BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 9 2005, pages 1927–1934
doi:10.1093/bioinformatics/bti251

Gene expression

Detecting clusters of different geometrical shapes in microarray gene expression data

Dae-Won Kim[1], Kwang H. Lee[1,2] and Doheon Lee[1,*]

[1]Department of BioSystems and Advanced Information Technology Research Center, and [2]Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, 373–1 Guseong-dong, Yuseong-gu, Daejeon, 305–701, Korea

Received on August 13, 2004; revised on December 16, 2004; accepted on December 23, 2004
Advance Access publication January 10, 2005

ABSTRACT

Motivation: Clustering has been used as a popular technique for finding groups of genes that show similar expression patterns under multiple experimental conditions. Many clustering methods have been proposed for clustering gene-expression data, including the hierarchical clustering, k-means clustering and self-organizing map (SOM). However, the conventional methods are limited to identify different shapes of clusters because they use a fixed distance norm when calculating the distance between genes. The fixed distance norm imposes a fixed geometrical shape on the clusters regardless of the actual data distribution. Thus, different distance norms are required for handling the different shapes of clusters.

Results: We present the Gustafson–Kessel (GK) clustering method for microarray gene-expression data. To detect clusters of different shapes in a dataset, we use an adaptive distance norm that is calculated by a fuzzy covariance matrix (F) of each cluster in which the eigenstructure of F is used as an indicator of the shape of the cluster. Moreover, the GK method is less prone to falling into local minima than the k-means and SOM because it makes decisions through the use of membership degrees of a gene to clusters. The algorithmic procedure is accomplished by the alternating optimization technique, which iteratively improves a sequence of sets of clusters until no further improvement is possible. To test the performance of the GK method, we applied the GK method and well-known conventional methods to three recently published yeast datasets, and compared the performance of each method using the *Saccharomyces* Genome Database annotations. The clustering results of the GK method are more significantly relevant to the biological annotations than those of the other methods, demonstrating its effectiveness and potential for clustering gene-expression data.

Availability: The software was developed using Java language, and can be executed on the platforms that JVM (Java Virtual Machine) is running. It is available from the authors upon request.

Contact: dhlee@bisl.kaist.ac.kr

Supplementary information: Supplementary data are available at http://dragon.kaist.ac.kr/gk

1 INTRODUCTION

The DNA microarray technology has enabled biologists to monitor the expression levels of thousand of genes in parallel under multiple experimental conditions. Since the diauxic shift (DeRisi *et al*., 1997), sporulation (Chu *et al*., 1998) and the cell cycle (Cho *et al*., 1998) in the yeast *Saccharomyces cerevisiae* were explored, many experiments have been made to monitor the gene-expression levels of various organisms during some biological process.

The present study focuses on the application of clustering methods for analyzing gene-expression datasets that are comprised of the expression patterns of specific environmental conditions, rather than time-course type of data. Since Eisen *et al*. (1998) first used the hierarchical clustering method to find groups of coexpressed genes, numerous methods have been studied for clustering gene-expression data: self-organizing map (SOM) (Tamayo *et al*., 1999), k-means clustering (Tavazoie *et al*., 1999), simulated annealing (Lukashin and Fuchs, 2001), graph-theoretic approach (Xu *et al*., 2001), mutual information approach (Steuer *et al*., 2002), fuzzy c-means clustering (Dembele and Kastner, 2003), kernel hierarchical clustering (Qin *et al*., 2003), diametrical clustering (Dhilon *et al*., 2003), quantum clustering (QC) with singular value decomposition (Horn and Axel, 2003), bagged clustering (Dudoit and Fridlyand, 2003) and CLICK (Sharan *et al*., 2003).

Of the clustering methods reported to date, the most widely used methods are the hierarchical, k-means and SOM due to their superiority and availability of several free software tools. However, these conventional clustering methods have a number of limitations. As Yeung *et al*. (2001), and Gibbons and Roth (2002) pointed out, the performance of the hierarchical clustering method was close to random, despite its wide-usage, and worse than the k-means and SOM. Such cases arise because the hierarchical clustering is likely to produce one single large cluster and several singletons. Furthermore, this method suffers from the defect that it can never repair what was done in previous steps.

The partitional clustering methods such as the k-means and SOM are also problematic when the clusters differ in shape (Babuska, 1998; Bezdek *et al*., 1999; Jain *et al*., 1999). The shape of clusters is determined by a distance norm, which is a fundamental factor when developing a clustering method. Let x_j be the expression vector for the j-th gene over different environmental conditions (or samples), and let v_i be the i-th cluster centroid (prototype). Then a typical distance norm between x_j and v_i is represented as:

$$D_{ij}^2 = \|x_j - v_i\|_A^2 = (x_j - v_i)^{\mathrm{T}} A (x_j - v_i) \qquad (1)$$

where A is a symmetric and positive definite matrix. Thus different norms can be induced by the choice of the matrix A. The Euclidean

*To whom correspondence should be addressed.

norm is induced when $A = I$ where I is an identity matrix. The Mahalonobis norm is induced when $A = M^{-1}$ where M^{-1} is the inverse of the covariance matrix of patterns in the system.

Although many clustering methods have been studied in the literature, a common limitation of conventional methods is to use a fixed distance norm for finding clusters; this fixed norm imposes a fixed geometrical structure and finds clusters of that shape even if they are not present. The Euclidean norm-based methods find only spherical shape of clusters, and the Mahalanobis norm-based methods find only ellipsoidal ones even if those shapes of clusters are not present in a dataset. For example, let us consider a dataset distributed in four clusters, plotted in Figure 1a, which is composed of the expression vectors of 400 genes measured at two conditions. Although the clusters differ in shape, the conventional clustering method using the Euclidean distance is limited to identify the four clusters because it imposes a spherical shape on the clusters with no regard to the actual data distribution, as depicted in Figure 1b. In such cases, it is of no help to use another distance norm.

To tackle the addressed problems, different distance norms are required for handling the different shapes of clusters; specifically, different As should be applied to the different clusters. In the present work, we present the Gustafson–Kessel (GK) method based on an adaptive distance norm (Gustafson and Kessel, 1979; Babuska, 1998; Bezdek *et al*., 1999) for clustering gene-expression data. Different norm-inducing matrix As are adapted by estimating the data covariance, and used to guess the associated structure of the data in each individual cluster. Recently, Guthke *et al*. (2002) showed that the GK method gave higher accuracy than other methods in predicting the gene function of *Escherichia coli*. However, the work of Guthke *et al*. was not very extensive and the experimental comparisons were made using a very small subset of data (265 genes). In the present study we provide a comprehensive analysis on the GK method, and show the superiority of the GK method to other methods through extensive experiments with various datasets. The remainder of this paper is organized as follows: Section 2 describes the formulation of the GK method; Section 3 highlights the potential of the GK method through several tests on the yeast datasets; and Section 4 presents our concluding remarks.

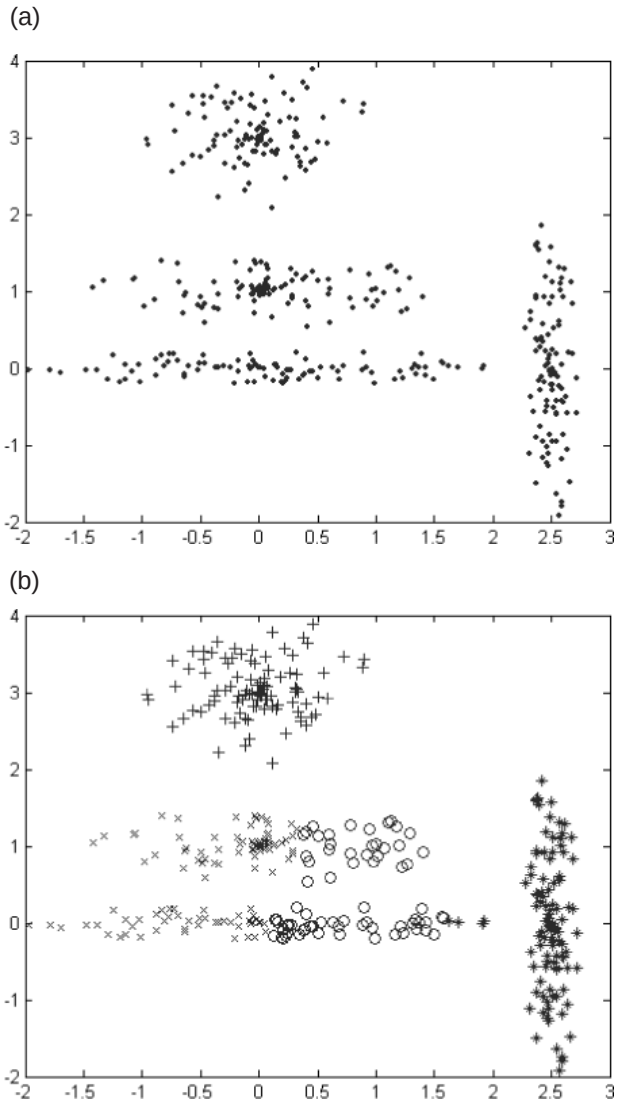

Fig. 1. Clustering of the conventional clustering method using the Euclidean distance: (**a**) plot of the original dataset measured at two conditions in which four clusters have different shapes; (**b**) conventional method detects clusters of spherical shape regardless of the actual data distribution. The horizontal axis represents the expression value at the first condition. The vertical axis represents the expression value at the second condition.

2 THE GUSTAFSON–KESSEL METHOD

The GK method generates a fuzzy partition that provides a degree of membership of each data point to a given cluster. The objective of the GK method is to classify a set of data points $X = \{x_1, x_2, \ldots, x_n\}$ in p-dimensional space into k disjoint and homogeneous clusters represented as $C = \{C_1, C_2, \ldots, C_k\}$. To detect clusters of different geometrical shapes in a dataset, this method introduces an adaptive distance norm for each cluster. Each cluster C_i has its own norm-inducing matrix A_i, which affects the distance norm given as the following.

$$D^2_{ijA_i} = \|x_j - v_i\|^2_{A_i} = (x_j - v_i)^{\mathrm{T}} A_i (x_j - v_i), \qquad (2)$$

where $V = [v_1, v_2, \ldots, v_k]$ is a vector of the centroids of the fuzzy clusters $C_1, C_2, \ldots, C_k$. Use of the matrix A_i makes it possible for each cluster to adapt the distance norm to the geometrical structure of the data at each iteration. Based on the norm-inducing matrices, the objective of the GK method is obtained by minimizing the function J_m.

$$J_m(U, V, A : X) = \sum_{i=1}^{k} \sum_{j=1}^{n} (\mu_{ij})^m D^2_{ijA_i}, \qquad (3)$$

where

$$A = (A_1, A_2, \ldots, A_k)$$

is a k-tuple of the norm-inducing matrices,

$$U = [\mu_{ij}] = \begin{bmatrix} \mu_{11} & \mu_{12} & \ldots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \ldots & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{k1} & \mu_{k2} & \ldots & \mu_{kn} \end{bmatrix} \qquad (4)$$

is a fuzzy partition matrix of X satisfying the following constraints,

$$\begin{aligned} &\mu_{ij} \in [0,1], \quad 1 \le i \le k, \;\; 1 \le j \le n, \\ &\sum_{i=1}^{k} \mu_{ij} = 1, \quad 1 \le j \le n, \\ &0 < \sum_{j=1}^{n} \mu_{ij} < n, \quad 1 \le i \le k, \end{aligned} \tag{5}$$

and

$$m \in [1, \infty) \tag{6}$$

is a weighting exponent that controls the membership degree μ_{ij} of each data point x_j to the cluster C_i. The choice of appropriate m value is of importance because the final clusters may vary depending on the m value selected. As $m \to 1$, J_1 produces a hard partition where $\mu_{ij} \in \{0, 1\}$. As m approaches infinity, J_∞ produces a maximum fuzzy partition where $\mu_{ij} = 1/c$.

To obtain a feasible solution by minimizing Equation (3), the additional constraint is required for A_i.

$$\det(A_i) = \rho_i, \quad \rho_i > 0, \quad 1 \le i \le k, \tag{7}$$

where ρ_i is a cluster volume for each cluster. Equation (7) guarantees that A_i is positive-definite, indicating that A_i can be varied to find the optimal shape of cluster with its volume fixed. Using the Lagrange multiplier method, minimization of Equation (3) with respect to A_i produces the following equation.

$$A_i = [\rho_i \det(F_i)]^{1/p} F_i^{-1}, \tag{8}$$

where

$$F_i = \frac{\sum_{j=1}^{n} (\mu_{ij})^m (x_j - v_i)(x_j - v_i)^{\mathrm{T}}}{\sum_{j=1}^{n} (\mu_{ij})^m} \tag{9}$$

is the fuzzy covariance matrix of cluster C_i. The set of fuzzy covariance matrices is represented as a k-tuple of $F = (F_1, F_2, \ldots, F_k)$. To show A_i in Equation (8) satisfies the constraint of a symmetric and positive-definite matrix, suppose that there are p linearly independent data points $\xi \in R^p$ in the dataset. Then, the matrices $\xi\xi^{\mathrm{T}}$ are symmetric and positive semi-definite and also their weighted sum (F_i), and hence A_i is symmetric and positive-definite.

There is no general agreement on what value to use for ρ_i; without any prior knowledge, a rule of thumb that many investigators use is $\rho_i = 1$ in practice (Bezdek *et al*., 1999). Moreover, based on the notion that ρ_i represents the cluster volume for each cluster, in the present study we estimated ρ_i, shown in the following equation, by exploiting the definition on the volume of fuzzy cluster C_i (Dumitrescu *et al*., 2000), making the GK method to be fully operational.

$$\rho_i = \sqrt{\det(F_i)} \tag{10}$$

Under this formulation, the fixed norm $D_{ij}^2 = \|x_j - v_i\|_A^2$ calculated for the distance between x_j and v_i is replaced in the GK method

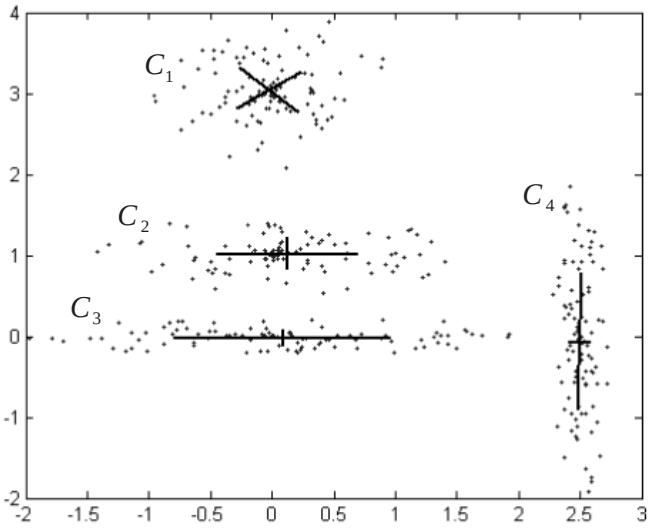

Fig. 2. The GK method successfully identifies the four clusters by exploiting the covariance matrices F_i for the clusters. Dots represent data points and thick lines represent the hyperellipsoides given by the eigenstructures of clusters. The horizontal axis represents the expression value at the first condition. The vertical axis represents the expression value at the second condition.

with the following distance,

$$\begin{aligned} D_{ijA_i}^2 &= \|x_j - v_i\|_{A_i}^2 \\ &= (x_j - v_i)^{\mathrm{T}} \left[\sqrt{\det(F_i)}\det(F_i)\right]^{1/p} F_i^{-1}(x_j - v_i) \\ &= \sqrt{\det(F_i)}\det(F_i)^{1/p}\|x_j - v_i\|_{F_i^{-1}}^2. \end{aligned} \tag{11}$$

The eigenstructure of the covariance matrices F_i identifies the shape of clusters. Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the eigenvalues of F_i, and they are arranged in decreasing order; then the length of the r-th axis of each cluster is given by $\sqrt{\lambda_r}$ for $r = 1, \ldots, p$. Figure 2 shows the eigenstructures of the clusters for the dataset from Figure 1a. The thick lines represent the eigenvectors of F_i, which correctly detect the shape of the clusters. The cluster centroids and eigenvalues obtained by the GK method for the four clusters are listed in Table 1. The last column is the ratio of the length of the first axis ($\sqrt{\lambda_1}$) to the length of the second axis ($\sqrt{\lambda_2}$), which describes the overall distribution of data. We can see that the ratio value of the cluster C_1 is about 1.0, indicating that C_1 has a spherical shape. In contrast, the cluster C_4 has a ratio value of 9.36, indicating that the shape of C_4 would be an elongated ellipsoid. The eigenvalue calculations, shown in Figure 2 and Table 1, indicate that the shape of each cluster is recognized by its eigenstructure, and therefore the adaptive distance norm is capable of identifying the inherent structure of the data.

The saddle point of J_m is obtained by considering the constraint Equation (5) as the Lagrange multipliers:

$$\begin{aligned} &\nabla J_m(U, V, A, \lambda : X) \\ &\quad = \sum_{i=1}^{k}\sum_{j=1}^{n} (\mu_{ij})^m D_{ijA_i}^2 + \sum_{j=1}^{n} \alpha_j \left[\sum_{i=1}^{k} \mu_{ij} - 1\right] \end{aligned} \tag{12}$$

Table 1. The cluster centroids (v_i) and eigenvalues (λ_i) obtained by the GK method for the dataset from Figure 2

Cluster	Centroid	λ_1	λ_2	$\sqrt{\lambda_1}/\sqrt{\lambda_2}$
C_1	$[-0.03, 3.05]^{\mathrm{T}}$	0.14	0.12	1.06
C_2	$[0.12, 1.02]^{\mathrm{T}}$	0.35	0.04	3.15
C_3	$[0.09, -0.02]^{\mathrm{T}}$	0.82	0.01	9.11
C_4	$[2.50, -0.01]^{\mathrm{T}}$	0.75	0.01	9.36

and by setting $\nabla J_m = 0$. If $D^2_{ijA_i} > 0$ for all i, j and $m > 1$, then (U, V) may minimize J_m only if,

$$\mu_{ij} = \left[\sum_{l=1}^{k}\left(\frac{D_{ijA_i}}{D_{ilA_i}}\right)^{2/(m-1)}\right]^{-1}, \quad 1 \le i \le k,\ 1 \le j \le n, \quad (13)$$

and

$$v_i = \frac{\sum_{j=1}^{n}(\mu_{ij})^m x_j}{\sum_{j=1}^{n}(\mu_{ij})^m}, \quad 1 \le i \le k. \quad (14)$$

This solution also satisfies the remaining constraints of Equation (5).

The GK method iteratively improves a sequence of sets of clusters until no further improvement in $J_m(U, V, A : X)$ is possible. This type of iteration is often referred to as the alternating optimization (AO) scheme. It loops through the estimates for $V_{t-1} \to U_t \to V_t$ (where t is the iteration step) and terminates on $\|V_t - V_{t-1}\| \le \epsilon$. Equivalently, the initialization of the algorithm can be done on U_0, and the iterates become $U_{t-1} \to V_t \to U_t$, with the termination criterion $\|U_t - U_{t-1}\| \le \epsilon$. This way of alternating optimization makes the GK method be less prone to falling into local minima than the k-means and SOM methods because it makes soft decisions in each iteration through the use of membership functions.

Algorithm 1. Gustafson–Kessel method

Given the dataset $X = \{x_1, \ldots, x_n\}, x_j \in R^p$, the number of clusters (k), the weighting exponent (m), and the termination criterion (ϵ), this method finds k disjoint and homogeneous clusters.

1. Initialize $U_{t-1} = [\mu_{ij}^{(t-1)}]$ (initially, $t \leftarrow 1$) of x_j belonging to cluster C_i for $1 \le i \le k, 1 \le j \le n$ such that:
$$\sum_{i=1}^{k} \mu_{ij} = 1.0.$$
2. Update the cluster centroids $V_t = [v_1^{(t)}, \ldots, v_k^{(t)}]$ for $1 \le i \le k$ using:
$$v_i^{(t)} = \frac{\sum_{j=1}^{n}(\mu_{ij}^{(t-1)})^m x_j}{\sum_{j=1}^{n}(\mu_{ij}^{(t-1)})^m}.$$
3. Update the cluster covariance matrices F_i for $1 \le i \le k$ using:
$$F_i = \frac{\sum_{j=1}^{n}(\mu_{ij}^{(t-1)})^m (x_j - v_i^{(t)})(x_j - v_i^{(t)})^{\mathrm{T}}}{\sum_{j=1}^{n}(\mu_{ij}^{(t-1)})^m}.$$
4. Compute the distances between x_j and $v_i^{(t)}$ for $1 \le i \le k, 1 \le j \le n$ using:
$$D^2_{ijA_i} = \rho_i \det(F_i)^{1/p}\|x_j - v_i^{(t)}\|^2_{F_i^{-1}}.$$
5. Update $U_t = [\mu_{ij}^{(t)}]$ by the following procedure. For each $x_j, 1 \le j \le n$,
   (a) if $D_{ijA_i} > 0, 1 \le i \le k$, then update the membership of x_j at t by:
$$\mu_{ij}^{(t)} = \left[\sum_{l=1}^{k}\left(\frac{D_{ijA_i}}{D_{ilA_i}}\right)^{2/(m-1)}\right]^{-1},$$
   (b) if $D_{ijA_i} = 0$ for some $i \in I \subseteq 1, \ldots, k$, then for all $i \in I$, set $\mu_{ij}^{(t)}$ to be between $[0, 1]$ such that:
$$\sum_{i \in I} \mu_{ij}^{(t)} = 1, \text{and}$$
$$\text{set } \mu_{ij}^{(t)} = 0 \text{ for other } i \notin I.$$
6. If $\|U_t - U_{t-1}\| \le \epsilon$, then stop; otherwise, $t \leftarrow t + 1$ and go to Step 2.

Algorithm 1 shows the procedural steps of the GK method for clustering the $n \times p$ gene-expression data where n is the number of genes and p is the number of experiments. A singularity can occur at Step 4 when the number of data is small or when the data within a cluster are much linearly correlated (Babuska, 1998). In such cases, the cluster covariance matrix becomes singular and cannot be inverted; thus we cannot compute the distances at Step 4. To deal with such cases, in the present study, the membership degrees are set to $\mu_{ij} = 0$ for non-singular classes, and arbitrary values for singular classes subject to the constraints of Equation (5).

THEOREM 1. *The GK method given in Algorithm 1 converges in a finite number of iterations.*

PROOF 1. We first show that a saddle point of J_m appears at most once by the GK method before it stops. Suppose that this is not true, i.e. $U_{t_1} = U_{t_2}$ for some t_1, t_2 where $t_1 \neq t_2$. By the AO scheme, we get V_{t_1+1} and V_{t_2+1} as optimal solutions for $U = U_{t_1}$ and $U = U_{t_2}$, respectively. Therefore, we have

$$\begin{aligned} J_m(U_{t_1}, V_{t_1+1}) &= J_m(U_{t_2}, V_{t_1+1}) \quad (\text{since } U_{t_1} = U_{t_2}) \\ &= J_m(U_{t_2}, V_{t_2+1}) \end{aligned} \quad (15)$$

However, the sequence $J_m(\cdot, \cdot)$ generated by the GK method is strictly decreasing (Selim and Ismail, 1984). Hence equation (15) is false and $U_{t_1} \neq U_{t_2}$. Since there are a finite number of saddle points of J_m (Selim and Ismail, 1984), the algorithm will converge in a finite number of iterations.

A similar proof concerning the convergence of the k-means-type algorithms to a local minimum has been stated by Selim and Ismail (1984).

3 RESULTS

3.1 Datasets and implementation parameters

To test the effectiveness with which the GK method clusters gene-expression data, we applied the GK method and five well-known clustering methods to three recently published yeast datasets and compared the performance of each method. In the present study, we used EXPANDER (Sharan *et al.*, 2003) software that implements many clustering methods, of which we investigated the results of

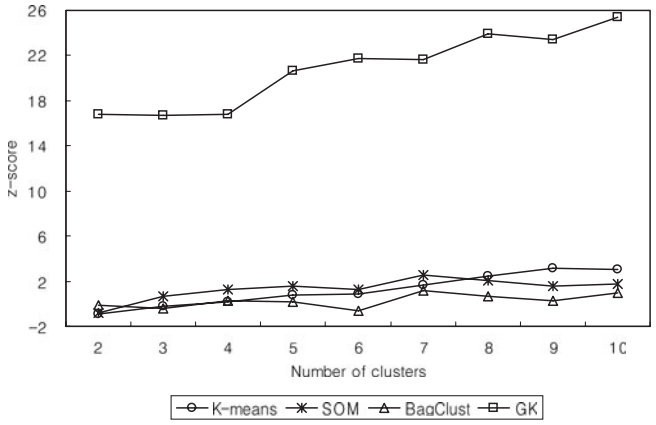

Fig. 3. Comparison of the clustering performance of the k-means, SOM, BagClust and GK methods for the yeast PHO-regulation dataset of Ogawa *et al*. (2000). The horizontal axis represents the number of clusters given, the vertical axis represents the z-score. The z-score is computed with the relation between a clustering result and the SGD functional annotation of the genes in the cluster (Gibbons and Roth, 2002).

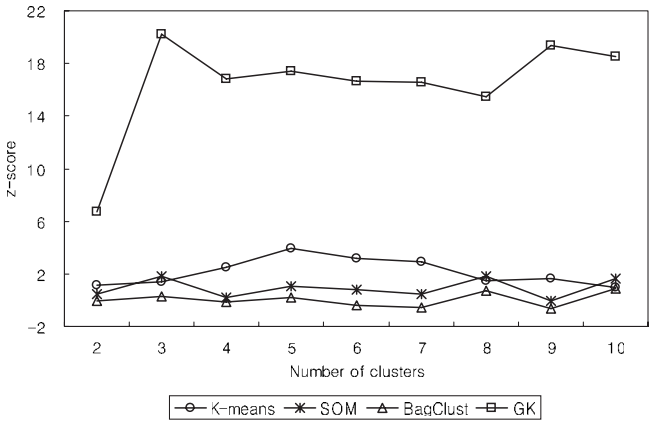

Fig. 4. Comparison of the clustering performance of the k-means, SOM, BagClust, and GK methods for the yeast ATP-regulation dataset of Mizuguchi *et al*. (2004). The horizontal axis represents the number of clusters given, the vertical axis represents the z-score. The z-score is computed with the relation between a clustering result and the SGD functional annotation of the genes in the cluster (Gibbons and Roth, 2002).

the k-means, SOM, and CLICK methods. Along with these, we examined the results of the bagged clustering (BagClust) (Dudoit and Fridlyand, 2003) and the QC (Horn and Axel, 2003).

The datasets employed were the yeast PHO-regulation dataset of Ogawa *et al*. (2000), the yeast ATP-regulation dataset of Mizuguchi *et al*. (2004), and the yeast Calcineurin-regulation dataset of Yoshimoto *et al*. (2002). The Ogawa's PHO dataset contains the expression profiles of 6013 yeast genes measured at eight PHO-regulated experiments. The Mizuguchi's ATP dataset consists of the expression levels of the 6215 yeast genes measured at three different ATPase experiments. The Yoshimoto's Calcineurin dataset contains the expression profiles of 6102 yeast genes at 24 experiments by the presence and absence of Ca^{2+}, Na^{+}, CRZ1 and FK506. These three datasets were obtained from a public website containing various published large-scale yeast datasets (http://sgdlite.princeton.edu/download/yeast_datasets/).

In these experiments, the parameters used in the GK method were $\epsilon = 0.001$, $m = 2.5$, and $\rho = 1$; these values were chosen because they have been overwhelmingly favored in many studies (Bezdek *et al*., 1999). In the tests reported here, we analyzed the performance of each method under changes in the number of clusters of k, with varied from $k = 2$ to $k = 10$.

3.2 Performance comparison

In the present study, the clustering results were assessed using two validation measures: z-score and Jaccard score.

First, the z-score (Gibbons and Roth, 2002) is calculated by investigating the relation between a clustering result and the functional annotation of the genes in the cluster. To achieve this, the score uses the *Saccharomyces* Genome Database (SGD) annotation of the yeast genes, along with the gene ontology developed by the Gene Ontology Consortium (Ashburner *et al*., 2000; Issel-Tarver *et al*., 2002). A higher score of z indicates that genes are better clustered by function, indicating a more biologically significant clustering result.

Figure 3 shows the clustering results of the k-means, SOM, BagClust and GK methods for the yeast PHO dataset. The z-scores of the four clustering methods are plotted with respect to $k = 2, 3, \ldots, 10$. The k-means method gave z-scores of ranging from -0.9 to 3.1, and SOM gave scores from -0.8 to 2.5. The z-scores of the BagClust were ranged from -0.6 to 1.1. The SOM method outperformed the k-means and BagClust methods at low k-values, and the k-means method showed better performance than the SOM and BagClust methods at high k-values. Compared to these three methods, the GK method provided superior clustering performance over a wide range of k-values; the z-scores were varied from 16.8 to 25.3.

Figure 4 shows the clustering results of the k-means, SOM, BagClust and GK methods for the yeast ATP dataset. The k-means method gave z-scores of ranging from 1.0 to 3.9. On the whole, the SOM and BagCluster showed similar tendency for all k values. In comparison to these methods, it is evident that the GK clustering method shows markedly better performance, giving z-scores of >15.0 for $k > 3$. This result is in agreement with the work of Mizuguchi *et al*. (2004), which reported that the optimal k-value lies ~ 3 for this dataset.

The clustering results achieved by the six clustering methods for each of the yeast PHO and ATP datasets are listed in Table 2. For the PHO dataset, the k-means method showed the best z-score of 3.15 at $k = 9$. The SOM and BagClust methods provided best z-values of 2.53 and 1.14 at $k = 7$, respectively. The CLICK and QC methods yielded z-values of 1.65 and 1.46 at $k = 31$ and $k = 16$, respectively; these two methods automatically find the optimal k. In contrast, the best value of $z = 25.38$ of the GK method was obtained at $k = 10$. Furthermore, we tested the performance of the GK method for two different values of $\rho = 1$ and $\rho = \sqrt{\det F}$. It is observed that the GK method for these two ρ values showed similarly better performance than other methods; the z-scores were varied from 5.77 to 29.80 over $k = 2, 3, \ldots, 10$. The best scores of the GK method were $z = 25.38$ for $\rho = 1$, and $z = 29.80$ for $\rho = \sqrt{\det F}$. For the ATP dataset, the best z-values of the k-means and SOM methods were 3.97 and 1.84 at $k = 5$ and $k = 3$, respectively. The BagClust and CLICK provided the best $z = 0.87$ and $z = -1.03$ at $k = 10$, respectively,

Table 2. Summary of clustering results obtained by six clustering methods for the yeast PHO-regulation dataset and the ATP-regulation dataset

Dataset	Genes/conditions	Method	z-score	k
Yeast PHO	6013/8	k-means	3.15	9
		SOM	2.53	7
		BagClust	1.14	7
		CLICK	1.65	31
		QC	1.46	16
		GK ($\rho = 1$)	25.38	10
		GK ($\rho = \sqrt{\det F}$)	29.80	6
Yeast ATP	6215 / 3	k-means	3.97	5
		SOM	1.84	3
		BagClust	0.87	10
		CLICK	−1.03	10
		QC	0.94	8
		GK ($\rho = 1$)	20.26	3
		GK ($\rho = \sqrt{\det F}$)	29.70	4

For each dataset, the highest z-score of each method from Figures 3 and 4 are specified.

whereas the QC method yielded the best $z = 0.94$ at $k = 8$. The GK method with respect to $\rho = 1, \sqrt{\det F}$ provided significantly better clustering performance than other methods, giving z-scores of >20.0.

In addition to the assessment using the z-score, we quantified the clustering result of each method using the Jaccard and Minkowski scores. Let T be the 'true' solution and C the solution a clustering algorithm generated. Let n_{11} be the number of pairs of data that are in the same cluster in both T and C. Let n_{10} be the number of pairs that are in the same cluster only in T, and n_{01} be the number of pairs that are in the same cluster only in C. Then the Jaccard score is defined as

$$S_J(T, C) = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}. \tag{16}$$

A higher value of $S_J(T, C)$ indicates a better clustering result; the Minkowski score is defined as

$$S_M(T, C) = \sqrt{\frac{n_{01} + n_{10}}{n_{11} + n_{10}}}. \tag{17}$$

In this case, a lower value of $S_M(T, C)$ indicates a well-clustered result. To measure the quality of clustering with these two scores, we applied five clustering methods to the yeast Calcineurin dataset that provides a putative true solution obtained through manual inspection by Yoshimoto *et al*. (2002). Table 3 lists the comparison results of five clustering methods for $k = 3$. The GK method is the most effective of the methods considered; it provides the highest Jaccard score, with the lowest Minkowski score. The k-means and BagCluster methods showed better scores than the SOM and QC methods, and the QC method proved the most ineffective of the methods considered.

The results of the comparison calculations indicate that the GK method gave markedly better clustering performance than the other five methods considered, highlighting the effectiveness and potential of the GK method.

3.3 Functional enrichment

The enriched functional categories for each cluster obtained by the GK method on the yeast PHO and ATP datasets are listed in Tables 4 and 5 respectively. The enrichment of each GO category in each of the clusters was calculated by its P-value. To compute the P-value, we employed the hypergeometric distribution used by Tavazoie *et al*. (1999) and Dembele and Kastner (2003) in order to obtain the probability of observing the number of genes from a specific GO functional category within each cluster. More detailed explanation on this P-value can be found in Tavazoie *et al*. (1999) and Dembele and Kastner (2003). A low P-value indicates that the genes belonging to the enriched functional categories are biologically significant in the corresponding clusters. In the present study, only functional categories with P-value $<5.0 \times 10^{-7}$ are reported.

Of the ten clusters obtained for the yeast PHO dataset (Table 4), the cluster C_9 contains several enriched categories on 'ribosome'. The highly enriched category in cluster C_9 is the 'ribonucleoprotein complex' with P-value of 4.83×10^{-35}. The GO category 'ribosome'

Table 3. Comparison of the clustering performance of the k-means, SOM, BagClust, QC and GK methods for the yeast Calcineurin-regulation dataset of Yoshimoto *et al*. (2002)

Dataset	Method	Jaccard	Minkowski
Yeast Calcineurin	k-means	0.55	0.72
	SOM	0.49	0.79
	BagClust	0.56	0.74
	QC	0.41	0.90
	GK	0.59	0.66

The number of clusters is $k = 3$. The Jaccard and Minkowski scores are computed with the putative solution of Yoshimoto *et al*. (2002).

Table 4. Enrichment of GO categories in each of the clusters obtained by the GK method for the yeast PHO-regulation dataset of Ogawa *et al*. (2000)

Cluster	GO category	GO number	P-value
C_1	Alcohol metabolism	GO:0006066	4.31E−16
	Sterol metabolism	GO:0016125	1.15E−10
	Steroid metabolism	GO:0008202	1.91E−11
	Ergosterol metabolism	GO:0008204	1.14E−09
	Lipid biosynthesis	GO:0008610	6.14E−09
	Lipid metabolism	GO:0006629	7.34E−08
C_2	RNA processing	GO:0006369	7.30E−44
	RNA metabolism	GO:0016070	8.01E−41
	35S primary transcript processing	GO:0006365	1.49E−16
	Processing of 20S pre-rRNA	GO:0030490	2.76E−14
	RNA modification	GO:0009451	1.43E−10
	SnoRNA binding	GO:0030515	4.13E−10
C_9	Ribonucleoprotein complex	GO:0030529	4.83E−35
	Ribosome	GO:0005840	4.20E−45
	Cytosol	GO:0005829	4.04E−43
	Cytosolic ribosome	GO:0005830	6.33E−43
	Protein biosynthesis	GO:0006412	7.07E−34
	Cytosolic large ribosomal subunit	GO:0005842	9.61E−28
	Large ribosomal subunit	GO:0015934	9.01E−27
	Small ribosomal subunit	GO:0015935	3.12E−19
	Cytosolic small ribosomal subunit	GO:0005843	7.52E−18

The number of clusters is ten. Only functional categories with P-values less than 5.0E−7 are reported.

Table 5. Enrichment of GO categories in each of the clusters obtained by the GK method for the yeast ATP-regulation dataset of Mizuguchi *et al*. (2004)

Cluster	GO category	GO number	*P*-value
C_1	Endoplasmic reticulum	GO:0005783	1.58E−12
	Amino acid metabolism	GO:0006520	5.19E−09
	Amine metabolism	GO:0009308	9.33E−09
	Amino acid and derivative metabolism	GO:0006519	1.35E−08
	Carboxylic acid metabolism	GO:0019752	5.01E−08
	Amine biosynthesis	GO:0009309	1.02E−07
C_2	Cytosolic ribosome	GO:0005830	8.40E−45
	Ribosome	GO:0005840	2.23E−30
	Cytosolic large ribosomal subunit	GO:0005842	2.34E−29
	Ribonucleoprotein complex	GO:0030529	3.16E−29
	Protein biosynthesis	GO:0006412	5.97E−25
	Large ribosomal subunit	GO:0015934	1.66E−21
	Eukaryotic 48S initiation complex	GO:0016283	5.28E−17
	Cytosolic small ribosomal subunit	GO:0005843	5.28E−17
	Eukaryotic 43S preinitiation complex	GO:0016282	2.84E−16
C_3	Oxidative phospholylation	GO:0006119	5.70E−14
	Energy derivation by oxidation of organic compounds	GO:0015980	4.75E−09
	Hydrogen-translocating F-type ATPase complex	GO:0045255	4.89E−09
	Proton-transporting ATP synthase complex	GO:0005753	4.89E−09
	Proton-transporting two-sector ATPase complex	GO:0016469	4.89E−09
	ATP metabolism	GO:0046034	4.89E−09
	ATP synthesis coupled proton transport	GO:0015986	4.89E−09
	Nucleoside phosphate metabolism	GO:0006753	4.89E−09
	Proton-transporting ATP synthase complex	GO:0045259	4.89E−09
	Carbohydrate metabolism	GO:0005975	1.01E−08
	Purine ribonucleoside triphosphate biosynthesis	GO:0009206	1.39E−07
	Purine nucleoside triphosphate biosynthesis	GO:0009145	1.39E−07

The number of clusters is three. Only functional categories with *P*-values less than 5.0E−7 are reported.

is also highly enriched in this cluster with *P*-value of 4.20×10^{-45}. The cluster C_2 contains an enriched category 'RNA processing' with *P*-value of 7.30×10^{-44}. In the case of the ATP dataset (Table 5), the cluster C_3 contains the yeast genes corresponding to the ATP-involved GO biological process. The highly enriched categories in cluster C_3 are the 'ATP metabolism' with *P*-value of 4.89×10^{-9} and the 'purine ribonucleoside triphosphate biosynthesis' with *P*-value of 1.39×10^{-7}. From the results of Tables 4 and 5, we see that the cluster obtained by the GK method shows a high enrichment of functional categories.

Besides, as mentioned earlier, the GK method produces a fuzzy clustering result, which provides a convenient way of selecting genes showing tight association to given clusters (Dembele and Kastner, 2003). Dembele and Kastner referred to this process as 'restricted' selection. Similar to the work of Dembele and Kastner, we removed the most loosely associated genes in each cluster using a threshold-based selection; specifically, genes with the highest membership degree $\mu_{ij} > 0.5$ were selected. To see the effect of the restricted selection, we compared the percentage of genes in the raw cluster and its corresponding restricted cluster for GO functional categories. Table 6 lists a comparison result for the cluster C_9 shown from Table 4. In comparison to the raw cluster, the percentage of genes in the restricted cluster were remarkably increased in most cases. For instance, the percentage for the 'ribonucleoprotein complex' was increased from 14.12% in the raw cluster to 20.36% in the restricted cluster; this indicates that the genes in this category were retained more frequently than the average genes in the cluster. We see from this example that selecting tightly associated genes using the threshold of membership degree can increase the biological relevance of the genes in the cluster.

Table 6. Enrichment of GO categories in the raw cluster C_9 from Table 4 and in the corresponding restricted cluster obtained by the GK method for the yeast PHO-regulation dataset of Ogawa *et al*. (2000)

Cluster	GO category	Raw cluster	Restricted cluster
C_9	Ribonucleoprotein complex	14.12	20.36
	Ribosome	11.96	17.81
	Cytosol	14.31	21.12
	Cytosolic ribosome	8.73	13.74
	Protein biosynthesis	15.20	21.37
	Cytosolic large ribosomal subunit	5.10	8.40
	Large ribosomal subunit	6.27	9.92
	Small ribosomal subunit	4.71	6.87
	Cytosolic small ribosomal subunit	3.63	5.34

For each GO category, the percentage (%) of genes in a cluster is specified.

4 CONCLUSIONS

It is well established that clustering is a useful technique for analysis of large amounts of gene-expression data, and a variety of clustering method have been used in biological research. However, conventional clustering methods suffer from a tendency to impose a fixed shape on the clusters even if the clusters in many real-life datasets may differ in shape. To address this problem, we have presented the GK clustering method using an adaptive distance norm. The adaptive norm is described in terms of the covariance matrix for each cluster in which the eigenstructure is exploited to identify the shape of the cluster. The present assessment has shown that the GK method is the superior and effective method for clustering gene-expression data.

Despite these benefits of the GK method, several issues require further investigation. The computational costs of the GK method are much higher than other clustering methods such as the *k*-means method. In each iteration step, $k \cdot n$ matrices are computed and subsequently inverted, and additionally, k determinants are computed. Especially, the GK method becomes computationally inefficient when applied to high dimensional data. Figure 5 shows the real run time of the GK method to cluster the yeast PHO-regulation and the ATP-regulation datasets. It is observed from the figure that the eight-dimensional PHO dataset shows a rapid increase in the computational time more than the three-dimensional ATP dataset as the number of clusters is increasing. In such a case, it maybe useful to initialize the cluster centroids of the GK method with the resulting centroids of the *k*-means method so that the GK method can converge fast to the

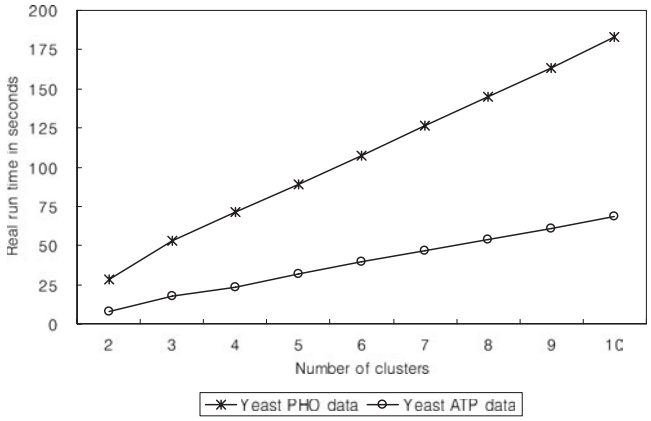

Fig. 5. The run time of the GK method for clustering the yeast PHO-regulation dataset and the ATP-regulation dataset. The horizontal axis represents the number of clusters given, the vertical axis represents the real run time in seconds.

saddle points of J_m with the reduced number of iterations. Second, the GK method is problematic when the number of data is small or when the data within a cluster are much linearly correlated (Babuska, 1998). In such cases, the cluster covariance matrix becomes singular and cannot be inverted; thus we cannot compute the distances at Step 4 in Algorithm 1. To tackle this singularity problem, it would be helpful to prevent the ratio of the maximum to the minimum eigenvalue from being larger than some predefined threshold.

ACKNOWLEDGEMENTS

This work was supported by the Korean Systems Biology Research Grant (M1-0309-02-0002) from the Ministry of Science and Technology. We would like to thank CHUNG Moon Soul Center for BioInformation and BioElectronics and the IBM SUR program for providing research and computing facilities.

REFERENCES

Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Davis,A.P., Dolinski,K., Dwight,S.S., *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Babuska,R. (1998) *Fuzzy Modeling for Control*. Kluwer Academy Publishers, Boston.

Bezdek,J., Keller,J., Krisnapuram,R. and Pal,N.R. (1999) *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer Academy Publishers, Boston.

Cho,R., Campbell,M., Winzeler,E., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhar,D.J. and Davis,R.W., *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Chu,S., DeRish,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

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

DeRisi,J., Iyer,V. and Brown,P. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **282**, 257–264.

Dhilon,I., Marcotte,E. and Roshan,U. (2003) Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, **19**, 1612–1619.

Dudoit,S. and Fridlyand,J. (2003) Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, **19**, 1090–1099.

Dumitrescu,D., Lazzerini,B. and Jain,L. (2000) *Fuzzy Sets and Their Applications to Clustering and Training*. CRC Press, Florida.

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

Gibbons,F. and Roth,F. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.*, **12**, 1574–1581.

Gustafson,E. and Kessel,W. (1979) Fuzzy clustering with a fuzzy covariance matrix. *Proc. IEEE Conf. Decision Control*, 761–766.

Guthke,R., Schmidt-Heck,W. Hahn,D. and Pfaff,M. (2002) Gene expression data mining for functional genomics using fuzzy technology. *Advances in Computational Intelligence and Learning: Methods and Applications*. Springer, New York, pp. 475–488.

Horn,D. and Axel,I. (2003) Novel clustering algorithm for microarray expression data in a truncated svd space. *Bioinformatics*, **19**, 1110–1115.

Issel-Tarver,L., Christie,K., Dolinski,K., Andrada,R., Balakrishnan,R., Ball,C.A., Binkley,G., Dong,S., Dwight,S.S. and Fisk,D.G. (2002) *Saccharomyces*, genome database. *Methods Enzymol.*, **350**, 329–346.

Jain,A., Murty,M. and Flynn,P. (1999) Data clustering: a review. *ACM Comput. Surv.*, **31**, 264–323.

Lukashin,A. and Fuchs,R. (2001) Analysis of temporal gene expression profiles: clustering by simuulated annealing and determining the optimal number of clusters. *Bioinformatics*, **17**, 405–414.

Mizuguchi,G., Shen,X., Landry,J., Wu,W.H., Sen,S. and Wu,C. (2004) ATP-driven exchange of histone h2az variant catalyzed by swr1 chromatin remodeling complex. *Science*, **303**, 343–348.

Ogawa,N., DeRisi,J. and Brown.P.O, (2000) New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by gemomic expression analysis. *Mol. Biol. Cell*, **11**, 4309–4321.

Qin,J., Lewis,D. and Noble,W. (2003) Kernel hierarchical gene clustering from microarray gene expression data. *Bioinformatics*, **19**, 2097–2104.

Selim,S. and Ismail,M. (1984) *K*-means type algorithms: a generalized convergence theorem and the caracterization of local optimality. *IEEE Trans. Pattern Anal. Mach. Intell.*, **6**, 284–288.

Sharan,R., Maron-Katz,A. and Shamir,R. (2003) Click and expander: a system for clustering and visualizing gene expression data. *Bioinformatics*, **19**, 1787–1799.

Steuer,R., Kurths,J., Daub,C., Weise,J. and Selbig,J. (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18**, S231–S240.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky.E., Lander,E.S. and Golub,T.R. (1999) Interpreting patters of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

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

Xu,Y., Olman,V. and Xu,D. (2001) Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, **17**, 309–318.

Yeung,K., Haynor,D. and Ruzzo,W. (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.

Yoshimoto,H., Saltsman,K., Gasch,A., Li,H.X., Ogawa,N., Botstein,D., Brown,P.O. and Cyert,M.S. (2002) Genome-wide analysis of gene expression regulated by the calcineurin/crz1p signaling pathway in *Saccharomyces cerevisiae. J. Biol. Chem.*, **277**, 31079–31088.