Mol. Cells, Vol. 25, No. 2, pp. 279-288

Communication

Molecules and Cells

Clustering Approaches to Identifying Gene Expression Patterns from DNA Microarray Data

Jin Hwan Do[1] and Dong-Kug Choi*

Department of Biotechnology, Konkuk University, Chungju 380-701, Korea;
[1] Bio-Food and Drug Research Center, Konkuk University, Chungju 380-701, Korea.

(Received July 30, 2007; Accepted November 5, 2007)

The analysis of microarray data is essential for large amounts of gene expression data. In this review we focus on clustering techniques. The biological rationale for this approach is the fact that many co-expressed genes are co-regulated, and identifying co-expressed genes could aid in functional annotation of novel genes, *de novo* identification of transcription factor binding sites and elucidation of complex biological pathways. Co-expressed genes are usually identified in microarray experiments by clustering techniques. There are many such methods, and the results obtained even for the same datasets may vary considerably depending on the algorithms and metrics for dissimilarity measures used, as well as on user-selectable parameters such as desired number of clusters and initial values. Therefore, biologists who want to interpret microarray data should be aware of the weakness and strengths of the clustering methods used. In this review, we survey the basic principles of clustering of DNA microarray data from crisp clustering algorithms such as hierarchical clustering, *K*-means and self-organizing maps, to complex clustering algorithms like fuzzy clustering.

Keywords: Co-Expression; DNA Microarray; Fuzzy Clustering; Hierarchical Clustering; *K*-means; Self-organizing Map.

Introduction

Microarray technologies are powerful techniques for simultaneously monitoring the expression of thousands of genes under different sets of conditions. These conditions may involve different cell lines, diverse physiological conditions, pathological versus normal tissues, or serial time points following a stimulus (Alizadeh et al., 2000; Eisen et al., 1998; Wen et al., 1998). Clustering of gene expression data can be divided into two main categories: gene-based clustering and sample-based clustering (Belacel et al., 2006). In gene-based clustering, genes are treated as objects and samples are treated as features or attributes for clustering. The goal of gene-based clustering is to identify differentially expressed genes and sets of genes or conditions with similar expression patterns or profiles, and to generate a list of expression measurements. As sets of genes with similar expression patterns may share biological functions and be under common regulatory control, genes are frequently clustered according to their expression patterns in genome-wide expression data analysis (Eisen et al., 1998). Sample-based clustering can be used to reveal the phenotypic structures or substructures of samples. Golub et al. (1999) have demonstrated that the phenotypes of samples can be discriminated by employing only a small subset of genes whose expression levels strongly correlate with the class distinctions. These genes are called *informative genes*. The remaining genes are irrelevant to the classification of samples of interest and thus are regarded as noise. Here, we will focus on gene-based clustering which identifies sets of genes that are co-expressed. Many clustering algorithms have been used to identify genes displaying similar expression patterns because such genes have a high probability of being co-expressed. The information on co-expression can be combined with other types of data to yield new conclusions such as functional annotation of novel genes and identification of transcription factor binding sites (TFBS).

* To whom correspondence should be addressed.
Tel: 82-43-840-3610; Fax: 82-43-840-3872
E-mail: choidk@kku.ac.kr

Abbreviations: ANN, artificial neural network; PAM, partitioning around medoids; SOFM, self-organizing feature map; TSVQ, tree-structured vector quantization; VNS, variable neighborhood search.

Clustering methods can also be divided into two general classes, designated supervised and unsupervised (Raychaudhuri et al., 2001). Supervised methods are widely used by biologists to locate the informative genes in sample-based clustering. They take known class patterns and create rules for reliably assigning genes or conditions into each cluster using various machine-learning techniques such as logistic regression, neural networks and linear discriminant analysis. Thus, the supervised methods cannot be applied unless the phenotypes of samples or class patterns are known in advance. Unsupervised methods, however, group similar patterns based on a distance metric without prior information about class patterns. Clustering of microarray gene expression data is performed mostly by unsupervised or hybrid (unsupervised followed by supervised) methods due to the absence of information on known expression patterns. A key weakness of unsupervised methods is that they assume the existence of an underlying pattern in the data. Thus, the output of unsupervised methods should be rigorously validated, both statistically and scientifically (Boutros and Okey, 2005). In addition, the results of unsupervised methods depend on the clustering algorithms and distance metrics used. Therefore, an understanding of the various unsupervised clustering algorithms is a prerequisite for proper grouping of genes according to their expression patterns. Here, we present the basic principles of unsupervised gene-based clustering, from crisp clustering algorithms such as hierarchical clustering, *K*-means and self-organizing maps, to complex clustering algorithms like fuzzy clustering, together with their pros and cons. Before describing unsupervised clustering algorithms for gene expression data, it is worth looking at methods for similarity measure between pairs of genes, because the basis of unsupervised clustering is to group together genes by similarity of expression. Thus, the various types of distance metric are investigated below.

Types of distance metric for similarity measures between pairs of genes To group together genes with similar expression patterns it is necessary to quantitatively measure the similarity of two expression patterns. Here, an expression pattern is considered to be the values that make up the expression profile for a single gene in a series of conditions. The similarity of two expression patterns can be measured by various metrics such as Euclidian distance, cosine-angle metric, Pearson correlation and Spearman rank correlation. Euclidian distance measures the absolute distance between two expression patterns, and thus considers the magnitude of changes in the gene expression levels, while the cosine-angle metric measures the angular separation of two expression patterns. The Pearson correlation measures the directional similarity of two expression patterns, and is thus insensitive to the amplitude of the expression vector. Expression patterns that are visually similar will have high Pearson correlation coefficients. The value of the Pearson correlation coefficient is always between -1 and 1, with 1 meaning that the two expression vectors have exactly same 'shape', 0 meaning that they are completely unrelated, and -1 meaning that they are perfect opposites. The Pearson correlation coefficient for two genes that are co-expressed should be near 1. This is the reason that the Pearson correlation metric is most commonly used in clustering DNA microarray expression data. It is, however, very sensitive to outliers. If the expression levels of two patterns are completely unrelated, but one component and both patterns have a high peak or valley at that component, then the Pearson correlation coefficient will be very high.

In order to address this problem, Heyer et al. (1999) proposed the jackknife correlation, which calculates the correlation n times, each repeat leaving out one dimension of the total n dimensions of the expression pattern, and calculates the correlation only on the remaining n-1. They define the jackknife correlation J_{ij} for the expression pair i, j as

$$J_{ij} = \min\{\rho_{ij}^{(1)},...,\rho_{ij}^{(n)}\}$$

where $\rho_{ij}^{(l)}$ denote the jackknife correlation of the pair i, j computed with the l-th component deleted. Thus, the jackknife correlation is not distinguished by a simple shift. Another correlation metric robust to outliers is the Spearman rank correlation, which uses the ranks, instead of their actual values in each expression pattern. Therefore, the Spearman correlation may not require normalization of the micro array expression data, which is an essential step for microarray data analysis because there are many variations that affect the measured gene expression levels (Do and Choi, 2006). The Spearman rank correlation may be a good compromise between numerical measures such as the Pearson correlation or Euclidian distance, and simple qualitative measures that consider only the relevant information like the 'ups' and 'downs' of a time series (Balasubramaniyan et al., 2005). The distance between two expression patterns in the correlation-based metric is generally obtained by subtracting the correlation coefficient from unity.

For a clearer understanding of each type of distance metric, five gene expression patterns are shown in Fig. 1. The distance for each pair of patterns is represented in Table 1, and the distance between two patterns is inversely proportional to their similarity. Patterns 1 and 2 have the highest similarity or shortest distance in Euclidian distance, while the pattern most similar or nearest to pattern 1 is pattern 5 in the Pearson correlation-based distance. Pattern 5 can be overlapped with pattern 1 by simple linear transformation, which generates zero for the Pearson correlation-based distance. The ranks for every component in patterns 1, 3 and 5 are the same; thus any pair of these three patterns yields a Spearman correlation-

Table 1. Pairwise distance according to four similarity measures. As the distance between two patterns decreases, their similarity increases.

Patterns	Euclidian distance	Correlation based distance		
		Cosine-angle metric	Pearson correlation	Spearman rank correlation
(1,2)	0.5148	0.0662	0.3134	0.4054
(1,3)	1.8901	1.1391	0.0536	0
(1,4)	2.1194	1.9916	1.9464	2
(1,5)	3.0541	1.8612	0	0
(2,3)	1.711	1.0825	0.2187	0.4054
(2,4)	1.9357	1.9517	1.7813	1.5946
(2,5)	2.8561	1.8092	0.3134	0.4054
(3,4)	1.2141	0.9078	2	2
(3,5)	1.3892	0.3979	0.0536	0
(4,5)	1.2048	0.1510	1.9464	2
Formula of distance**	$\sqrt{\sum_{i=1}^{n}(a_i-b_i)^2}$	$1-\frac{\sum_{i=1}^{n}a_ib_i}{\|a\|\|b\|}$	$1-\frac{\sum_{i=1}^{n}(a_i-\bar{a})(b_i-\bar{b})}{\sigma(a)\sigma(b)}$	$1-\frac{\sum_{i=1}^{n}(a_i^*-\overline{a^*})(b_i^*-\overline{b^*})}{\sigma(a^*)\sigma(b^*)}$

** $a=(a_1,...,a_n), b=(b_1,...,b_n), \bar{a}=\frac{a_1+...+a_n}{n}, \bar{b}=\frac{b_1+...+b_n}{n}, \|a\|=\sqrt{\sum_i^n a_i^2}, \|b\|=\sqrt{\sum_i^n b_i^2}$

$\sigma(a)=\sqrt{\frac{\sum_i^n(a_i-\bar{a})^2}{n}}, \sigma(b)=\sqrt{\frac{\sum_i^n(b_i-\bar{b})^2}{n}}, a^*=rank\,a\,(\text{if } a_1<a_2<...<a_{n-1}<a_n \text{ then } a^*=(n,n-1,...,2,1)), b^*=rank\,b$

based distance of zero, while the ranks in pattern 4 are the reverse of these, and hence the Spearman correlation-based distance from pattern 4 to any of the others has the maximum value, namely, 2.

Recently, Hardin et al. (2007) proposed Tukey's biweight as a resistant correlation measure for two genes on a microarray. They demonstrated that the biweight is a powerful technique to use when computing correlations between pairs of genes regardless of whether there is a significant amount of contamination. By separately modeling the shape and the magnitude parameters in a gene expression profile, Kim et al. (2007) proposed a new measure for clustering genes when the profile shape is a key factor, and when the expression magnitude also needs to be accounted for in determining the gene relationships. This approach is to use the estimated shape and magnitude parameters to define a Chi-square-statistic based on distance in a new feature space.

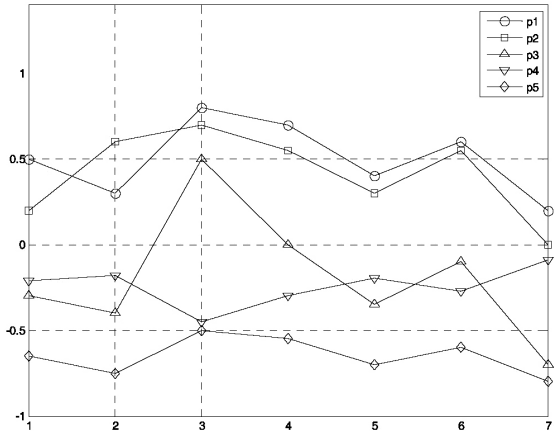

Fig. 1. Five gene expression patterns consisting of seven expression values. Vertical and horizontal axes represent normalized expression values and experimental conditions, respectively.

Hierarchical clustering Hierarchical clustering methods have been extensively used in the analysis of many types of microarray data, such as gene expression data, CGH arrays, and protein arrays (Ikota et al., 2006; Wu et al., 2006; Xing et al., 2007). The goal of hierarchical clustering is to yield the definitive clustering that characterizes a set of patterns in the context of a given distance metric. Graphic representation of the results of hierarchical clustering allows users to visualize global expression patterns in DNA microarray data, which makes this method a favorite among biologists (Tseng, 2004). This clustering can be divided into two splitting methods, namely agglomerative (bottom-up) and divisive (top-down), depending on how the comparisons are made (Azuaje, 2003). The agglomerative method calculates a table containing the distances from each cluster to every other cluster, and initially consists of a single pattern. Then the two most similar clusters are merged into a single super-cluster and the distances between clusters are re-calculated. This process of distance calculation and merging is repeated until the

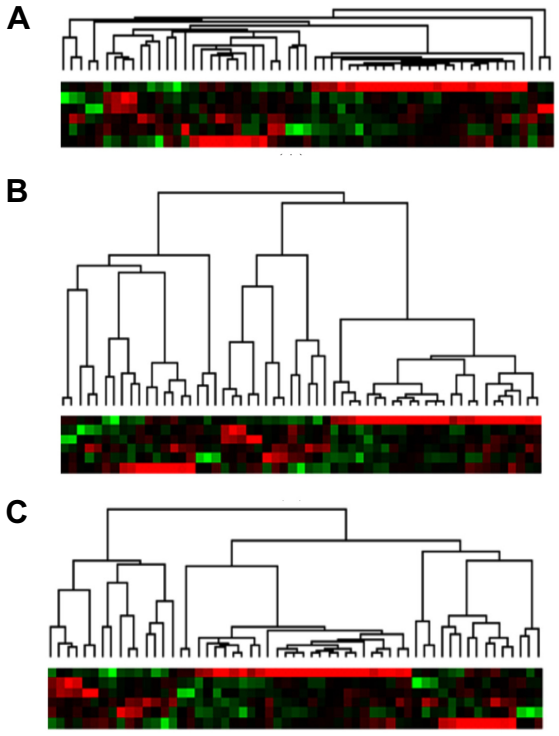

Fig. 2. Hierarchical clustering of 56 genes selected from a microarray data set. (A) Single linkage clustering. (B) Complete linkage clustering. (C) Average linkage clustering.

entire tree is constructed. Thus, this approach operates on a dissimilarity matrix instead of directly on patterns, and is computationally expensive for large data ($O(n^2)$ calculations for n patterns).

The distances between clusters can be measured by various methods such as single linkage, complete linkage, centroid linkage and average linkage. The distance between the closest neighbors in different groups is taken for single linkage, while the distance between furthest neighbors is taken for complete linkage. Centroid and average linkages use the distances between cluster centroids, and the average of all distances between patterns in the two clusters, respectively. It is not appropriate to use centroid linkage when the distance metric is based on the Pearson correlation. The reason is that no normalization is included when the cluster centroid is calculated, whereas the patterns are implicitly normalized when the Pearson correlation-based distance is calculated. The selection of a distance between clusters will affect the result of clustering. Single linkage tends to produce long stringy clusters while complete linkage tends to produce compact clusters (Fig. 2). Centroid and average linkages are compromises between single and complete linkages (Chipman, 2006).

The divisive method divides the set successively into two clusters using a number of non-hierarchical clustering algorithms. Thus, this approach is the opposite of that taken with the agglomerative method in that it views the data top down, rather than bottom up. Each division is a two-group partitioning problem. Thus, recursive application of K-means with $K = 2$, known as 'tree-structured vector quantization' (TSVQ), is a kind of top-down clustering (Gersho and Gray, 1992). The TSVQ algorithm is fast, requiring $O(n\log_2 n)$ calculations (worst-case $O(n^2)$). Another divisive method uses the Macnaughton-Smith algorithm (Macnaughton-Smith et al., 1964), which starts with searching for the pattern with the highest average distance from all other patterns. The selected pattern forms a splinter group. Patterns that are closer to the splinter group are moved to the splinter group one at a time until no pattern in the original group is closer to the splinter group. This algorithm takes $O(n^2)$ to split a group, and is thus probably slower than TSVQ. The agglomerative method, which successively joins objects, is good at identifying small clusters, but can give sub-optimal performance for identifying a few large clusters, while the devisive method recognizes a few large clusters but is weak in identifying many small clusters. Hierarchical clustering methods, including agglomerative and divisive approaches, follow a strategy that prevents cluster refinement once a decision is made to merge or split clusters. In addition, the iterative merging of clusters is determined locally at each step by the pairwise distances rather than a global criterion (Tan et al., 2005). Chipman et al. (2006) have proposed a hybrid hierarchical clustering method that combines the strengths of agglomerative and devisive methods by modifying the former with information gained from a preliminary use of the latter. This combination is facilitated by the concept of 'mutual cluster', which is defined as a group of patterns collectively closer to each other than any other object. The mutual clusters are not broken no matter whether single, complete or average linkage is used in the agglomerative approach, but can be broken by the divisive approach.

***K*-means clustering** K-means is one of the most commonly used partitioning methods. This clustering starts by randomly choosing k patterns as initial means for each cluster. After that the patterns are assigned to the clusters by finding the patterns' closest mean. The new mean is calculated for each cluster and the patterns are reassigned to new means. This process is iterated until the cluster means are such that no pattern moves from one cluster to another. The K-means algorithm is fast, because at each iteration KN distances are evaluated and K means updated when N is the total number of patterns. It is not appropriate to use the K-means algorithm with Pearson correlation-based distances because the distance measures based on the Pearson correlation effectively normalize the patterns when calculating distance, whereas no normalization is used when calculating the cluster mean. Theoretically, it is best to use K-means with the Euclidean distance.

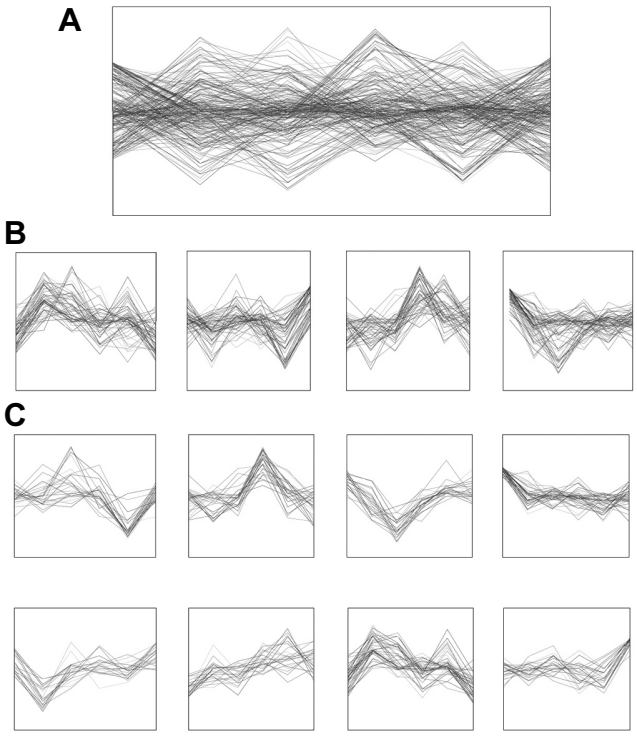

Fig. 3. *K*-means clustering of 230 genes with different *K* values. **A.** Normalized expression patterns consisting of six expression values ($K = 1$). **B.** *K*-means clustering with $K = 4$. **C.** *K*-means clustering with $K = 8$.

An issue that needs careful consideration in using *K*-means is the choice of the number of clusters, *k*, which must be specified by the user. Prior knowledge of the number of clusters is not available in most gene expression clustering. Figure 3 gives the results of *K*-means clustering with various *K* values. The algorithm can even be repeated with different *k* values to obtain the optimum number of clusters, but this process is not practical for a large amount of gene expression data. Dudoit et al. (2002) proposed a prediction-based re-sampling method, Clest, for estimating the number of clusters in a dataset. The algorithm estimates the number of clusters *K* by repeatedly dividing at random the original dataset into two non-overlapping sets, a learning set and a test set. After a predictor is built using the class labels obtained by clustering the learning set, it is applied to the test set and the similarity between predicted labels and those produced by clustering the test set is measured using an external index such as the Fowlkes and Mallows (FM) (1983) index. The number of clusters is estimated by comparing the observed similarity statistic for each *K* with its expected value under a suitable null distribution with $K = 1$. The drawback of Clest is that a large number of parameters need to be set by the user. Ben-Hur et al. (2002) have proposed a stability-based method for estimating the number of clusters, where stability is characterized by the distribution of pairwise similarities between clusters of subsamples of the dataset. The pairwise similarity between clusters can be measured by the FM index or the matching coefficient (Jain and Dubes, 1988). The basic idea of this approach is that a dataset has *K* clusters if its different subsamples have similar *k* clusters. The algorithm searches for the largest *K* such that partitions into *k* clusters are stable, but also gives information on how well-defined the structure in the data is via the sharpness of the transition from stable to unstable solutions. The stability-based method can be combined with any clustering algorithm, but proves to be most useful in conjunction with hierarchical clustering.

Another weakness of *K*-means clustering is its sensitivity to noise and outliers, which are frequently present in gene expression data. This problem can be overcome to some extent by using medoids instead of means. The medoid of a cluster of patterns is the pattern with smallest average distance to all other patterns. The calculation of the medoid of each cluster is computationally more expensive than that of the centroid because it uses all pairwise distances within each cluster. However, use of the medoid instead of the mean is more robust to outliers in the same way that a univariate median is more robust than a mean. Partitioning around medoids (PAM) (Kaufman and Rousseeuw, 1990) and its extension PAMSIL (van der Laan et al., 2003) are popular *K*-medoid algorithms. The PAM function takes as its arguments a dissimilarity matrix like the Euclidian distance matrix and a prespecified number of clusters *K*, and searches for *K* representative medoids, which are taken advantage of in constructing *K* clusters. The objective function of PAM finds *K* medoids that minimize the sum of the dissimilarities of the patterns to their nearest medoid, whereas the objective function of PAMSIL searches for *K*-medoids that maximize the average silhouette width of a cluster. The silhouette width for the *i*-th gene in the cluster *j* is defined as

$$s(i) = \frac{b_l(i) - a_j(i)}{\max\{a_j(i), b_l(i)\}}, (1 \le j, l \le k, j \ne l)$$

where k is the total number of cluster and gene i belongs to cluster j

In this equation, $a_j(i)$ is the average distance between the *i*-th gene and all of the genes in the *j*-th cluster, and $b_l(i)$ is the smallest average distance between the *i*-th gene and all of the genes in the *l*-th cluster. Figure 4 shows an example of clustering by PAM. The best cluster number in this clustering is estimated to be 2 because the maximum of the average silhouette of the cluster is obtained at $K = 2$.

Finally, *K*-means clustering has a tendency to converge to a local optimum due to random initial clustering. Therefore, the *K*-means clustering is generally repeated with a different initial cluster assignment to obtain a global optimum that minimizes the sum of within-cluster

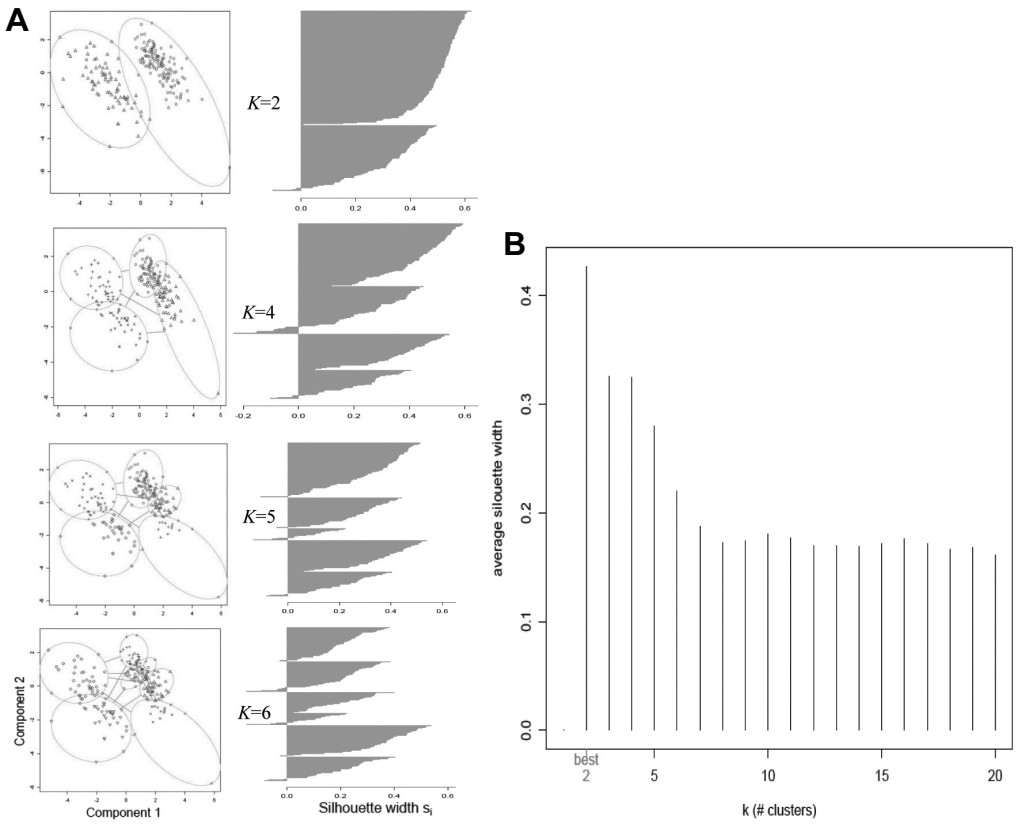

Fig. 4. PAM clustering of 206 gene expression patterns consisting of eight expression values. **A.** Plane (left) and silhouette plot (right) of the two principal components with various cluster numbers. **B.** Effect of cluster number on average silhouette width.

distances. Various optimization techniques such as simulated annealing and a genetic algorithm can be used to find the global optimal solution. Krishna and Murty (1999) have proposed a new genetic *K*-means algorithm (GKA) that combines a genetic algorithm with a gradient descent algorithm and the *K*-means algorithm. The GKA, like the genetic algorithm, is insensitive to the initiation process, but still suffers from high computational costs in areas such as gene expression data analysis. A faster version of GKA (FGKA) (Lu et al., 2004a) and an extended version of FGKA, incremental genetic *K*-means algorithm (IGKA) (Lu et al., 2004b), give better time performance than the original GKA. The FGKA partitions the data set into *k* clusters, such that it minimizes the total within-cluster variation (TWCV), whereas IGKA clusters centroids incrementally when the mutation probability is small after calculation of TWSV. Both FGKA and IFGA, like GKA, always converge to the global optimum.

Self-organizing feature maps The self-organizing feature map (SOFM or SOM) was proposed by Kohonen (1990). The SOFM divides the input patterns into groups of similar patterns, like *K*-means and hierarchical clustering, but the output of SOFM is a grid of spatially located maps according to features of the input patterns. A one- or two-dimensional grid is most widely used for SOFM. The SOFM is based on a single-layered artificial neural network (ANN) that consists of simple elements called units or neurons. Clustering is started after each neuron is initially associated with a random weight pattern or reference pattern. When an input pattern is presented to the neuron network, each neuron calculates the distance between its weight pattern and the input pattern presented. The neuron closest to the presented input pattern is the winning neuron. The weight pattern of the winning neuron is modified in such a way that the weight pattern becomes more similar to the presented input pattern. The winning neuron's neighbors are also changed in the same way but to a lesser extent. As presentation of all the patterns in the input space to the neuron network is repeated, the change of weight patterns decreases. The mapping of all input patterns to the output neurons leads to the identification of clusters. Figure 5 shows a two-dimensional SOFM with a rectangular grid of 10 × 10. The weight patterns of neighboring neurons have similar shapes. SOFM can also provide an intuitive view for mapping high-dimensional datasets.

However, SOFM requires the user to predefine a number of parameters such as the number of initial clusters and the topology of the neurons, as well as learning rates

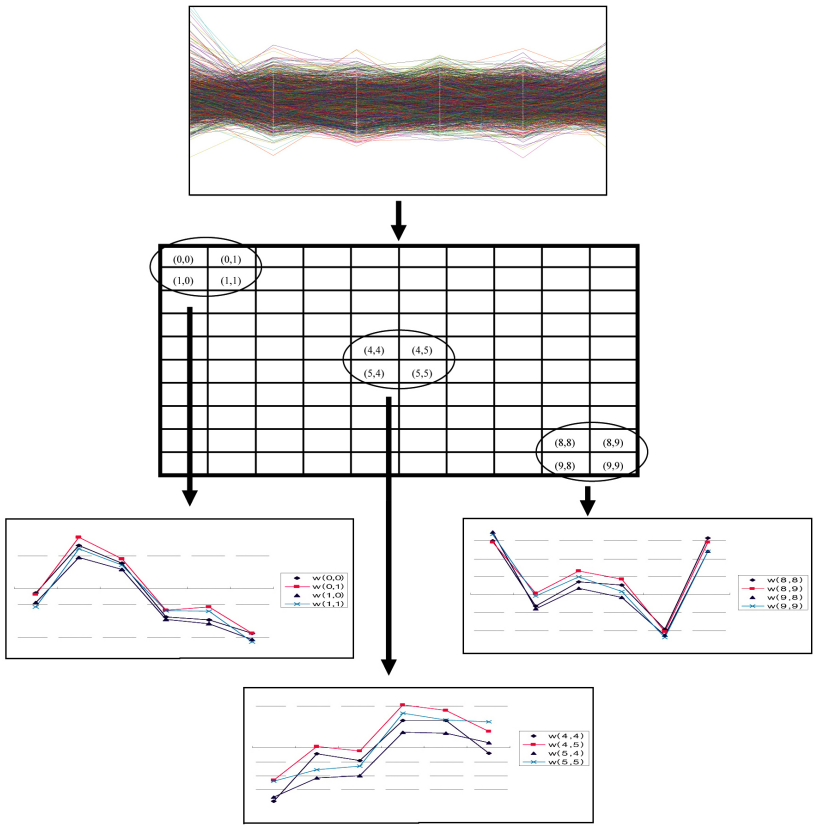

Fig. 5. Two-dimensional (rectangular grid of 10 by 10) SOFM of 2470 gene expression patterns consisting of six expression values. Although the neuron network initially has random weight patterns, the repeated presentation of all the gene expression patterns to a rectangular neuron network increases the similarity of the weight patterns of neighboring neurons. The adapted weight pattern for each neuron is used for clustering all gene expression patterns.

and a neighborhood function. Su and Chang (2001) have proposed a model called double SOM (DSOM) to determine the number of clusters. In DSOM, each neuron is associated with a weight pattern and a two-dimensional position pattern. After the learning process, the number of groups of two-dimensional position patterns is taken to be the number of clusters. Although DSOM addresses the issue of the number of clusters detected, it is still impeded by the selection of its free parameters, such as learning rates and neighborhood functions for updating the weight patterns as well as the corresponding position patterns. To tackle this problem, Ressom et al. (2003) have proposed adaptive DSOM (ADSOM), which updates the free parameters involved in DSOM during the learning process by analyzing the mathematical relationships between the parameters and the updating process. Another approach to estimating clusters is the dynamic SOM tree algorithm, which is a robust hierarchical clustering application based on a growing SOM (GSOM) (Hsu et al., 2000). The GSOM can grow according to its own growth criterion like the growth threshold, whereas SOM is not capable of growing. A dynamic SOM tree is constructed using input tracing with various spread factors (SFs).

Fuzzy clustering Up to now we have considered only crisp clustering methods such as hierarchical clustering, *K*-means and SOM. While these crisp clustering approaches can accurately identify distinct expression patterns by grouping genes with similar expression patterns, they are unable to identify genes whose expression levels are similar to multiple distinct groups of genes. In addition, crisp clustering methods may yield inaccurate clusters that lead to incorrect conclusions when analyzing large gene expression data sets collected under different conditions, since genes are likely to be co-expressed with different groups of genes under different conditions (Gasch and Eisen, 2002). Many approaches to the complex relationships between objects have been developed (Friedman et al., 2000; Ihmels et al., 2002; Sheng et al., 2003; Woolf and Wang, 2000) and the fuzzy clustering method is one of them. This employs the fuzzy logic method for grouping patterns, and provides a systematic and unbiased way to change precise values into several descriptors of cluster memberships (Bezdek, 1981). An indicator variable showing whether a pattern is a member of a given cluster is extended to a weighting factor called membership. Membership value ranges from 0 to 1, where

Fig. 6. The F-CM clustering of 37 gene expression patterns consisting of seven expression values. The column and row represent the centroids of each cluster and gene designations, respectively. The membership value of a gene is inversely proportional to the distance of the gene from each cluster centroid. As the membership value increases, the color changes from dark to deep red. Four genes, 18X, 19X, 20X and 28X have membership values greater than 0.4 in two clusters, namely centroids 6, 7, and the membership value of 21X is greater than 0.2 in three clusters including centroids 5, 6 and 7. The remaining genes clearly belong to clusters having a membership values of 0.5 or above.

membership values close to 0 indicate weak association with the cluster, and membership values close to 1 indicate strong association.

The membership values obtained from fuzzy clustering methods can be used to assign genes that are tightly clustered to one, two or several groups. Therefore, fuzzy clustering uncovers information about the relative likelihood of each gene belonging to each of a predefined number of clusters. The applications of the fuzzy logic method to microarray data analysis have been Fuzzy C-Means (F-CM) (Dembélé and Kastner, 2003; Wang et al., 2003), Fuzzy J-Means (F-JM) and variable neighborhood search (VNS) (Belacel et al., 2004). F-CM is the fuzzy logic extension of the *K*-means heuristic used for crisp clustering, and it searches for the membership degrees and centroids via the objective function. Figure 6 shows an output of F-CM for 37 gene expression patterns. In this clustering, eight centroids have been identified, from centroid 0 to centroid 7. The four genes 18X, 19X, 20X and 28X have membership values of more than 0.4 in two clusters, namely centroid 6, and 7, and the membership value of 21X exceeds 0.2 in three clusters, centroids 5, 6 and 7. The remaining genes clearly belong to a cluster having a membership value of 0.5 or more.

In F-JM methods, the centroid moves to the neighborhood of the current solution defined by all possible centroid-to-pattern relocations, and then the centroids obtained are used to calculate memberships. Both F-CM and F-JM are local heuristic algorithms that cannot guarantee the global optimum. VNS searches for distant, possibly more appropriate, cluster arrangements. It has been demonstrated that the combination of VNS and F-JM gives superior cluster quality and accuracy with four cDNA microarray datasets (Belacel et al., 2004). The main drawback of fuzzy clustering is that the assignment of genes to each cluster depends on membership cutoff. The application of a very high membership cutoff may leave out genes with highly correlated expression patterns in all of the experiments, while a very low membership cutoff will assign most of the genes to every cluster. Therefore, membership cutoff should be carefully chosen to obtain the desired outcome.

Recently, Fu and Medico (2007) proposed a new fuzzy clustering algorithm combining simplicity with good performance and robustness. This algorithm is mainly based on two assumptions: (a) clusters should be identified in the relatively dense part of the dataset; (b) neighboring patterns with similar expression profiles must have similar cluster memberships so that the membership of one object is constrained by the memberships of its neighbors. Thus, the membership of each pattern is not determined with respect to all other patterns in the dataset or some cluster centroids, but is determined with respect to its neighboring objects only. This method can capture non-linear relationships that are lacking in fuzzy C-means-derived clustering approaches. There are other complex clustering methods such as probabilistic clustering in addition to fuzzy clustering. Probabilistic clustering allows each pattern to belong to multiple clusters with given probabilities, like fuzzy clustering. However, this clustering method has some limitations because it relies on the assumption that the dataset fits a statistical distribution like the Gaussian distribution. This assumption may not always be applicable. For example, the Gaussian distribution model may not be effective for time-series data since it ignores the inherent dependence of gene expression on time (Jiang et al., 2003).

Results

Clustering is a mathematical approach to identifying groups of genes that have similar expression patterns in a group of DNA microarray experiments. The grouping of genes according to their expression patterns is performed

by measuring the similarity of genes with respect to expression pattern. A number of measures of similarity in the behavior of two genes can be used, such as Euclidian distance, cosine-angle metric, Pearson correlation and Spearman rank correlation. The most commonly used similarity measures for gene expression data are the Pearson correlation and Euclidean distance. The Pearson correlation is overly sensitive to the shape of an expression profile while Euclidean distance mainly considers the magnitude of the changes of gene expression. The Spearman rank correlation can be less sensitive to outliers than the Pearson correlation, but it can interpret a pattern mistakenly due to the use of rank. Tukey's biweight as a resistant correlation measure for two genes on a microarray is a powerful technique to use when computing correlations between pairs of genes regardless of whether or not there is a significant amount of contamination.

In addition to the distance metric, the output of clustering is affected by the choice of clustering algorithm. The crisp cluster algorithms such as hierarchical clustering, *K*-means and SOM are able to identify genes with highly correlated expression patterns in all the experiments, but cannot identify co-expression with different groups of genes, each governed by a distinct regulatory mechanism. This is due to the fact that the crisp algorithms are based on an assumption about where a gene should belong in a cluster. Information about gene multi-functionality can be provided by fuzzy clustering, which uncovers information about the relative likelihood of each gene belonging to each of a predefined number of clusters. In other words, fuzzy clustering facilitates the identification of overlapping groups of genes by allowing the genes to belong to more than one group. However, the membership cutoff, which determines the assignment of genes to each cluster, should be set by the user.

In conclusion, there are no perfect clustering algorithms suited to clustering genes into functional groups by expression profiling for all data sets. However, the application of several clustering methods to the same dataset may yield a consensus if the patterns in the dataset have distinctive shapes or characteristics. We recommend the application of a number of clustering algorithms to the same microarray data sets and their evaluation by external measures that employ existing biological knowledge like information about gene ontology.

Acknowledgments This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2006-353-D00006) and the Regional Research Universities Program/Chungbuk BIT Research-Oriented University Consortium. This work was also supported by the Regional Innovation Center Program of the Ministry of Commerce, Industry and Energy through the Bio-Food and Drug Research Center at Konkuk University.

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503–511.

Azuaje, F. (2003). Clustering-based approaches to discovering and visualizing mciroarray data patterns. Brief. Bioinform. *4*, 31–42.

Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., and Kämper, J. (2005). Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics *21*, 1069–1077.

Belacel, N., Čuperlović-Culf, M., Laflamme, M., and Ouellette, R. (2004). Fuzzy J-means and VNS methods for clustering genes from microarray data. Bioinformatics *20*, 1690–1701.

Belacel, N., Wang, Q., and Cuperlovic-culf, M. (2006). Clustering methods for microarray gene expression data. Omics *10*, 507–531.

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. Pac. Symp. Biocomput. *7*, 6–17.

Bezdek, J.C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms (New York: Plenum Press).

Boutros, P.C., and Okey, A.B. (2005). Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. Brief. Bioinform. *6*, 331–343.

Chipman, H. (2006). Hybrid hierarchical clustering with applications to microarray data. Biostatistics *7*, 286–301.

Dembélé, D., and Kastner, P. (2003). Fuzzy C-means for clustering microarray data. Bioinformatics *19*, 973–980.

Do, J.H., and Choi, D.K. (2006). Normalization of microarray data: single-labeled and dual-labeled arrays. Mol. Cells *22*, 254–261.

Dudoit, S., and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol. *3*, research0036.

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

Fowlkes, E.B., and Mallows, C.L. (1983). A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. *78*, 553–584.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. *7*, 601–620.

Fu, L., and Medico, E. (2007). FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics *8*, 3.

Gasch, A.U., and Eisen, M.B. (2002). Exploring the conditional co-regulation of yeast gene expression through fuzzy K-means clustering. Genome Biol. *3*, 1–22.

Gersho, A., and Gray, R. (1992). Vector Quantization and Signal Compression (Boston USA: Kluwer Academic Publishers).

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaassenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531–537.

Hardin, J., Mitani, A., Hicks, L., and VanKoten, B. (2007). A

robust measure of correlation between two genes on a microarray. BMC Bioinformatics *8*, 220.

Heyer, L.J., Kruglyak, S., and Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. Genome Res. *9*, 1106–1115.

Hsu, A.L., Tang, S.-L., and Halgamuge, S.K. (2003). An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data. Bioinformatics *19*, 2131–2140.

Ihmels, J., Friedlander, G., Bergman, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. Nat. Genet. *31*, 370–377.

Ikota, H., Kinjo, S., Yokoo, H., and Nakazato, Y. (2006). Systematic immunohistochemical profiling of 378 brain tumors with 37 antibodies using tissue microarray technology. Acta Neuropathol. (Berl) *111*, 475–482.

Jain, A.K., and Bubes, R.C. (1988). Algorithms for Clustering Data (NJ: Prentice Hall, Englewood Cliffs).

Jiang, D., Pei, J., and Zhang, A. (2003). Towards interactive exploration of gene expression patterns. ACM SIGKDD Explor Newslett *5*, 79–90.

Kaufman, L., and Rousseeuw, P. (1990). Finding groups in data (New York, NY: Wiley).

Kim, K., Zhang, S., Jiang, K., Cai, L., Lee, I.-B., Feldman, L.J., and Huang, H. (2007). Measuring similarities between gene expression profiles through new data transformation. BMC Bioinformatics *8*, 29.

Kohonen, T. (1990). The self-organizing map. Proc. IEEE *78*, 1464–1479.

Krishna, K., and Narasimha Murty, M. (1999). Genetic K-means algorithm. IEEE Trans. Syst. Man Cybern. Part B *29*, 433–439.

Lu, Y., Lu, S., Fotouhi, F., Deng, Y., and Brown, S.J. (2004a). FGKA: a fast genetic K-means clustering algorithm. Proceedings of the 2004 ACM symposium on Applied computing (SAC), Nicosia, Cyprus.

Lu, Y., Lu, S., Fotouhi, F., Deng, Y., and Brown, S.J. (2004b). Incremental genetic K-means algorithm and its application in gene expression data analysis. BMC Bioinformatics *5*, 172.

Macnaughton-Smith, P., Williams, W.T., Dale, M.B., and Mockett, L.G. (1964). Dissimilarity analysis: a new technic of hierarchical subdivision. Nature *202*, 1034–1035.

Raychaudhuri, S., Sutphin, P.D., Chang, J.T., and Altman, R.B. (2001). Basic microarray analysis: grouping and feature reduction. Trends Biotechnol. *19*, 189–193.

Ressom, H., Wang, D., and Natarajan, P. (2003). Adaptive double self-organizing maps for clustering gene expression profiles. Neural Netw. *16*, 633–640.

Sheng, Q., Moreau, Y., and De Moor, B. (2003). Biclustering microarray data by Gibbs sampling. Bioinformatics *19* (Suppl. 2), ii196–ii205.

Slonim, D.K. (2002). From patterns to pathways: gene expression data analysis comes of age. Nat. Genet. *32* (Suppl. 2), 502–508.

Su, M., and Chang, H. (2001). A new model of self-organizing neural networks and its application in data projection. IEEE Trans. Neural Netw. *12*, 153–158.

Tan, P.N., Steinbach, M., and Kumar, V. (2005). Introduction to Data Mining (Boston: Addison-Wesley).

Tseng, G. (2004). A comparative review of gene clustering in expression profile. eighth international conference on control, automation, robotics and vision (ICARCV). 1320– 1324.

Van der Laan, M., Pollard, K.S., and Bryan, J. (2003). A new partitioning around medoids algorithm. J. Stat. Comput. Simul. *73*, 575–584.

Wang, J., Bo, T.H., Jonassen, I., and Hovig, E. (2003). Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. BMC Bioinformatics *4*, 60.

Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., and Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. Proc. Natl. Acad. Sci. USA *95*, 334–339.

Woolf, P.J., and Wang, Y. (2000). A fuzzy logic approach to analyzing gene expression data. Physiol. Genomics *3*, 9–15.

Wu, F.X., Zhang, W.J., and Kusalik, A.J. (2006). Determination of the minimum number of microarray experiments for discovery of gene expression patterns. BMC Bioinformatics *7* (Suppl. 4), S13.

Xing, B., Greenwood, C.M., and Bull, S.B. (2007). A hierarchical clustering method for estimating copy number variation. Biostatistics *8*, 632–653.