BIOINFORMATICS

Vol. 20 no. 10 2004, pages 1535–1545
doi:10.1093/bioinformatics/bth124

Evaluation and optimization of clustering in gene expression data analysis

A. Fazel Famili*, Ganming Liu and Ziying Liu

Institute for Information Technology, National Research Council of Canada, Ottawa, ON, Canada K1A 0R6

Received on September 22, 2003; revised on December 11, 2003; accepted on December 17, 2003
Advance Access publication February 12, 2004

ABSTRACT

Motivation: A measurement of cluster quality is needed to choose potential clusters of genes that contain biologically relevant patterns of gene expression. This is strongly desirable when a large number of gene expression profiles have to be analyzed and proper clusters of genes need to be identified for further analysis, such as the search for meaningful patterns, identification of gene functions or gene response analysis.

Results: We propose a new cluster quality method, called stability, by which unsupervised learning of gene expression data can be performed efficiently. The method takes into account a cluster's stability on partition. We evaluate this method and demonstrate its performance using four independent, real gene expression and three simulated datasets. We demonstrate that our method outperforms other techniques listed in the literature. The method has applications in evaluating clustering validity as well as identifying stable clusters.

Availability: Please contact the first author.

Contact: fazel.famili@nrc-cnrc.gc.ca

INTRODUCTION

A general question facing researchers in many areas, where large amounts of data are produced, is how to organize observed data into meaningful structures and search for useful patterns. Unsupervised learning techniques, such as clustering, have been the most popular method applied to this problem. Over the last five years, with the advances in genomics and microarray technologies and large amounts of microarray data produced, clustering has been applied to identifying groups of genes with meaningful properties. For example, Eisen *et al.* (1998) applied a hierarchical clustering algorithm to identify groups of co-regulated yeast genes. Tamayo *et al.* (1999) used self-organizing maps to identify clusters of genes with similar expression patterns in the yeast cell cycle and human hematopoietic differentiation datasets. In addition, Yeung *et al.* (2001) have considered clustering as an useful technique because of the large number of genes and the complexity of biological networks.

Given a large number of clusters, biologists are faced with the problem of choosing the smallest number of clusters, which potentially contain biologically relevant patterns of gene expressions. Quantitative methods are preferred when assessing whether a cluster of genes is potentially related to a problem or, amongst all clusters, which ones would result in meaningful patterns if more investigation were done. Our paper provides a quantitative, data-driven approach to selecting the most promising clusters of genes that contain biologically relevant information (i.e. meaningful patterns of expression).

There are many publications related to the optimal number of clusters, the optimal clustering algorithm or the optimal similarity/dissimilarity measure for a given gene expression dataset. However, only a few papers address the problem of identifying high-quality clusters that potentially contain biologically relevant patterns. In this paper, we investigate and compare techniques that could be used to assess the cluster quality. We further introduce a new stability-based technique based on clusters' immovability on partition. Immovability of a cluster is the rate at which the contents of a cluster remain unchanged during a clustering process for $K = i$ to $i + n$, where $n \geq 1$, and K is the number of clusters. The advantage of our method over existing methods is that it takes into account all factors that affect the clustering process and at the same time uses the complete dataset to determine the cluster quality (the original information is kept intact). Other methods use only some of the factors that determine quality, such as the silhouette index (Rousseeuw, 1987); or use only part of the complete dataset such as the re-sampling validation method (Dudoit and Fridlyand, 2002; Ben-Hur *et al.*, 2002).

In the following sections, we first describe related work and compare it with techniques used in our studies. We then introduce our stability-based technique and present the results of applying this method to four gene expression datasets from biological applications and three simulated datasets. To show the effectiveness of our technique, we also apply some well-known and efficient cluster validation methods to our datasets, and compare them with our method. The final section of the paper contains the conclusions of our studies and potential future work.

*To whom correspondence should be addressed.

RELATED WORK

Choosing the highest quality clusters of genes from the results of clustering gene expression data is not a well-studied topic. Only a few researchers have addressed this problem. Vilo *et al.* (2000) created a large number of independent clusters of gene expression data and simultaneously assessed the 'goodness' of each cluster by its average object silhouette value (Rousseeuw, 1987). Raychaudhuri and Altman (2003) evaluated a method called neighbor divergence per gene (NDPG), which uses scientific literature to assess whether a group of genes are functionally related. This method needs a corpus of documents and an index connecting the documents to genes. Zhang and Zhou (2000) proposed a parametric bootstrap re-sampling method (PBR) to incorporate information on variations in gene expression levels to assess the reliability of gene clusters identified from large-scale gene expression data. For each re-sampling, a set of 'new' observations is generated by replacing the true observation for each gene under each condition with a random variable sampled using the observed expression level and estimated uncertainty in gene expression measurements. More recently, Smolkin and Ghosh (2003) assessed the stability of a cluster using their Cluster Stability Score, by which a cluster's stability is calculated through clustering on a random subspace of the attribute space.

There are also a number of papers that refer to the cluster validation problem for gene expression data (the optimal number of clusters). The most common cluster validation techniques are based on one of the following three principles: external criteria, internal criteria and replication (Fiske, 1983). In most cases, external information is not known, and so internal criteria and replication techniques are more often used for cluster validation. Azuaje (2002) evaluated the validation of three internal indices, the silhouette index, Dunn's index and the Davies–Bouldin (DB) index, for estimating the optimal number of clusters with two gene expression datasets: leukemia samples and B-cell lymphoma samples. Dudoit and Fridlyand (2002) proposed a re-sampling method called Clest to estimate the number of clusters (K) by repeatedly and randomly dividing the original dataset into two non-overlapping sets. Ben-Hur *et al.* (2002) proposed a stability-based re-sampling method for estimating the number of clusters, where stability is characterized by the distribution of pair-wise similarities between clusters obtained from subsamples of the data. Yeung *et al.* (2001) applied a clustering algorithm to all but one experimental conditions in a dataset. They used the left-out condition to assess the predictive power (Figure-of-Merit, FOM) of the clustering algorithm. The basic idea is to calculate the mean expression level of all the objects (genes) at the left-out condition in one cluster and then calculate the difference between each gene's expression level and the mean expression level. The FOM of this cluster is then the average sum of this difference. More recently, Datta and Datta (2003) formulated three other validation measures using the left-outone condition strategy to evaluate the performances of six clustering algorithms. Lukashin and Fuchs (2001) proposed a clustering algorithm based on a simulated annealing procedure and determined the optimal number of clusters simultaneously with the optimization of the distribution of the genes over clusters. In addition, Giurcăneanu *et al.* (2003) introduced a stability index to estimate the quality of clusters for randomly selected subsets of the data. A decision based on the correct number of clusters was made from the statistics of the index. Lange *et al.* (2002), introduced a model assessment scheme that is based on the notion of stability. The approach results in an upper bound to cross-validation in the supervised learning, with extensions to semi-supervised and unsupervised applications.

METHODS

Among the methods discussed for a cluster's quality in the literature, NDPG needs external information (scientific literature) and PBR requires generating 'new' observations through re-sampling and is time consuming. Cluster Stability Score repeatedly subsamples the attribute space to do clustering. If two subsets of attributes that are randomly sampled from the attribute space, happen to contain independent information, stability of a cluster formed from one subset of attributes is not expected when it is formed from the other subset of attributes. The silhouette index cannot always determine the optimal number of clusters when using genes as objects. Also, the silhouette value cannot identify proper clusters containing informative genes for a disease when using patients as objects (this will be illustrated as part of our evaluation strategy in the Results section).

Among the techniques for clustering validity, the stability-based re-sampling method and FOM could also be used to assess a cluster's quality. The stability-based re-sampling approach generally involves repeatedly re-sampling of the dataset, each time using only a subset of the whole data, We expected the potential for some re-sampled subsets to have a different underlying data structure compared with the original dataset. In addition, this technique has high run-time complexity due to multiple re-samplings. FOM has a limitation that it is not applicable if the experiment conditions from which data are generated contain independent information.

A new method for proper assessment of cluster quality is therefore preferred so that (i) the dataset is kept intact during clustering; (ii) one is able to determine the optimal number of clusters and, in particular, clusters with meaningful patterns of gene expressions and (iii) it can work for various gene expression data including time series, labeled and non-labeled datasets. In this paper, we introduce a novel stability-based technique that assesses the immovability of objects in each cluster when it is partitioned. This method satisfies the above three conditions. We call it cluster's stability on partition, in which no re-sampling is required and is different from the stability indices described in the last section.

Suppose we have a set of clustering results with the number of clusters from 2 to n, which are obtained from the same clustering algorithm. Let $C_{c,l}$ be a set of objects in cluster l resulting from a clustering result with c ($2 \le c \le n$) clusters. Let k ($0 < k \le n - c$) be the threshold (we call it the partition threshold) at which the stability calculation of a cluster will stop. Then the cluster stability of cluster l is

$$S_{c,l} = \min_{i=c+1}^{c+k} \left\{ \max_{j=1}^{i} \left\{ \frac{|C_{c,l} \cap C_{i,j}|}{|C_{c,l}|} \right\} \right\}.$$

$S_{c,l}$ calculates the k values for the maximum number of overlapping objects between the considered cluster l and each of the clusters in a clustering result with the number of clusters $i(c < i \le c+k)$. Then it takes the minimum of the k maximum values as the stability of the considered cluster l. To make the stabilities of different clusters comparable, the stability value is normalized to the range from 0 to 1 by dividing it with the number of objects in l. The closer the stability to 1, the more stable the cluster is.

Let $S_{c,l}$ be the stability of cluster l resulting from a clustering result with c clusters; then the general stability of the entire clustering with c clusters is

$$GS_c = \frac{1}{c} \sum_{l=1}^{c} S_{c,l}.$$

This is the average of the stabilities of all clusters in the clustering result. The optimal number of clusters is a value q at which the general stability is the largest. The larger the cluster stability, the better the cluster quality. The clusters of genes with the best stabilities can be taken as candidates containing reliable patterns, which are valuable for further analysis for biological pattern recognition.

To illustrate the value of our stability-based method, we evaluated it with four gene expression and three simulated datasets and compared it with the result of the silhouette index, which is defined as follows (Rousseeuw, 1987).

The silhouette of a cluster A is measured on its compactness and how far it is from the next closest cluster. Let i be an arbitrary object in A. We define $a(i)$ as the average distance between the i-th object and all the other objects in the same cluster as i.

$$a(i) = \frac{\sum_{j \in A, j \neq i} d(i,j)}{|A| - 1}.$$

For any cluster C other than A, we define

$$d(i,C) = \frac{1}{|C|} \sum_{j \in C} d(i,j)$$

and

$$b(i) = \min_{C \neq A} \{d(i,C)\}.$$

Then object silhouette of object i is

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

in which the range of $S(i)$ is between -1 and 1.

The cluster silhouette is the average of the object silhouette for all objects in cluster A:

$$\text{cluster\_silhouette} = \frac{\sum_{i=1}^{|A|} s(i)}{|A|}.$$

The general silhouette of a clustering result with c clusters is

$$\text{general\_silhouette} = \frac{1}{c} \sum_{i=1}^{c} \text{cluster\_silhouette}_i.$$

The optimal number of clusters is a value q at which the general silhouette is the largest. The larger the cluster silhouette, the better the cluster quality.

DATA USED FOR THIS STUDY

The real data used for this study consisted of four gene expression datasets, each containing gene expression measurements for various numbers of genes that were collected for different problems under study. Two of these datasets are publicly available. We provide references for all these datasets for which more information can be obtained.

Yeast: Consists of 2321 genes as objects with 16 time points as attributes. This data is a subset from the original 6220 genes with 17 time points listed by Cho *et al*. (1998), from which we selected 2321 genes based on the largest variance in their expression. One abnormal time point was also removed from the dataset (suggested by Tamayo *et al*., 1999). These data have been used extensively in the literature for clustering and unsupervised pattern recognition. A large number of genes contained in this dataset have been biologically characterized and assigned to different cell cycle phases.

Leukemia: Approximately 7000 genes as objects, consisting of data for 38 ALL and AML patients as attributes (Golub *et al*., 1999; Famili and Ouyang, 2003). The objective of the original research was to identify the most informative genes for the purpose of disease modeling and more accurate classification of ALL/AML patients. The most informative genes exhibit expression patterns strongly correlated with the class distinction (Golub *et al*., 1999).

Hepatitis C virus: Containing 5756 genes for six repeated experiments (Famili *et al*., 2003) related to Hepatitis C transgenic mice. These data were originally used for gene identification. The expression level of the most informative genes should exhibit a large deviation between experiment and control.

TGF modulated: Consisting of 331 genes (selected from an original list of 15 264 genes) of cells under experimental conditions: Stimulus-transforming growth factor (TGF-β1), p38MAPK inhibitor SB203580 (SB) or both. Each experimental condition was repeated six times. The gene expression level is the ratio of the experimental sample divided by the control sample. These data were generated to isolate and characterize a murine mammary epithelial tumor cell line designated as BRI-JM01. Exposure of this cell line to TGF-β1 resulted in inducing an epithelial-to-mesenchymal transition (EMT) and increased motility, a phenotype critical to tumor progression in cancer (O'Connor-McCourt *et al.*, 2003). The most informative genes exhibit expression patterns that strongly correlated with the experimental conditions (stimulus, inhibitors).

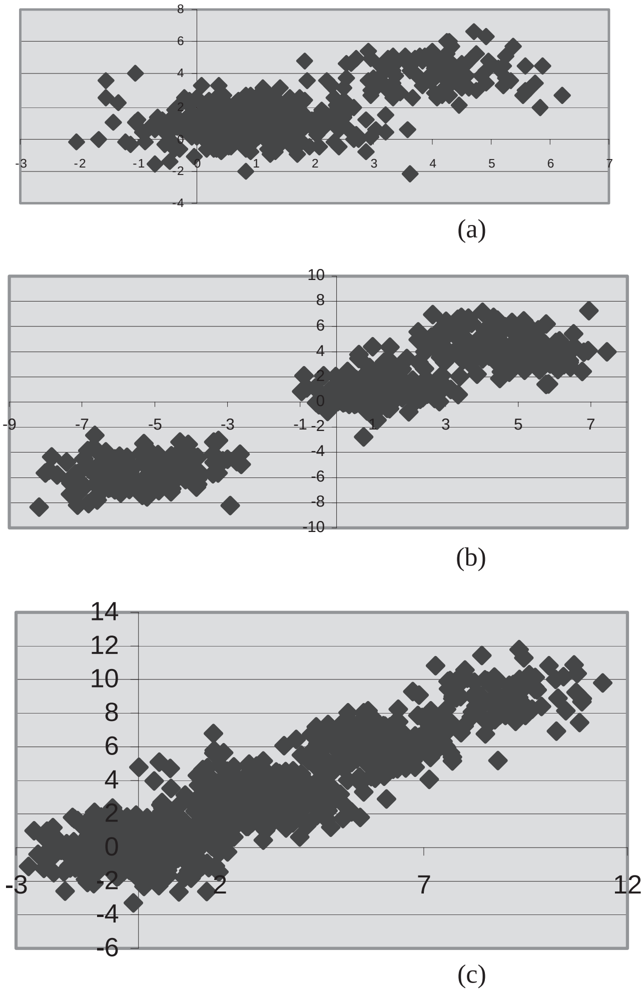
Simulated data: In addition to the above datasets, we generated three simulated datasets, S1, S2 and S3, that contained a bivariate normal distribution, and these were used for this study. Following is a description of these datasets.

S1: Consisting of two overlapped clusters. One contained 300 objects with means (1, 1) and SDs (1, 1). The other contained 100 objects with means (4, 4) and SDs (1, 1). An additional 15% of noise was added to the dataset.

S2: Consisting of three clusters, two of them overlapped. Each has 150 objects. The middle cluster in the plot has means (1, 1) and SDs (1, 1). The other two have means (4, 5), (−5.5, −5.5) and SDs (1.3, 1.3), (1.3, 1.3), respectively. The deviation is so designed that the objects in the middle cluster in the plot are co-expressed better than those of the other two. This dataset can be used to verify the larger stability of a cluster, indicating its objects are co-expressed well. This dataset also contained 10% noise.

S3: Consisting of four overlapped clusters with objects 400, 300, 200 and 100; means (0, 0), (3, 3), (6, 6) and (9, 9); SDs (1, 1), (1, 1), (1, 1) and (1, 1), respectively. An additional 12% noise objects were added to the dataset.

Figure 1(a–c) shows the three graphs from our simulated datasets.



(a)

(b)

(c)

Fig. 1. (**a**–**c**) Simulated datasets.

Table 1. Summary of clustering experiments

Dataset	Range of clusters	Distance measure
Yeast	2–70	Pearson correlation
Leukemia	2–40	Difference-in-shape
Hepatitis	2–50	Difference-in-size
TGF	2–30	Difference-in-shape

RESULTS

To evaluate the performance of our new stability measure, a number of clustering experiments were performed. These all used K-means with a random seed selection and Euclidean as distance measure. All experiments were performed using our BioMiner data mining software (Walker *et al.*, 2004). Table 1 contains the summary of these experiments. The distance measures listed in this table were selected from amongst 21 different distance measures available in this software. They were selected because they resulted in the highest general silhouette values.

Figure 2 shows the experimental and evaluation procedure. After K-means clustering, stability and silhouette values were calculated. The optimal partition was determined by the general stability and silhouette values. The meaningful and non-meaningful clusters were selected based on domain knowledge and were used to validate the effectiveness of stability over silhouette measures.

Optimal partition—Stability versus Silhouette

Figures 3–6 show the general stability and silhouette values versus the number of clusters for all four datasets.

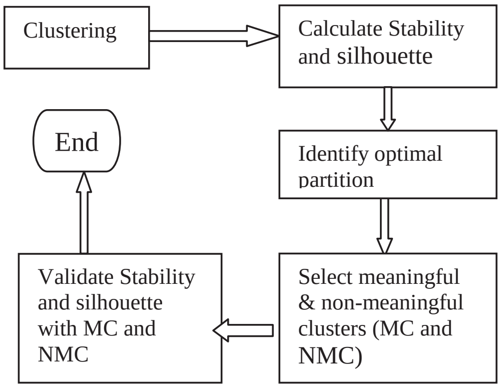

Clustering
Calculate Stability and silhouette
Identify optimal partition
Select meaningful & non-meaningful clusters (MC and NMC)
Validate Stability and silhouette with MC and NMC
End

Fig. 2. Experimental procedure.

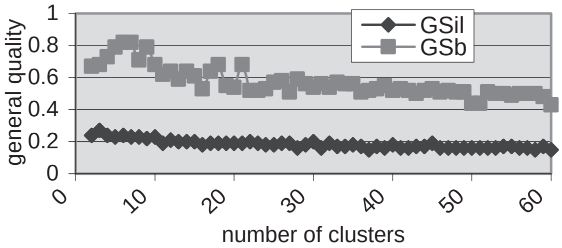


Fig. 3. General stability and silhouette versus the number of clusters for Yeast data.

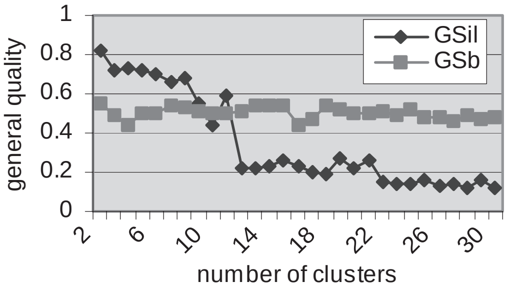


Fig. 4. General stability and silhouette versus the number of clusters for Leukemia data.

These graphs show interesting results that are described below:

(i) In Yeast data (Fig. 3), comparing silhouette versus stability, we noticed that the silhouette values do not indicate any significant changes as the number of clusters increases. Considering the stability values, we checked the partitions with the number of clusters less than 10. In each partition, no clusters could be found exhibiting periodic behaviors that correspond to the five known cell cycle phases: Early G_1, Late G_1, S, G_2 and M. Therefore, we preclude them from the partition with the optimal number of clusters. We noticed that clusters 18 and 21 had the highest general stability (0.68) among clustering results with the number of clusters greater than 10. Therefore, we choose 21 (the average number of genes in each cluster is less than that of 18) as the optimal number of clusters. Figure 7 shows mean expression levels at 16 time points of the 21 clusters corresponding to the clustering results. The gene expression patterns of the 21 clusters are distinctive. This is evidence to support that 21 is the optimal number of clusters identified by general stability.

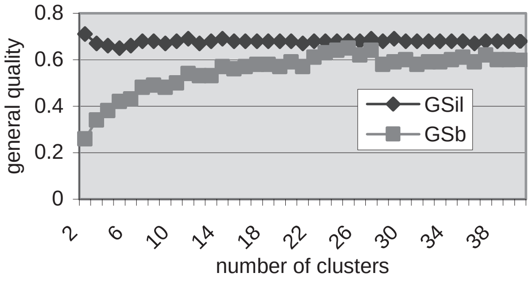


Fig. 5. General stability and silhouette versus the number of clusters for Hepatitis data.

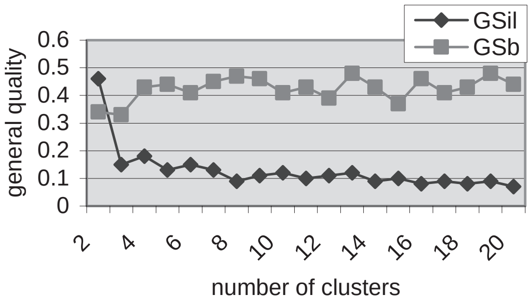


Fig. 6. General stability and silhouette value versus the number of clusters for TGF data.

(ii) In Leukemia data (Fig. 4), the general stability was between 0.4 and 0.6 throughout the clustering experiments compared with silhouette values that dropped substantially after 11 clusters. Using the stability index, clusters 13, 14 and 15 had the highest general stabilities (0.54) among clustering results with the number of clusters from 3 to 30.

(iii) For Hepatitis data (Fig. 5), although the silhouette values were fairly high, they did not change significantly, while the general stability showed an upward trend from the beginning, with the values very close to the silhouette value after 12 clusters. Considering the stability values, the clustering result with 25 clusters had the highest general stability (0.65), and so we choose 25 as the optimal number of clusters.

(iv) In the case of TGF data (Fig. 6), the silhouette values dropped after three clusters and remained very low. However, the stability was much higher, almost

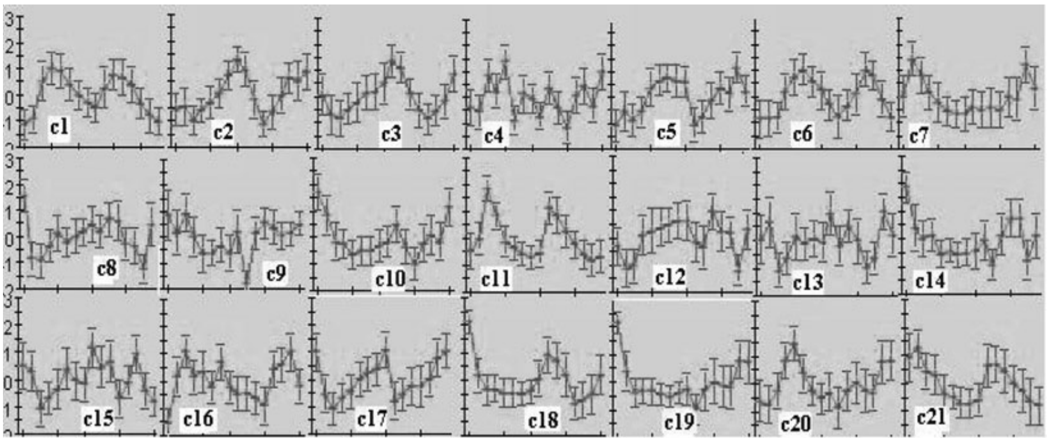


Fig. 7. Mean expression levels at 16 time points for the 21 clusters corresponding to the best clustering results as determined by stability on Yeast data.

from the beginning of the experiments. As with other datasets, the clustering process with number of clusters 13 resulted in the highest general stability (0.48), and so we choose 13 as the optimal number of clusters.

According to the stability and silhouette formula, the dataset reaches its optimal partition at the point where the general stability and general silhouette reach peak values. For all four real datasets tested, the silhouette values reached peaks at very small number of clusters and decreased (in three out of four experiments) with the number of clusters increasing. This is obviously not reasonable when clustering datasets containing large numbers of genes. Therefore, the stability is a more reliable measure for the optimal partition.

Meaningful clusters

The clusters in the best partitions for all four datasets were further evaluated.

(i) *Yeast data*: Figure 7 shows mean expression levels at 16 time points of the 21 clusters corresponding to the clustering results. From this figure, five clusters clearly exhibit periodic behaviors (Table 2) that correspond to cell cycle phases: i.e. C3 corresponds to early G_1 phase, C11 corresponds to late G_1 phase, C1 corresponds to S phase, C6 corresponds to G_2 phase and C2 corresponds to M phase. These are consistent with patterns identified by Cho *et al.* (1998). Figure 8 shows cluster stabilities and silhouettes of the 21 clusters corresponding to the clustering results with the number of clusters at 21. The stabilities of cluster C1, C6 and C11 are 0.45, 0.43 and 0.59, respectively. They are among the clusters with the highest stabilities. The silhouettes of C1, C2 and C11 are 0.2, 0.21 and 0.4, which are among the set of clusters with the highest silhouettes. The silhouettes of clusters C3 and C6 are 0.17 and 0.13, which is not high. The stabilities of C2 and C3 are 0.36 and 0.35, respectively. These are not very high. Here, both stability and silhouette measures correctly identify three clusters among the five clusters with biologically relevant expression patterns.

Table 2. Proportion of biologically characterized genes in meaningful clusters versus those in Cho *et al*. (1998)

Cell cycle	Proportion	Meaningful cluster
Early G_1	16/32	C3
Late G_1	72/87	C11
S phase	23/48	C1
G_2 phase	14/28	C6
M phase	17/30	C2

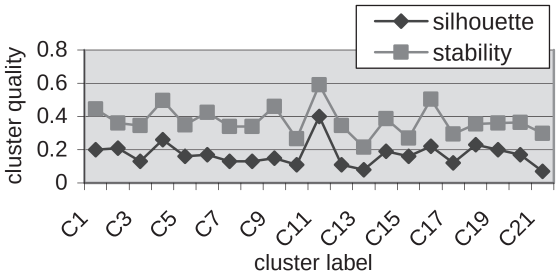


Fig. 8. Cluster stabilities and silhouettes of the 21 clusters formed on Yeast data.

Table 2 shows the proportion of biologically characterized genes listed by Cho *et al*. (1998), contained in our meaningful clusters. There are 225 genes (from the original 415 genes listed by Cho) that passed our variation filter. Among them, 142 genes were found in the five meaningful clusters we identified. This verified that the cell cycle-regulated patterns exist in our meaningful clusters. Figure 9 displays the intensity spectrum plot of the 142 genes.

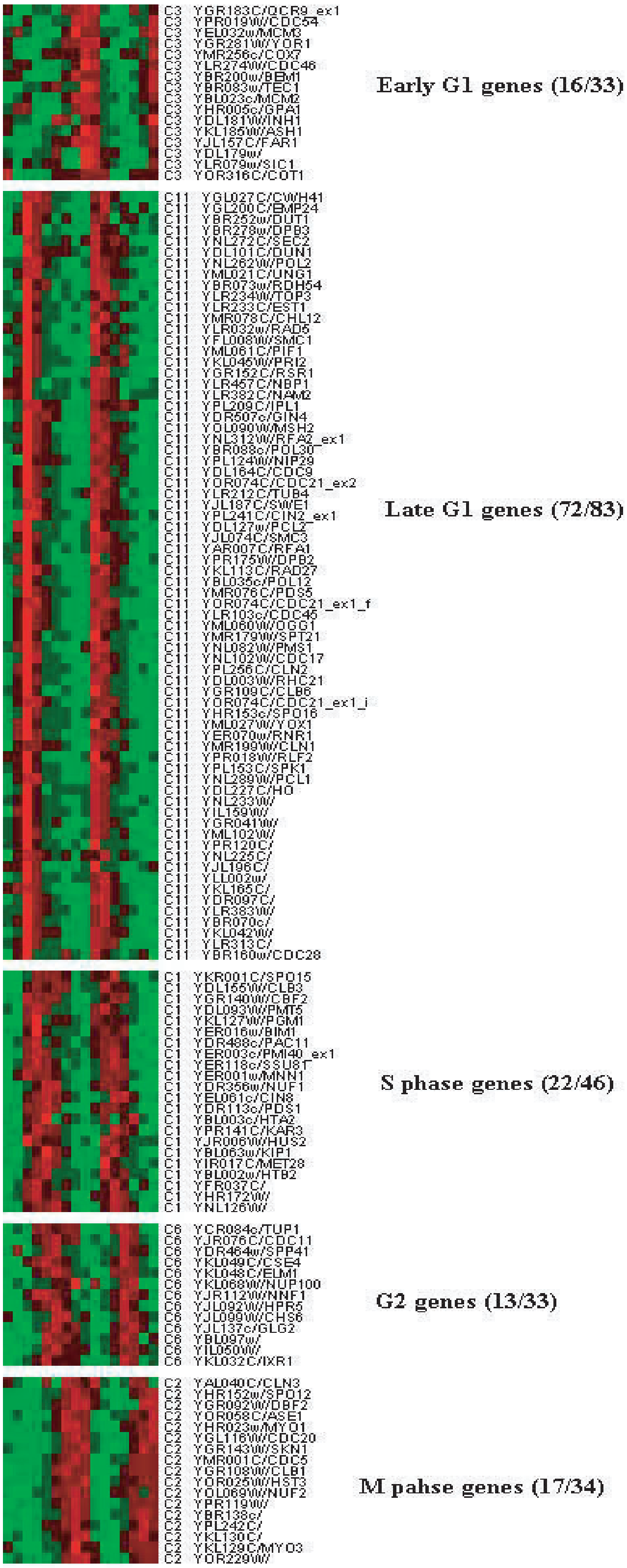


Fig. 9. The intensity spectrum plot of biologically characterized genes (listed by Cho *et al*., 1998) that are found in our meaningful clusters.

(ii) *Leukemia data*: Figure 10 illustrates mean expression levels of the 13 clusters corresponding to the clustering results with the number of clusters at 13. Among them, clusters C3, C5 and C13 exhibit obviously high expression levels (meaningful expression patterns) in AML samples (on the x-axis the last 11 points are AML patients, the others are ALL). Figure 11 shows cluster stabilities and silhouettes of the 13 clusters corresponding to the clustering results as discussed above. The stabilities of clusters C3, C5 and C13 are 0.67, 0.56 and 0.80, respectively. They are in the set of clusters with the highest stabilities. The silhouettes of clusters C3, C5 and C13 are −0.14, −0.09 and −0.02, respectively, which are very poor. Here, compared with the silhouette measure, the stability identified the clusters with gene expression patterns that are class distinctive (ALL and AML). Among the 25 most informative genes highly expressed in AML identified by Golub *et al*. (1999), 14 are found in the three clusters we identified with highly expressed patterns in AML. Figure 12 displays the intensity spectrum plot of the 14 genes. Furthermore among the 25 most informative genes highly expressed in ALL identified by Golub *et al*. (1999), none was included in the three clusters.

(iii) *Hepatitis data*: Figure 13 shows mean expression levels of the 25 clusters created from the hepatitis data. The x-axis represents the six repeated experiments. The y-axis represents the log ratio of experiment divided by control. So the farther the mean log ratio value of a cluster of genes is from zero, the more regulated the cluster of genes is by experimental conditions. Among all the clusters in Figure 13, cluster C1 is the most up-regulated cluster and C15 is the most down-regulated cluster. Figure 14 shows cluster stabilities and silhouettes of the 25 clusters corresponding to the clustering results with the number of clusters set to 25. Clusters C1 and C15 have the highest stability value of 1. The other clusters with high stabilities are C16, C2 and C24, with their stability values being 1.0, 0.98 and 0.84, respectively. Cluster C17 has the lowest stability of 0.42. By comparing these clusters with Figure 13, C16 and C2 are strongly down-regulated; C24 is strongly up-regulated; and C17 does not exhibit any regulated character whose average expression level is just a straight line located near zero. Overall, we notice that while the stability measure works very well for this dataset, the silhouette values do not indicate any additional information.

(iv) *TGF data*: Figure 15 shows the mean expression levels of the 13 clusters from the TGF data. Among them, clusters C1, C6 and C13 exhibit obviously different expression levels (meaningful expression patterns), for the three experimental conditions of TGF-β1, TGF-β1 + SB and SB. Figure 16 shows cluster stabilities and silhouettes for the 13 clusters. The stabilities of clusters C1, C6 and C13 are 1.0, 1.0 and 0.67, respectively. They are in the set of clusters with the highest stabilities. The silhouettes of C1, C6 and C13 are 0.0, 0.0 and 0.29, respectively. Clusters C1 and

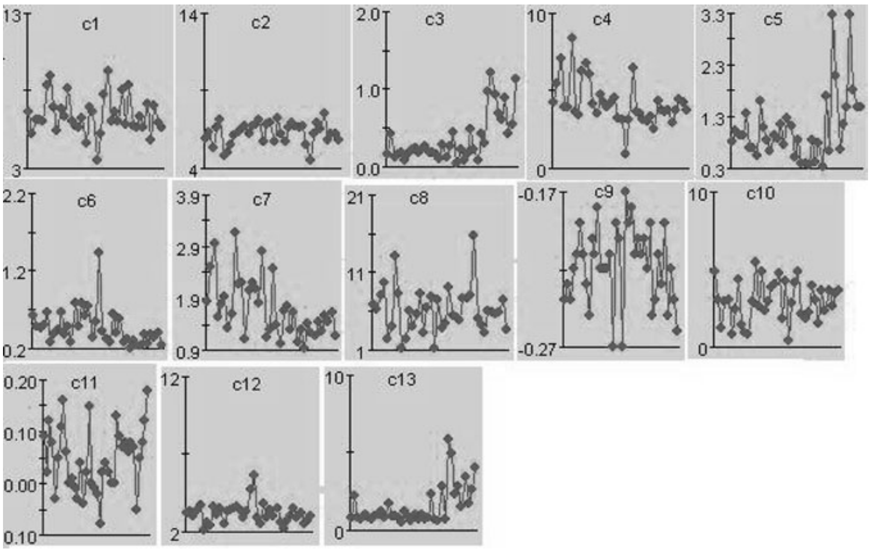


Fig. 10. Mean expression levels of the 13 clusters corresponding to Leukemia data.

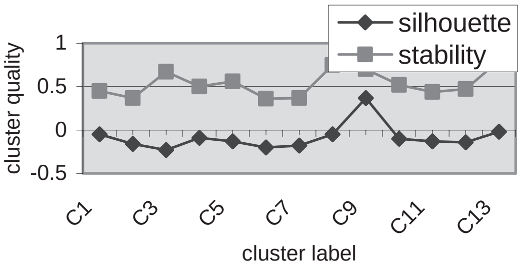


Fig. 11. Cluster stabilities and silhouettes of the 13 clusters from Leukemia data.

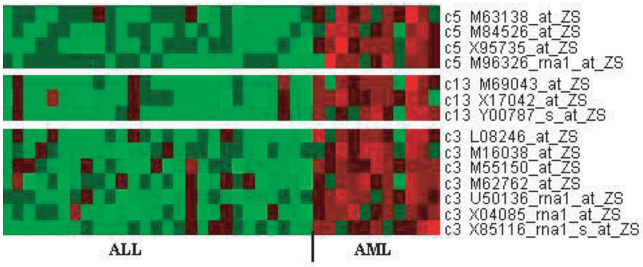


Fig. 12. The intensity spectrum plot of the 14 genes, among the 25 genes highly expressed in AML (Golub *et al*., 1999) that are found in our meaningful clusters.

C6 have only one gene (the silhouette algorithm assigns a value of 0.0 to the cluster), and therefore we exclude them from meaningful clusters. Cluster C13 has the highest silhouette value. For this dataset, both stability and silhouette measures correctly identified the clusters of genes with gene expression patterns that are distinctive with the three different experimental conditions (TGF-β1, TGF-β1 + SB and SB).

Table 3. Summary of experimental results

Dataset	Yeast	Leukemia	Hepatitis	TGF
Optimal clusters	21	13	25	13
Meaningful clusters (MC)	5	2	4	1
Non-meaningful clusters (NMC)	N/A	N/A	1	N/A
MC with high stability	3	2	4	1
MC with high silhouette	3	0	0	1
NMC with low stability	N/A	N/A	1	N/A
NMC with low silhouette	N/A	N/A	0	N/A

N/A, not available.

Table 3 is a summary of the experimental results. It shows that the number of known clusters with biologically relevant patterns that have high stability values, is larger than that of the ones with high silhouette values. The known non-regulated cluster in Hepatitis data has the lowest stability, but its silhouette is not low. As shown in this table, the stability measure outperformed silhouette. The reason is that cluster's stability is measured when it is partitioned in the clustering process, and so it is the result of all factors affecting the clustering.

On the other hand, the silhouette measure considers a cluster as a good cluster if it is compact and separated from other clusters. It would appear that there are other factors that are not taken into account by silhouette, such as the shape of a cluster.

Results from simulated datasets

Table 4 shows the general quality for simulated data. In dataset 1, all four indices correctly identified the correct

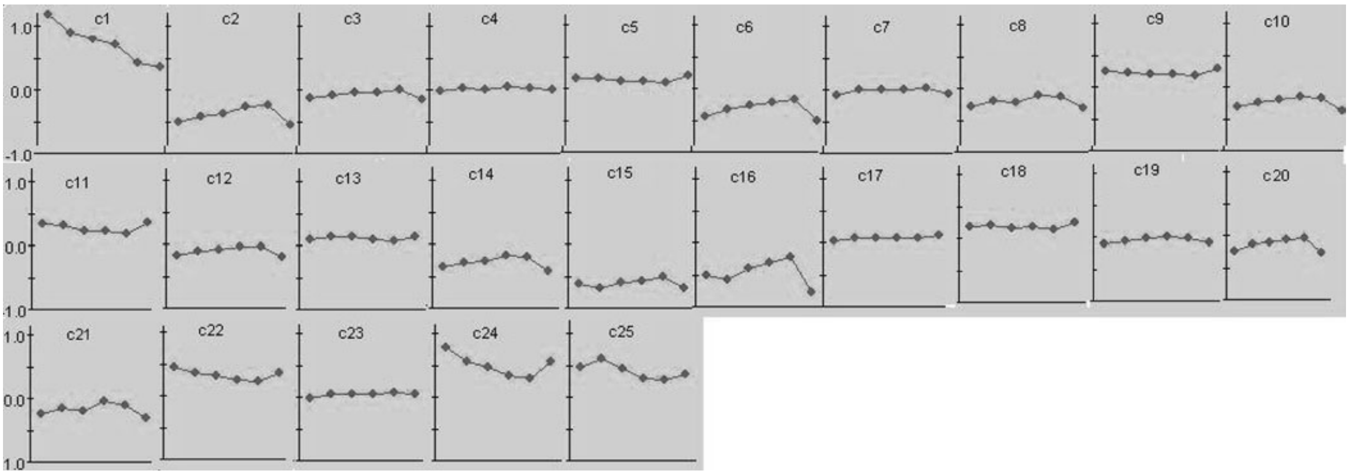


Fig. 13. Mean expression levels of the 25 clusters created from Hepatitis data.

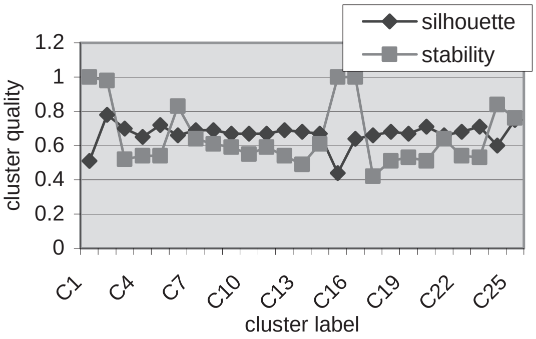


Fig. 14. Cluster stabilities and silhouettes of the 25 clusters created with Hepatitis data.

number of clusters, which was two. In dataset 2, stability index correctly identified the correct number of clusters, which was three. The other three indices wrongly identified it as two. And finally in dataset 3, stability index correctly identified the number of clusters, which was four. The other three indices wrongly identified it as two.

As for cluster quality in simulated data, the middle cluster in the plot (Fig. 1b) had a smaller SD. This dataset was used to verify that a larger stability of a cluster indicated that its objects were co-expressed well. The cluster stability of the middle cluster was 1. The other two had stabilities of 0.95 and 0.71, respectively.

Table 4. General quality of simulated data

Dataset	c	GSb	GSil	Dunn	DB
Data S1	**2**	**0.67**	**0.61**	**2.8**	**−1.85**
	3	0.55	0.39	0.56	−0.56
	4	0.5	0.39	0.2	0.07
	6	0.55	0.32	0.08	0.34
	7	0.52	0.32	0.08	0.07
	8	0.54	0.32	−0.41	0.33
	9	0.5	0.31	−0.49	0.41
	10	0.54	0.33	−0.48	0.28
Data S2	**2**	0.76	**0.73**	**2.94**	**−2.05**
	3	**0.89**	0.66	1.17	−1.4
	4	0.64	0.51	−0.13	−0.19
	5	0.84	0.46	−0.27	−0.01
	6	0.8	0.45	−0.06	0.29
	7	0.65	0.38	−0.27	0.03
	8	0.75	0.32	−0.19	0.38
	9	0.7	0.42	−0.39	0.14
	10	0.7	0.35	−0.5	0.58
Data S3	**2**	0.61	**0.59**	**2.11**	**−1.37**
	3	0.74	0.55	0.44	−0.41
	4	**0.78**	0.55	1.12	−0.86
	5	0.67	0.42	−0.12	0.07
	6	0.67	0.37	−0.19	0.61
	7	0.74	0.39	−0.31	−0.04
	8	0.75	0.39	−0.39	−0.16
	9	0.62	0.37	−0.65	0.34
	10	0.65	0.34	−0.37	0.35

Values in bold indicate the optimal number of clusters.

CONCLUSIONS

Long-standing problems in the analysis of large amounts of microarray data are how to properly cluster the data, how to decide on the correct number of clusters and, more importantly, what are the clusters with the most meaningful information. This is extremely important when one wants to reduce data dimensionality (i.e. dealing with a large number of genes) or choose the right clusters for labeling and pattern recognition. Here, we investigated some of the existing techniques and identified their deficiencies. We then introduced a new, simple and robust method that allows us to quantitatively evaluate any gene expression clustering processes and identify clusters with meaningful patterns. We evaluated the

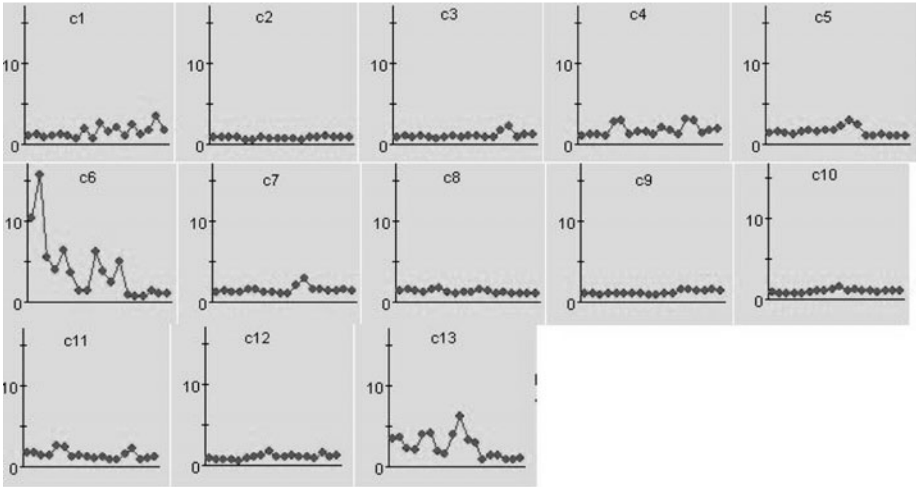


Fig. 15. Mean expression levels of the 13 clusters obtained with TGF data.

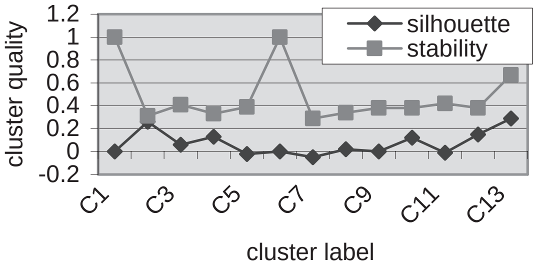


Fig. 16. Cluster stabilities and silhouettes of the 13 clusters from TGF data.

method and its performance using four large gene expression datasets, all collected from real-world applications and three simulated datasets. To summarize:

(1) The new stability on partition measure provided a simple and robust quantitative measure allowing us to identify clusters of genes that contain biologically relevant patterns of gene expression.

(2) It is shown that the stability on partition is a good measure to indicate the optimal number of clusters when genes are treated as objects. In addition to providing useful information about the stability of clusters, the approach solves the problem of cluster validity.

(3) Compared with other techniques, our procedure does not use any external information and does not require subsampling the original dataset.

The new cluster quality evaluation method was tested using the K-means clustering algorithm. As part of our future studies, we plan to use other clustering techniques (such as SOM and Hierarchical clustering) to evaluate our cluster quality index. The new cluster evaluation method allows researchers to perform a meaningful clustering of data, focusing only on genes with the highest information value. This would be a valuable support for gene identification, gene response analysis, disease modeling using microarray data and many other genomics data mining tasks that require a complex data analysis process.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of all members of the BioMine project: Alan Barton, Julio Valdes, Youlian Pan and Junjun Ouyang from IIT, Roy Walker, Qing Yan Liu and Brandon Smith from the IBS at NRC and a number of former students. Special thanks to Marko Kryworuchko, Maureen O'Connor and Anne Lenferink for providing the Hepatitis and TGF data. We would like to thank Bob Orchard for reviewing an earlier version of this paper.

REFERENCES

Azuaje,F. (2002) A cluster validity framework for genome expression data. *Bioinformatics*, **18**, 319–320.

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.

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

Datta,S. and Datta,S. (2003) Comparisons and validation of clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459–466.

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.

Famili,A. and Ouyang,J. (2003) Data mining: understanding data and disease modeling. *Proceedings of the 21st IASTED International Multi-Conference on Applied Informatics (AI 2003)*, Innsbruck, Austria, February 10–13. IASTED/ACTA Press, Anaheim, USA, pp. 32–37.

Famili,A., Ouyang,J., Kryworuchko,M., Alvarez-Maya,I., Smith,B. and Diaz-Mitoma,F. (2003) Knowledge discovery in Hepatitis C virus transgenic mice. Submitted to the *17th International Conference on Industrial and Engineering Applications of Artificial Intelligence*, Ottawa, ON, Canada, May 17–20.

Fiske,D. (1983) *Cluster Analysis for Social Scientists*. Jossey-Bass Publishers, San Francisco, pp. 104–109.

Giurcăneanu,C.D., Tabus,I., Shinulevich,I. and Zhang,W. (2003) Stability-based cluster analysis applied to microarray data. *Proceedings of the ISSPA 2003, EURA SIP-IEEE Seventh International Symposium on Signal Processing and its Applications*, Paris, France, July 1–4.

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

Lange,T., Braun,M., Roth,V. and Buhmann,J. (2002) Stability-based model selection. *Advances in Neural Information Processing Systems* (*NIPS 2002*), in press.

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

O'Connor-McCourt,M., Lenferink,A., Nantel,A., Cantin,C., Magoon,J., Ouyang,J., Liu,G. and Famili,A. (2003) Analysis of transforming growth factor (TGF)-β modulated genes involved in the epithelial to mesenchymal transdifferentiation of murine mammary epithelial cells. *Poster presentation at American Society of Classical Realism*, Washington DC, USA.

Raychaudhuri,S. and Altman,R. (2003) A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, **19**, 396–401.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Smolkin,M. and Ghosh,D. (2003) Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, **4**, 36.

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

Walker,P.R., Smith,B., Liu,Q., Famili,A., Valdes,J., Liu,Z. and Lach,B. (2004) Data mining of gene expression changes in Alzheimer brain. *Artif. Intell. Med.*, in press.

Vilo,J., Brazma,A., Jonassen,I., Robinson,A.J. and Ukkonen,E. (2000) Mining for putative regulatory elements in the yeast genome using gene expression data. *ISMB-2000, August*. pp. 384–394.

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

Zhang,K. and Zhou,H. (2000) Assessing reliability of gene clusters from gene expression data. *J. Func. Integr. Genomics*, **1**, 156–173.