BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 19 2006, pages 2405–2412
doi:10.1093/bioinformatics/btl406

Gene expression

Evaluation and comparison of gene clustering methods in microarray analysis

Anbupalam Thalamuthu[1,†], Indranil Mukhopadhyay[1,†], Xiaojing Zheng[1] and George C. Tseng[1,2,*]

[1]Department of Human Genetics and [2]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA

Received on February 4, 2006; revised on May 23, 2006; accepted on July 21, 2006
Advance Access publication July 31, 2006
Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Microarray technology has been widely applied in biological and clinical studies for simultaneous monitoring of gene expression in thousands of genes. Gene clustering analysis is found useful for discovering groups of correlated genes potentially co-regulated or associated to the disease or conditions under investigation. Many clustering methods including hierarchical clustering, *K*-means, PAM, SOM, mixture model-based clustering and tight clustering have been widely used in the literature. Yet no comprehensive comparative study has been performed to evaluate the effectiveness of these methods.

Results: In this paper, six gene clustering methods are evaluated by simulated data from a hierarchical log-normal model with various degrees of perturbation as well as four real datasets. A weighted Rand index is proposed for measuring similarity of two clustering results with possible scattered genes (i.e. a set of noise genes not being clustered). Performance of the methods in the real data is assessed by a predictive accuracy analysis through verified gene annotations. Our results show that tight clustering and model-based clustering consistently outperform other clustering methods both in simulated and real data while hierarchical clustering and SOM perform among the worst. Our analysis provides deep insight to the complicated gene clustering problem of expression profile and serves as a practical guideline for routine microarray cluster analysis.

Contact: ctseng@pitt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Microarray gene expression data allow us to quantitatively and simultaneously monitor the expression of thousands of genes under different conditions (Brown and Botstein, 1999). Genes with similar expression pattern under various conditions or time course may imply co-regulation or relation in functional pathways. Identification of such groups of genes with similar expression patterns is usually achieved by exploratory techniques such as cluster analysis (a.k.a. unsupervised machine learning). A total of n genes are assigned into K clusters of similar expression patterns given a dissimilarity measure (usually correlation-based or distance-based) between any two genes. Similarly one can cluster samples to look for patients with similar expression signature in order to discover unknown subtypes of a disease (Sorlie *et al*., 2003). A further aggressive approach known as bi-clustering or two-way clustering searches for groups of genes that have similar expression pattern only in a subset of samples or time periods (Cheng and Church, 2000). These analyses usually provide a good initial investigation in most of the microarray data before actually scrutinizing into specific pathways or genetic mechanisms.

In addition to cluster analysis, several advanced methods are developed to discover gene co-regulation or interactions that direct correlation of gene pairs may not be able to capture. Zhou *et al*. (2002) proposed an idea of transitive expression similarity where two genes with low direct correlation but with high correlation along a transitive pathway can be identified. Li (2002) proposed another idea of liquid association to capture pairs of genes with low direct correlation but their correlation becomes high when conditional on certain cellular state or the expression of a third gene. Although these methods help to capture many potential co-regulation information that gene clustering may not reveal, gene clustering remains an important and essential component in almost every microarray data analysis project. It usually serves as a first-step categorization and exploration of the genes in the data. Visualization of the clusters also provides initial validation of the data quality and helps to elucidate the inter-correlation structure and expression changes across the samples.

Many gene clustering methods have been proposed and applied in the literature. Hierarchical clustering (Eisen *et al*., 1998), *K*-means (MacQueen 1967; Hartigan and Wong, 1979), partitioning around medoids (PAM; a.k.a. *K*-memoids) (Kaufman and Rousseeuw, 1990), self-organizing maps (SOM) (Kohonen, 1990; Tamayo *et al*., 1999) are traditional algorithms and are among the most popular ones in microarray analysis. Recently some methods have been proposed to allow a noise set of genes (or so-called scattered genes) without being clustered. This is in view of the fact that very often a significant number of genes in an expression profile do not play any role in the disease or perturbed conditions under investigation. Forcing all these genes into cluster formation can introduce more false positives and distort the structure of identified clusters. Model-based clustering (Yeung *et al*., 2001a; Fraley and Raftery, 2002b; Medvedovic and Sivaganesan, 2002; McLachlan *et al*., 2002; Medvedovic *et al*., 2004) and tight clustering (Tseng and Wong, 2005) are two examples among this category. The former one is based on a mixture Gaussian

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

model with a component of homogeneous Possion process for scattered genes (Fraley and Raftery, 2002b) and rigorous statistical inference tools including parameter estimation and model selection are available for the method. The later one utilizes a repeated resampling approach to provide better robustness and directly searches for tight clusters. Descriptions of various gene clustering methods and their pros and cons are discussed in Supplementary Material.

A common practical issue in the problem of gene clustering of microarray data is the choice from many available methods and the choice of the corresponding parameters in the methods. It is well-known that varying parameters such as the number of clusters or even varying the random seed in the optimization routines in some clustering methods can produce very different results. Several resampling techniques have been applied to validate the clustering robustness and to provide more stable clustering (Tibshirani *et al*., 2001; Dudoit and Fridlyand, 2002; Smolkin and Ghosh, 2003; Monti *et al*., 2003; Tseng and Wong, 2005). Wu *et al*. (2002) have proposed a pooled analysis of thousands of clusters accumulated from various clustering methods and different parameter settings applied on the same data. By combining information of multiple clustering results, such pooled analysis is shown to provide better annotation prediction. It, however, still remains unclear what methods should be chosen to generate clusters for a better pooled analysis. In general, there is a need to understand the effectiveness of different clustering methods applied in gene clustering of microarray data.

To the authors' knowledge, none has performed such comprehensive evaluation. Yeung *et al*. (2001b) proposed a jackknife approach for within-system validation and comparison. Recently Handl *et al*. (2005) gave a general review of issues in clustering validation. In this paper, we have compared six widely used gene clustering methods by simulated data and real data. A weighted Rand index for comparing two clustering results with possible scattered genes was developed for performance evaluation in simulated data. In the real datasets we applied a predictive accuracy analysis without estimating the number of clusters (similar to Wu *et al*., 2002) to compare different methods. The result shows that methods allowing scattered genes seem to provide better accuracy and robustness in gene clustering while popular methods with visualization advantages, such as hierarchical clustering and SOM, have worse performance and should be used with caution in practice. The study not only demonstrates effectiveness of the methods but also provides deeper insights to the nature of the algorithms and their feasibility to gene clustering of microarray data.

2 METHODS

Suppose X denotes the microarray data matrix with n genes on the rows and d samples on the columns. In the following discussion we assume that the data matrix X is pre-processed and normalized (Tseng *et al*., 2001; Yang *et al*., 2002). Each gene vector is standardized to have mean 0 and SD 1 so that Pearson correlation and Euclidean distance of any two gene vectors are essentially equivalent (Tamayo *et al*., 1999). Cluster analysis aims at grouping these n genes into K clusters such that genes in the same cluster have similar expression patterns. Formally let $x_1, x_2, \ldots, x_n$ denote n gene vectors, each of dimension d; the problem is to assign these d-dimensional vectors into K disjoint subsets $C_1, C_2, \ldots, C_k$ of sizes $n_1, n_2, \ldots, n_K$ respectively, such that $\sum_{i=1}^{K} n_i = n$. We denote a clustering result as a partition $P_K(X,C)$, which is characterized by the data matrix X, the number of clusters K and $C = (C_1, C_2, \ldots, C_k)$. Most of the commonly used clustering algorithms require the number of clusters K to be known a priori. The problem of estimating K will be illustrated in a later subsection. A brief discussion of the clustering methods including hierarchical clustering, K-means, PAM, SOM, model-based clustering and tight clustering is given in Supplementary Material.

Implementation of clustering methods

We have used 'hclust' and 'kmeans' functions in 'stats' library of R (R Development Core Team, 2004) for Hierarchical clustering and K-means respectively. Libraries 'cluster', 'som' and 'mclust' (Fraley and Raftery, 2002a) were used for PAM, SOM and model-based clustering respectively. Tight clustering routine was obtained from the original author's website: http://www.pitt.edu/~ctseng/research/tightClust_download.html. We discuss more details of computational issues and implementation of 'mclust' in the Supplementary Material. In general application of 'mclust' in real datasets needs special care to avoid singularity or naïve local minimum results.

Estimation of number of clusters

Many methods for estimating the number of clusters have been proposed in the literature. Milligan and Cooper (1985) performed comprehensive comparison of >30 methods. In general a method may perform better than another method in a particular probability distribution setting but becomes worse in another data setting. Dudoit and Fridlyand (2002) introduced a resampling-based method known as CLEST to estimate the number of clusters and applied it in microarray analysis. However, estimating K is usually found very difficult, if not impossible, in a real microarray data especially when performing gene clustering. This fact also agrees to the biological intuition that the underlying genetic interactions in an organism are so complex that the definition of gene clusters and the exact number of clusters K are vague. In this paper, the true K ($K = 15$) is supplied to the algorithms in the simulated data. For the real data, we try to avoid such a task by running clustering with varying K ($K = 5$–30) and pooling the results to assess the effectiveness of different methods.

External indices for comparison of two partitions

One of the evaluation criteria for gene clustering methods is based on their ability to reconstruct the true underlying cluster structure, if known. In simulation studies the true underlying cluster structure or the partition of the data is known. The performance of a clustering method can be evaluated by the similarity of its resulting partition and the true partition. Several external indices are available in the literature (Dudoit and Fridlyand, 2002) for this purpose. Here we describe a popular similarity measure of two partitions known as Rand index (Rand, 1971; Hubert and Arabie, 1985) and extend the measure to accommodate situations where a set of scattered genes may exist without being clustered.

Consider two partitions $P_R(X,C_R)$ and $P_C(X,C_C)$ with the group labels $C_R = \{u_1, u_2, \ldots, u_R, u_{R+1}\}$ and $C_C = \{v_1, v_2, \ldots, v_C, v_{C+1}\}$ where u_{R+1} and v_{C+1} are the scattered gene sets in respective partitions. Suppose $P_R(X,C_R)$ is the underlying true clustering structure and we want to evaluate the performance of $P_C(X,C_C)$. The cross tabulation of the two partition can be represented as in the contingency table (Table 1). The entry n_{ij} denotes the number of genes belonging to the i-th group u_i in partition $P_R(X,C_R)$ and j-th group v_j in partition $P_C(X,C_C)$. The original Rand index (Rand, 1971) was proposed for the situation when no scattered gene sets exist in both partitions (i.e. $u_{R+1} = v_{C+1} = \phi$ and $n_{(R+1)\bullet} = n_{\bullet(C+1)} = 0$). A pair of genes is called concordant if they are in the same cluster in both partitions or if they are not in the same cluster in both partitions. Rand index is then defined as the proportion of concordant gene pairs in two partitions among all possible gene pairs. It is easily seen that, Rand index can be simplified to the following

Table 1. Cross tabulation of two different partitions

v_1		v_2	...	v_C	v_{C+1}	Total
u_1	n_{11}	n_{12}	...	n_{1C}	$n_{1(C+1)}$	$n_{1\bullet}$
u_2	n_{21}	n_{22}	...	n_{2C}	$n_{2(C+1)}$	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
u_R	n_{R1}	n_{R2}	...	n_{RC}	$n_{R(C+1)}$	$n_{R\bullet}$
u_{R+1}	$n_{(R+1)1}$	$n_{(R+1)2}$	...	$n_{(R+1)C}$	$n_{(R+1)(C+1)}$	$n_{(R+1)\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet C}$	$n_{\bullet(C+1)}$	$n_{\bullet\bullet}=n$

formula:

$$r(R,C)=1+\frac{\left[\sum_{i=1}^{R}\sum_{j=1}^{C}n_{ij}^2-0.5\left(\sum_{i=1}^{R}n_{i\bullet}^2+\sum_{j=1}^{C}n_{\bullet j}^2\right)\right]}{\binom{n}{2}}.$$

It is usually preferred that such an external index is standardized to have expected value zero when the partitions are randomly generated and takes maximum value one if two partitions are perfectly identical. This results in the modified Rand index (Hubert and Arabie, 1985):

$$\mathrm{Rand}(R,C)=\frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2}-\sum_{i=1}^{R}\binom{n_{i\bullet}}{2}\sum_{j=1}^{C}\binom{n_{\bullet j}}{2}\Big/\binom{n}{2}}{0.5\left[\sum_{i=1}^{R}\binom{n_{i\bullet}}{2}+\sum_{j=1}^{C}\binom{n_{\bullet j}}{2}\right]-\sum_{i=1}^{R}\binom{n_{i\bullet}}{2}\sum_{j=1}^{C}\binom{n_{\bullet j}}{2}\Big/\binom{n}{2}} \quad (1)$$

When at least one of the partitions generate scattered gene sets ($u_{R+1}\neq\phi$ or $v_{C+1}\neq\phi$), definition of modified Rand index has to be extended. A possible extension is to consider both scattered gene sets as regular clusters in the two partitions and define

$$\mathrm{Rand}_1^*(R,C)=\frac{\sum_{i=1}^{R+1}\sum_{j=1}^{C+1}\binom{n_{ij}}{2}-\sum_{i=1}^{R+1}\binom{n_{i\bullet}}{2}\sum_{j=1}^{C+1}\binom{n_{\bullet j}}{2}\Big/\binom{n}{2}}{0.5\left[\sum_{i=1}^{R+1}\binom{n_{i\bullet}}{2}+\sum_{j=1}^{C+1}\binom{n_{\bullet j}}{2}\right]-\sum_{i=1}^{R+1}\binom{n_{i\bullet}}{2}\sum_{j=1}^{C+1}\binom{n_{\bullet j}}{2}\Big/\binom{n}{2}}.$$

This index, however, treats scattered genes with equal importance as the clustered genes in concordance evaluation and results in bias against methods without scattered genes especially when $n_{(R+1)\bullet}$ is large.

Another simple alternative is to ignore all scattered genes in either partitions (i.e. considering the Supplementary Table 1) and define the new modified Rand index only based on intersection of clustered genes of the two partitions:

$$\mathrm{Rand}_2^*(R,C)=\frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2}-\sum_{i=1}^{R}\binom{\tilde{n}_{i\bullet}}{2}\sum_{j=1}^{C}\binom{\tilde{n}_{\bullet j}}{2}\Big/\binom{\tilde{n}}{2}}{0.5\left[\sum_{i=1}^{R}\binom{\tilde{n}_{i\bullet}}{2}+\sum_{j=1}^{C}\binom{\tilde{n}_{\bullet j}}{2}\right]-\sum_{i=1}^{R}\binom{\tilde{n}_{i\bullet}}{2}\sum_{j=1}^{C}\binom{\tilde{n}_{\bullet j}}{2}\Big/\binom{\tilde{n}}{2}},$$

where $\tilde{n}_{i\bullet}=n_{i\bullet}-n_{i(C+1)}, \tilde{n}_{\bullet j}=n_{\bullet j}-n_{(R+1)j}$ and $\tilde{n}=n-\sum_{\{i=R+1 \text{ or } j=c+1\}}n_{ij}$. This index is, however, biased against methods with a scattered gene set. For example, suppose $n_{(R+1)\bullet}$ is large and $\{v_1, v_2,\ldots v_C, v_{C+1}\}$ is a clustering result that contains empty scattered gene set $[n_{\bullet(C+1)}=0]$. If we have an extreme case that $R=C$ and $u_i\subseteq v_i (i=1,\ldots,R)$, then $\mathrm{Rand}_2^*(R,C)$ always equals 1 despite the fact that $\{v_1, v_2,\ldots v_C\}$ contains many scattered genes (false positives) in the clusters.

Here we propose a weighted Rand index using weighted average of the two measures. The weighted Rand index given in Equation (2) will be applied in the analysis hereafter:

$$\mathrm{Rand}^*(R,C)=\lambda\cdot\mathrm{Rand}_1^*(R,C)+(1-\lambda)\cdot\mathrm{Rand}_2^*(R,C) \quad (2)$$

where $\lambda=|u_{R+1}\cup v_{C+1}|/n=(n_{(R+1)\bullet}+n_{\bullet(C+1)}-n_{(R+1)(C+1)})/n$ and $|\cdot|$ denotes the number of genes in the gene set. Note that both $\mathrm{Rand}_1^*(R,C)$ and $\mathrm{Rand}_2^*(R,C)$ and thus $\mathrm{Rand}^*(R,C)$ take maximum value 1 when $P_R(X,C_R)$ and $P_C(X,C_C)$ are perfectly identical and have expected value 0 when $P_C(X,C_C)$ is a random partition. When $\mu_{R+1}=v_{C+1}=\phi$, $\mathrm{Rand}^*(R,C)$, $\mathrm{Rand}_1^*(R,C)$, and $\mathrm{Rand}_2^*(R,C)$ all reduce to the original modified Rand index, $\mathrm{Rand}(R,C)$ in Equation (1).

Annotation prediction by cluster analysis

Annotation prediction of novel genes is one of the initial and useful applications for gene clustering results. Intuitively if an unexpectedly large number of genes in a cluster belong to a specific functional category '*F*', then genes in this cluster are more likely to be relevant to function '*F*'. Suppose a total of *G* genes in the genome are analyzed in the microarray experiment among which *m* genes are known to belong to a particular functional category '*F*'. Within a cluster of size *D* genes, *h* genes belong to the functional category '*F*'. Under the null hypothesis that annotated genes are randomly distributed in clusters, *h* follows a hypergeometric distribution (Tavazoie *et al*., 1999). The *P*-value (i.e. the probability of observing *h* or more annotated genes in the cluster) is calculated as

$$P[X\geq h]=1-\sum_{i=0}^{h-1}\binom{D}{i}\binom{G-D}{m-i}\Big/\binom{G}{m}$$

Intuitively unexpected large *h* will result in small *P*-value indicating that majority of the genes in the cluster might belong to the functional category '*F*'. Given a pre-defined threshold δ which is determined after multiple comparison correction, all genes in the cluster are assigned (predicted) to '*F*' if its *P*-value is less than δ. It is noted that a cluster can be annotated to more than one functional category by this procedure.

Evaluation of gene clustering by functional prediction accuracy

Since the underlying true partition is unknown in real microarray data, the weighted Rand index cannot be used to evaluate different gene clustering methods. Instead the validated annotation from biological databases can be used. Another complexity commonly encountered while comparing clustering methods is the choice of the number of clusters *K*. Wu *et al*. (2002) proposed a method for functional annotation prediction of clusters by pooling results from several clustering algorithms and various *K*. To avoid the sensitivity of estimation of *K*, here we propose a similar pooling criterion and a plot of predictive accuracy for comparing clustering methods.

Suppose, based on a biological database, a number of genes in the microarray data are annotated as belonging to *M* distinct functional categories and the remaining genes are 'unannotated'. For example, six disjoint functional categories containing 104 genes are presented in Supplementary Table 2 (Spellman *et al*., 1998). Consider a *K*-cluster solution from a particular clustering algorithm. The clustering result and the functional categories of the 104 genes can be cross tabulated as in Table 2 ($K=5$). Here some genes belong to a cluster of 'unannotated' category (F_7) and the clustering method may or may not group some genes as 'scattered' genes (v_{noise}). For a given *K*-cluster solution and a specified δ, let n_{ij} ($i=1,2,\ldots,K$ and $j=1,2,\ldots,M$; 'noise' genes and 'unspecified' functional category are not considered) denote the number of genes in cluster *i* and functional category *j* and p_{ij}, the corresponding *P*-values obtained from the null hypergeometric distribution. We assign all the genes in cluster *i* to functional category *j*, if the corresponding p_{ij} values are below the threshold level δ. Thus for the cluster *i*, with total number of genes $n_{i\bullet}$, define, $vp_{Ki}=\sum_{\{j:p_{ij}<\delta\}}n_{ij}$ as the 'verified predictions'. Therefore for the entire *K*-cluster solution, the total number of

Table 2. Cross tabulation of clusters and functional annotation

	F_1	F_2	F_3	F_4	F_5	F_6	F_7
v_1	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	137
v_2	0(1)	0(1)	0(1)	0(1)	0(1)	0(1)	143
v_3	2(0.52)	6(0.0005)	23(0)	0(1)	0(1)	0(1)	136
v_4	0(1)	4(0.013)	5(0.074)	8(1.7E-9)	5(6.5E-5)	0(1)	115
v_5	0(1)	1(0.34291)	0(1)	0(1)	0(1)	0(1)	56
v_{noise}	15	1	0	0	2	15	989

Crosss tabulation of cluster assignment and functional annotation. The corresponding *P*-values are shown in parentheses.

'verified predictions' is given by $VP_K(\delta) = \sum_{i=1}^{K} vp_{Ki} = \sum_{i=1}^{K} \sum_{\{j:p_{ij}<\delta\}} n_{ij}$. The total number of 'predictions made' is given by $PM_K(\delta) = \sum_{i=1}^{K} \sum_{\{j:p_{ij}<\delta\}} n_{ij}$. We define the accuracy of a clustering method to predict functional annotation of genes as $A_K(\delta)=VP_K(\delta)/PM_K(\delta)$. Since the number of clusters K is not known, the accuracy of a clustering method is estimated by pooling solutions from multiple K; the overall accuracy for a given δ is defined as $A(\delta) = VP(\delta)/PM(\delta) = \sum_K VP_K(\delta)/\sum_K PM_K(\delta)$. By varying the P-value threshold δ, we obtain a curve of the total number of 'predictions made' ($PM(\delta)$) versus accuracy ($A(\delta)$). In general, a smaller δ results in fewer numbers of 'predictions made' but higher 'accuracy' in a given clustering method. Clustering methods that generate curves with higher accuracy (i.e. above the other curves) have better performance (Fig. 3). We especially note that success of this comparative tool greatly depends on the quality of annotations from independent resources and whether the sample perturbations in the experiment induce distinct expression patterns for each investigated functional category.

Stochastic model for simulated data and perturbation

The general dependence structure of a microarray data is very complex (Klebanov *et al*., 2006). We have simulated large datasets and perturbed them in such a way to resemble real microarray gene expression data for evaluation purpose. Totally 15 clusters of genes $C^* = (C_1, \ldots, C_{15})$ with dimension $d = 50$ samples are simulated. The cluster size $n_c (c = 1, \ldots, 15)$ is generated from $n_c \sim 4 \times Poisson(\lambda)$. Analyses of many real datasets have reported that the empirical distribution of expression levels is approximately log-normal or sometimes with a slightly heavier tailed t-distribution depending on the biological samples under investigation (Li, 2002). In this paper we use a hierarchical log-normal model for simulation of expression values in a cluster Cc ($c = 1, \ldots, 15$) as the following:

(1) Periods of constant expression: A vector of cluster template for cluster Cc is created with four periods of constant expression of size $m1$, $m2$, $m3$ and $m4$. The sizes mk ($k = 1, \ldots, 4$) is from a uniform distribution such that $\Sigma mk = d$ and $mk > 2$. An initial template with constant pattern in four periods is simulated from $\log(\mu_k^{(c)}) \sim N(\mu, \sigma^2)$ (the initial template in Fig. 1A).

(2) Sample variability and gene variability: Sample variability (σ_s^2) is introduced and the cluster template $T_j^{(c)}$ $(j = 1, \ldots 15)$ is generated from $\log(T_j^{(c)}) \sim N(\log(\mu_k^{(c)}), \sigma_s^2)$, where j is such that $m_1 + \ldots + m_{k-1} < j \leq m_1 + \ldots + m_k (m_0 = 0)$ (the cluster template in Fig. 1A). Then for each gene vector $i (1 \leq i \leq n_c)$ in sample j, the gene variability is added and expression values are generated as $\log(x_{ij}) \sim N(\log(T_j^{(c)}), \sigma_0^2)$ (the cluster of genes in Fig. 1A).

(3) Repeat steps 1 and 2 to simulate each cluster of genes Cc ($c = 1, \ldots, 15$). In this paper $\mu = 6$, $\sigma = 1$, $\sigma_S = 1.0$, $\sigma_0 = 0.1$, and $\lambda = 10$ are used.

In addition to genes with cluster patterns, a number (0, 5, 10, 20, 60, 100 and 200% of the original total number of clustered genes) of randomly simulated scattered genes are added. For sample j ($j = 1, \ldots, 50$) in a scattered gene, the expression level is randomly sampled from the empirical distribution of expressions of all clustered genes in sample j. This is our first perturbation model of simulated data. This contains seven datasets, including the base dataset without any scattered genes. These datasets are called Type I perturbed data for later reference.

A total of 25 different parameter settings (percentage of scattered genes and standard deviation of random Gaussian errors) in three types applied to generate the simulated data.

Another perturbation model is introduced to evaluate robustness of clustering methods against potential random errors introduced from experimental procedures including sample acquisition, labeling hybridization and scanning. For each element of the log-transformed expression matrix, a small random error from normal distribution (SD = 0.05, 0.1, 0.2, 0.4, 0.8, 1.2) is added. These are called Type II perturbed datasets in our analysis, which has six datasets. Type III perturbation model is the one in which scattered genes are added to Type II perturbed data. It is a combination of Type I and Type II perturbation. In Type III we add 100% and 200% scattered genes to the datasets in Type II perturbation. Therefore there are 12 datasets in Type III. Combining all the three perturbation models, there are 25 datasets from the original clustered gene expression matrix. A tabular representation of the schemes for simulated datasets is given in Table 3. All these datasets are replicated 100 times which makes the total number of simulated datasets for this study to 2500. Heatmaps of a simulated example

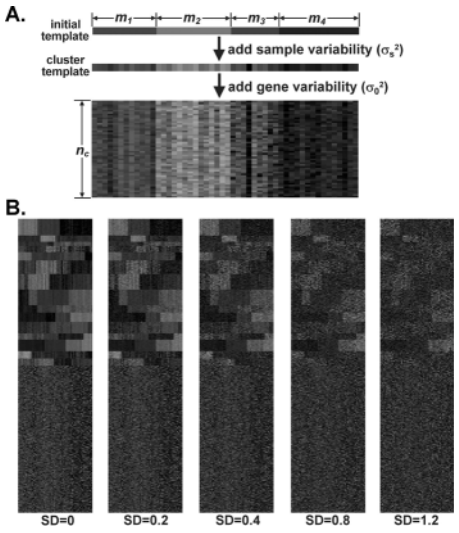

Fig. 1. (**A**) Diagram of cluster simulation. (**B**) Heatmaps of simulated data with increasing perturbations of experimental variability.

Table 3. Schematic representation of simulated data

Scattered genes (%)	SD for random normal error added to each gene 0	0.05	0.1	0.2	0.4	0.8	1.2
0	√ (I)	√ (II)	√ (II)	√ (II)	√ (II)	√ (II)	√ (II)
5	√ (I)						
10	√ (I)						
20	√ (I)						
60	√ (I)						
100	√ (I)	√ (III)	√ (III)	√ (III)	√ (III)	√ (III)	√ (III)
200	√ (I)	√ (III)	√ (III)	√ (III)	√ (III)	√ (III)	√ (III)

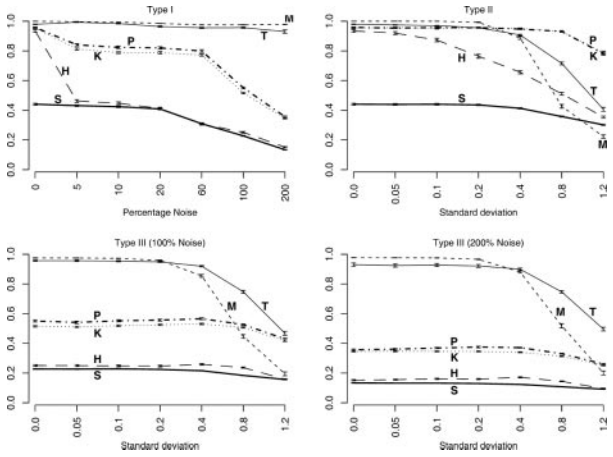

Fig. 2. Weighted Rand index (on *y*-axis) for SOM, Hierarchical, *K*-means, PAM, Mclust and Tight clustering results of simulated data. The means and standard errors are presented.

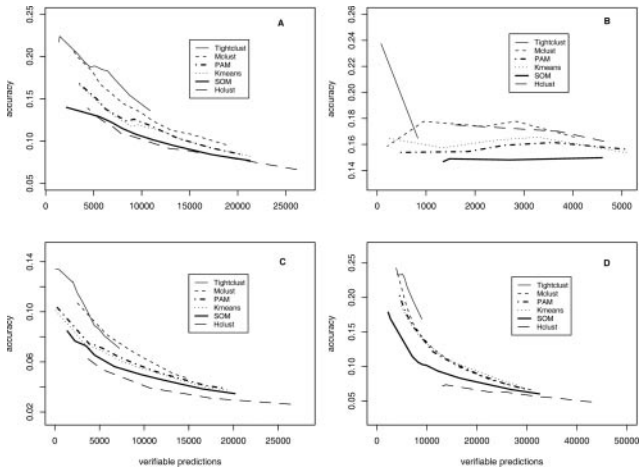

Fig. 3. Prediction accuracy analysis plots in (**A**) yeast cell cycle; (**B**) yeast environmental changes; (**C**) human cell cycle; (**D**) human lung cancer.

with perturbations (scattered genes = 100% and SD = 0, 0.2, 0.4, 0.8 and 1.2) are shown in Figure 1.

3 RESULTS

We have studied six commonly used clustering methods including hierarchical clustering, *K*-means, PAM, SOM, model-based clustering and tight clustering for simulated datasets as well as real microarray gene expression data. All the datasets are properly pre-processed (filtering, missing value imputation and normalization). Each gene vector is standardized to have mean 0 and SD 1 before gene clustering. The standardization makes Pearson correlation and Euclidean distance equivalent. Performance of the methods in simulated data is validated by our proposed weighted Rand index. In the real data sets, the predictive accuracy plots of annotation prediction are used for evaluation.

3.1 Simulated data

Performance of the clustering methods based on the simulated data is presented in this section. Supplementary Figure 1 shows heatmaps of clustering results of an example data set. The proposed weighted Rand index given in Equation (2) is used for evaluation. Mean and standard error of the index values over 100 replicated datasets in each simulation setting is shown in Figure 2 for comparison. A high weighted Rand index value for a clustering method implies that the particular method is able to recover the true underlying cluster structure and enjoys better performance.

As described in Supplementary Material, hierarchical clustering can be implemented using single linkage, complete linkage or average linkage. It is well-known that single linkage tends to generate elongated clusters while complete linkage normally obtains more spherical clusters. Supplementary Figure 2 gives the results of different linkages of hierarchical clustering in simulated data. Complete linkage is found to provide better results consistently and will be used in hierarchical clustering hereafter. It is also known that *K*-means (as well as PAM) is sensitive to random initial values used for optimization. The more random initial values searched, the better *K*-means clustering result can be obtained. We tested 1, 100 and 1000 random initial values in *K*-means and the results are shown in Supplement Figure 3. Clearly *K*-means with 1 random initial value easily falls into an inadequate local minimum and gives poor performance. Using 100 random initial values gives similar result to 1000 random initial values. In all analyses in this paper 100 random initial values are used in the implementation of *K*-means and PAM.

From Figure 2, it is immediately seen that all the methods, except for SOM, were able to recover the underlying cluster structure well when no scattered genes and no perturbation exist (Fig. 2, 0% noise in Type I). Therefore the comparison of the methods will be based on how the performances are affected as the level of complexity including number of scattered genes and degree of perturbation increases.

From Type I simulated model, it is seen that hierarchical clustering, *K*-means and PAM are very vulnerable to the presence of scattered genes. The mean weighted Rand indexes drop steeply as the percentage of noise genes increases (SOM already performed poorly in the ideal well-separated case). Tight clustering and model-based clustering are both very robust up to existence of 200% scattered genes. In the Type II and Type III simulated models, we see that SOM consistently performs far worse than all others. Hierarchical clustering, *K*-means and PAM are relatively more sensitive to the existence of scattered genes than random perturbations (Fig. 2, Type I and II). In the Type II comparison, tight clustering and model-based clustering start to drop when SD > 0.4 while *K*-means and PAM still perform well. This is because when the experimental variabilities (SD) becomes large, tight clustering and model-based clustering begin to consider some outlying clustered genes as scattered genes which, in fact, is biologically acceptable. In the Type III comparison, tight clustering and model-based clustering performed equally well up to SD = 0.4 even when 200% of scattered genes exist. To provide a fair evaluation, the true number of clusters ($K = 15$) was given to all the clustering methods in the above comparison. We also note that when experimental variabilities SD are large, model-based clustering performs even worse than *K*-means and PAM. This reflects potential computation

Table 4. Summary of real data sets

Organism, sample perturbation and reference	Dimension of data	Annotation
Yeast; cell cycle; Spellman *et al*., 1998	1663 genes × 77 samples	M/G1 boundary; Late G1, SCB regulated; Late G1, MCB regulated; S-phase; S/G2-phase; G2/M-phase
Yeast; environmental changes; Causton *et al*., 2001	1744 genes × 45 samples	response to stress
Human; cell cycle; Whitefield *et al*., 2002	2570 genes × 114 samples	G1/S; S; G2; G2/M; histone genes
human; lung cancer; Bhattacharjee *et al*., 2001	1920 genes × 203 samples	Keratin; metallothionein; melanoma antigen family; major histocompatibility complex (MHC); interferon; immunoglobulin heavy constant; G antigen; collagen

difficulties in model-based clustering in application of a complex data.

3.2 Real datasets

We have used four real datasets to compare the clustering methods for functional predictions of the annotated genes. Prediction accuracy (described in the Methods section) is used for comparing the clustering methods. Table 4 lists a summary of datasets used in our analysis. See Supplementary Material for more detailed data description and preprocessing.

The six clustering methods are implemented with K varying from 5 to 30 for a pooled analysis of functional predictions evaluation. A cross tabulation of the tight clustering result with five clusters and six functional categories, and the corresponding P-values are given in Table 2; the predictive accuracy in Table 2 is calculated as $A_5(0.01) = (6 + 23 + 8 + 5)/(2 \times 167 + 2 \times 137) = 6.9\%$. Overall accuracy for each of the methods is estimated by pooling the results with the number of clusters $K = 5, 6, \ldots, 30$. In Figure 3 the curves of prediction accuracy (y-axis) of all six clustering methods versus the total number of predictions made (x-axis) for varying p-value thresholds $\delta = (10^{-2}, \ldots, 10^{-10}, 10^{-15}, 10^{-20})$ are presented for different datasets. In all four datasets, the predictive performance of tight clustering is the best, followed by model-based clustering. Consistent with the simulation study, SOM and hierarchical clustering perform among the worst in the real datasets. We, however, notice that the total number of predictions made by tight clustering is much smaller than the other methods as many genes are assigned to the scattered gene set in this method.

4 DISCUSSION

Compared with supervised machine learning (or classification) problems, an empirical and unbiased comparison of clustering methods has always been difficult. The underlying true clustering assignment in a real data is generally unknown and the concept of a cluster is not mathematically well-defined in unsupervised learning. Very often a method works well in some datasets but may perform poorly in other datasets owing to different data structure and characteristics. In this paper we focused on the comparison of gene clustering methods of microarray data and evaluated six popular methods by both simulated and real datasets. In the simulated data, various proportions of scattered genes and various degrees of simulated experimental variabilities were simulated to mimic the nature of a microarray data and to examine the robustness of each clustering methods. A weighted Rand index was developed to compare two clustering results with possible scattered genes and to evaluate the simulated data. In the evaluation of the four real datasets, a predictive accuracy plot was utilized to compare the annotation prediction power of different clustering methods. To the authors' knowledge, this is the first comprehensive comparison of popular gene clustering methods in microarray analysis. The results not only provide practical guides to the application of the clustering methods but also elucidate much insight behind the algorithms.

In both simulated and real datasets, model-based clustering and tight clustering consistently performed among the best with model-based clustering slightly less robust in some situations (Fig. 2). This is not surprising in light of the fact that these two methods allow a set of scattered genes not being clustered which helps to outperform other methods (K-means, PAM, hierarchical clustering, and SOM) that assign all genes into clusters. It is worth noted that attempts for estimating the number of clusters were usually not successful especially in datasets with many scattered genes and large perturbations. As a result the correct number of clusters was supplied to all the methods compared in simulated data.

Model-based clustering enjoys full probabilistic modeling and rigorous statistical inference tools including BIC for selecting the number of clusters and the complexity of covariance structure. However, BIC criterion may in practice fail to select the correct model even if the model assumptions are true. The problem is 2-fold. First, BIC is an approximate measure of the Bayesian posterior probability. The performance of BIC depends on the goodness of the approximation. Second, local optimum is usually obtained when estimating the parameters in the model, making the estimation of maximum likelihood and BIC measure vulnerable. Supplementary Figure 4 shows histograms of estimated number of clusters using BIC criterion in the replicated simulation data under various settings. It indicates that when the number of scattered genes and experimental variabilities (SD) increase, the probability that the estimated number of clusters deviates from the truth ($K = 15$) becomes higher. In the comparative study of simulated data, the correct number of clusters is given to each method. We, however, still observed that the performance of model-based clustering dropped steeply when SD becomes large in Figure 2 (Type II and III). As discussed in more detail in Supplementary Material, implementation of model-based clustering often needs special care to avoid computational issues such as singularity and undesirable local minimum.

Conceptually tight clustering can be viewed as a higher-order machinery that can be built upon any other clustering method (e.g. K-means in the original paper). Through evaluation of repeated clustering on subsamples, tight clusters are sequentially generated and the remaining genes are left as scattered genes. From the results of simulated data, it was seen that the resampling evaluation helped

tight clustering to provide better robustness and better ability to deal with scattered genes. On the other hand, tight clustering had the tendency to include fewer clustered genes than the other methods (Fig. 3). Two additional methods considering cluster stability by resampling approach (consensus clustering, Monti *et al*., 2003 and HOPACH, van der Laan and Pollard, 2003) were also evaluated and presented in Supplementary Figure 5. Their inferior performance in simulated data suggested that although clustering methods considering stability are preferred, the resulting performance still depends on how the stability information is effectively utilized.

K-means and PAM were found with similar performance which is expected because PAM only replaces cluster centers with the median points in the loss function of *K*-means. They did not perform as well as model-based clustering and tight clustering owing to their inability of allowing a set of scattered genes. Compared with hierarchical clustering and SOM, they consistently performed better and provided better robustness. Hierarchical clustering and SOM are known to provide better visualization of the clustering results while at the same time they seem to have sacrificed performance (Figs 2 and 3). We especially noticed that SOM did not perform well even in the well-separated simulated clusters with no scattered genes and no perturbation (Fig. 2, 0% noise in Type I). Hierarchical clustering although could perform well in this pure case, the method was very sensitive to both the existence of scattered genes and perturbation. In conclusion, tight clustering and model-based clustering are recommended for gene clustering in expression profile. To date, hierarchical clustering and SOM remain two of the most popular gene clustering methods in many biological studies. Our comparative evaluation, however, suggests cautious use of the two methods. If identifying biologically meaningful cluster patterns and pursuing better annotation prediction are the primary goal and data visualization is secondary, these two methods should be avoided.

Our comparative study has its own limitation. There are many more methods published and used in microarray analysis that we cannot exhaust. In Supplementary Figure 5, we also performed evaluation of simulated data on three other methods including consensus clustering (Monti *et al*., 2003), HOPACH (van der Laan and Pollard, 2003) and fuzzy c-means (Dembélé and Kastner, 2003). The simulation model in this paper applied a hierarchical log-normal model. However, in real situation a heavy-tail or skewed distribution may be more appropriate. Other types of inter-gene and inter-sample correlations may also be considered. In real datasets, we evaluated four representative datasets from yeast and human with various kinds of sample perturbations while other datasets of different types of organism and sample perturbations may be further explored. Despite the above minor limitations, our comparative study has elucidated much insight of the clustering methods and provided a practical guideline for their applications to microarray analysis.

ACKNOWLEDGEMENTS

A.T. and I.M. acknowledge the support of the Fogarty/NIH grant 5D43TW006180 'India-US Research Training Program in Genetics' and the University of Pittsburgh, USA. A.T. thanks the University of Madras and I.M. thanks the University of Burdwan, India for granting leave to work in the University of Pittsburgh, USA. G.C.T. is supported by NSABP, NIH grant CA069974 and CA069651. The authors would like to thank the anonymous reviewers for their helpful comments.

Conflict of Interest: none declared.

REFERENCES

Bhattacharjee,A. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Brown,P.O. and Botstein,D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, **21**, 33–37.

Causton,H.C. *et al*. (2001) Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell*, **12**, 323–337.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **8**, 93–103.

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

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. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fraley,C. and Raftery,A.E. (2002a) MCLUST:Software for model-based clustering, density estimation and discriminant analysis. Technical Report, Department of Statistics, University of Washington, WA.

Fraley,C. and Raftery,A.E. (2002b) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc*., **97**, 611–631.

Handl,J. *et al*. (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics,*, **21**, 3201–3212.

Hartigan,J.A. and Wong,M.A. (1979) A *K*-means clustering alrorithm. *Appl. Stat*., **28**, 126–130.

Hubert,J. and Arabie,P. (1985) Comparing partitions. *J. Classif*., **2**, 193–218.

Kaufman,L. and Rousseeuw,P. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Klebanov,L. *et al*. (2006) A new type of stochastic-dependence revealed in gene expression aata. *Stat. Appl. Genet. Mol. Biol*., **5**, Article 7.

Kohonen,T. (1990) The self-organizing map. *Proc. IEEE*, **78**, 1464–1480.

Li,K.C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl Acad. Sci. USA*, **99**, 16875–16880.

MacQueen,J.B. (1967) Some methods for classification and analysis of multivariate observations. *Proc. fifth Berkeley Symp. Math. Stat. Prob*., **1**, 281–297.

McLachlan,G.J., Bean,R.W. and Peel,D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

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

Medvedovic,M. *et al*. (2004) Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20**, 1222–1232.

Milligan,G.W. and Cooper,M.C. (1985) An examination of procedures for determining number of clusters in a data set. *Psychometrika*, **50**, 159–179.

Monti,S. *et al*. (2003) A resampling-based method for class discovery and visualization of gene-expression microarray data. *Machine Learning*, **52**, 91–118.

Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc*., **66**, 846–856.

RDevelopmentCoreTeam (2004) *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Sorlie,T., Tibshirani,R., Parker,J., Hastie,T., Marron,J.S., Nobel,A., Deng,S., Johnsen,H., Pesich,R., Geisler,S., Demeter,J., Perou,C.M., Lonning,P.E., Brown,P.O., Borresen-Dale,A.-L. and Botstein,D. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA*, **100**, 8418–8423.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

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.

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

Tibshirani,R., Walther,G. and Hastie,T. (2001) Estimating the number of clusters in a dataset via the Gap statistic. *J. R. Stat.Soc. B*, **63**, 411–423.

Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Bostein,D. and Altman,R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tseng,G.C., Oh,M.K., Rohlin,L., Liao,J.C. and Wong,W.H. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*., **29**, 2549–2557.

Tseng,G.C. and Wong,W.H. (2005) Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**, 10–16.

van der Laan,M. and Pollard,K. (2003) A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *J. Stat. Plann. Infer*., **117**, 275–303.

Whitfield,M.L. *et al*. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.

Wu,L.F. *et al*. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet*., **31**, 255–265.

Yang,Y.H. *et al*. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*., **30**, e15.

Yeung,K.Y. *et al*. (2001a) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Yeung,K.Y. *et al*. (2001b) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.

Zhou,X. *et al*. (2002) From the cover: transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.