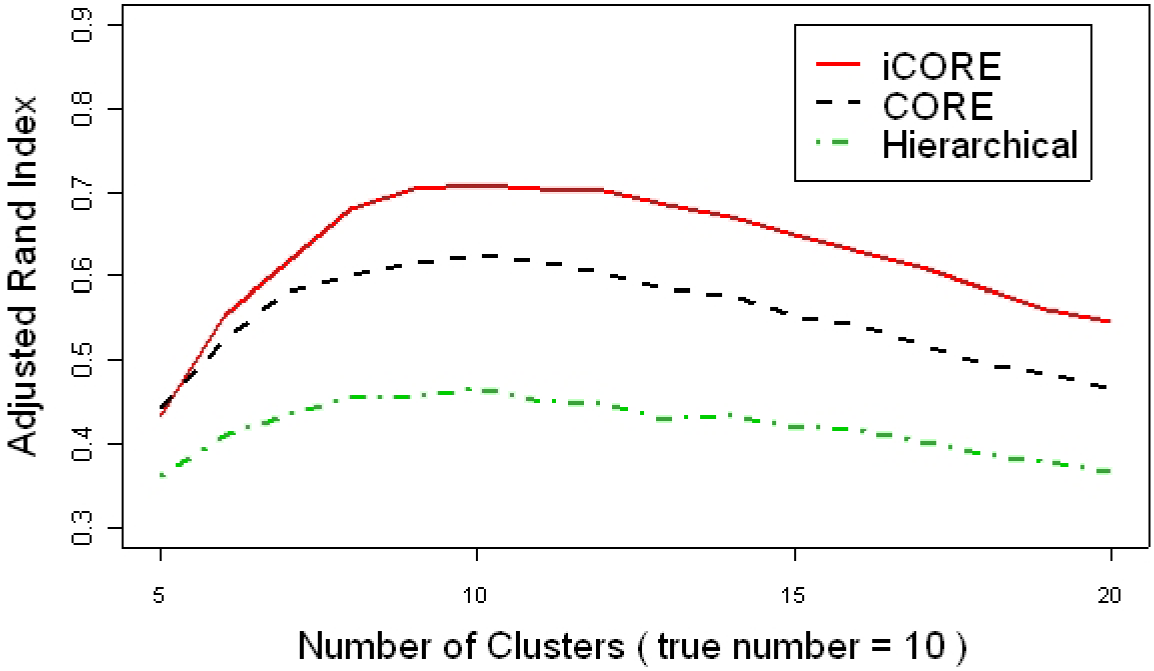

Figure 4. A comparison of the clustering accuracy for iCORE, CORE and hierarchical clustering (average gene expression across the replicates for each gene in every experiment) using the adjusted Rand Index and simulated data. Each curve reflects the average adjusted Rand index of clustering quality over a varying number of clusters. Each data point on a curve is an average of 100 simulated and clustered data sets. The true number of clusters is 10.

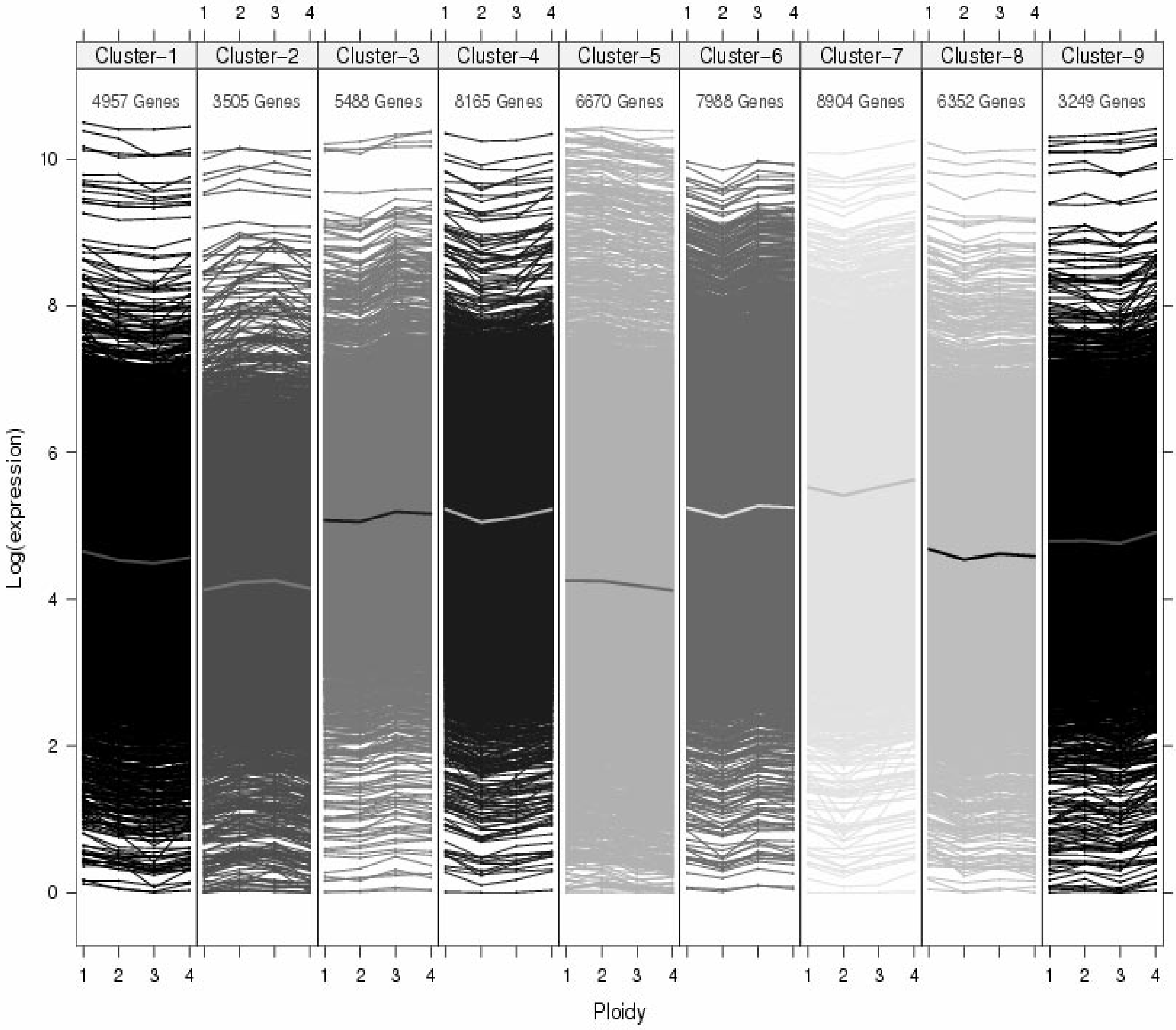

Figure 5. Clustering results by iCORE algorithm for maize data. The 9 panels represent 9 gene clusters. In each panel the curve with opposing color is the representative cluster profile for that group. The number of genes in each cluster is displayed at the top of each panel.