BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 16 2006, pages 1988–1997
doi:10.1093/bioinformatics/btl284

Gene expression

Clustering microarray gene expression data using weighted Chinese restaurant process

Zhaohui S. Qin

Center for Statistical Genetics, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA

Received on February 16, 2006; revised on April 20, 2006; accepted on May 31, 2006
Advance Access publication June 9, 2006
Associate Editor: John Quackenbush

ABSTRACT

Motivation: Clustering microarray gene expression data is a powerful tool for elucidating co-regulatory relationships among genes. Many different clustering techniques have been successfully applied and the results are promising. However, substantial fluctuation contained in microarray data, lack of knowledge on the number of clusters and complex regulatory mechanisms underlying biological systems make the clustering problems tremendously challenging.

Results: We devised an improved model-based Bayesian approach to cluster microarray gene expression data. Cluster assignment is carried out by an iterative weighted Chinese restaurant seating scheme such that the optimal number of clusters can be determined simultaneously with cluster assignment. The predictive updating technique was applied to improve the efficiency of the Gibbs sampler. An additional step is added during reassignment to allow genes that display complex correlation relationships such as time-shifted and/or inverted to be clustered together. Analysis done on a real dataset showed that as much as 30% of significant genes clustered in the same group display complex relationships with the consensus pattern of the cluster. Other notable features including automatic handling of missing data, quantitative measures of cluster strength and assignment confidence. Synthetic and real microarray gene expression datasets were analyzed to demonstrate its performance.

Availability: A computer program named Chinese restaurant cluster (CRC) has been developed based on this algorithm. The program can be downloaded at http://www.sph.umich.edu/csg/qin/CRC/

Contact: qin@umich.edu

Supplementary information: http://www.sph.umich.edu/csg/qin/CRC/

INTRODUCTION

Genome wide expression analysis with DNA microarray technology (Schena *et al*., 1995; Lockhart *et al*., 1996) has become an indispensable tool in genomics research. Owing to its high intrinsic variability, extracting insightful biological knowledge from microarray experiments remains a grand challenge. Building on the hypothesis that functionally related genes tend to display correlated gene expression patterns, clustering analysis has emerged as a fruitful approach for revealing mechanisms underlying various molecular and cellular processes. The goal of clustering is to identify groups of genes that show correlated expression patterns across a series of experimental conditions (Eisen *et al*., 1998; Hughes *et al*., 2000; Spellman *et al*., 1998; Cho *et al*., 1998).

Most of the clustering approaches implemented today are distance-based, such as Hierarchical clustering (Eisen *et al*., 1998), *K*-means clustering (Tavazoie *et al*., 1999) and Self Organizing Map (Tamayo *et al*., 1999). Although simple and visually appealing, the performances of these methods are sensitive to noise, which is extensive in microarray data. In addition, they have difficulty providing useful information, such as total number of clusters and confidence measures for individual clusters, and they are not flexible enough to accommodate missing data, which are common in microarray data analysis.

Alternative methods are model-based, which are able to circumvent the aforementioned shortcomings. Finite mixture models (FMM) have been proposed in the context of clustering and provide a principled statistical approach (McLachlan and Basford, 1988; Banfield and Raftery, 1993; Fraley and Raftery, 2002). They have been applied to clustering gene expression microarray data (Yeung *et al*., 2001; McLachlan *et al*., 2002; Ghosh and Chinnaiyan, 2002). Using FMM, determining the number of clusters is separated from estimating parameters in the mixture model and cluster assignments. The former can be regarded as a model selection problem and can be estimated using Bayesian information criterion (BIC) (Schwarz, 1978). Subsequently, parameter estimation conditional on the selected number of components is typically achieved by applying the EM algorithm (Dempster *et al*., 1977). Because these two steps are separated, Medvedovic and Sivaganesan showed that the results from the FMM approach are sensitive to the selected 'optimal' number of clusters, which is due to the fact that calculated confidence in a particular clustering does not take into account the uncertainties related to the choice of cluster sizes based on the BIC. Consequently, they are valid only under the model where a specific number of clusters is assumed known (Medvedovic and Sivaganesan, 2002). In practice, without necessary prior knowledge, this condition can hardly be satisfied.

The model-based clustering approach based on the Bayesian infinite mixture model, also known as the Dirichlet process mixture model (Ferguson, 1973, Neal, 2000, Rasmussen, 2000), provides an attractive alternative. This model does not require specifying the number of the mixture components. The clustering procedure can be viewed as a Chinese restaurant process (CRP) (Aldous, 1985, Pitman, 1996). This process gets its name because it can be viewed as a sequential restaurant 'seating arrangement' described as follows. Assume customers arrive sequentially at a Chinese restaurant and are randomly assigned to an infinite number of tables which have unlimited seating capacities. When a new customer arrives,

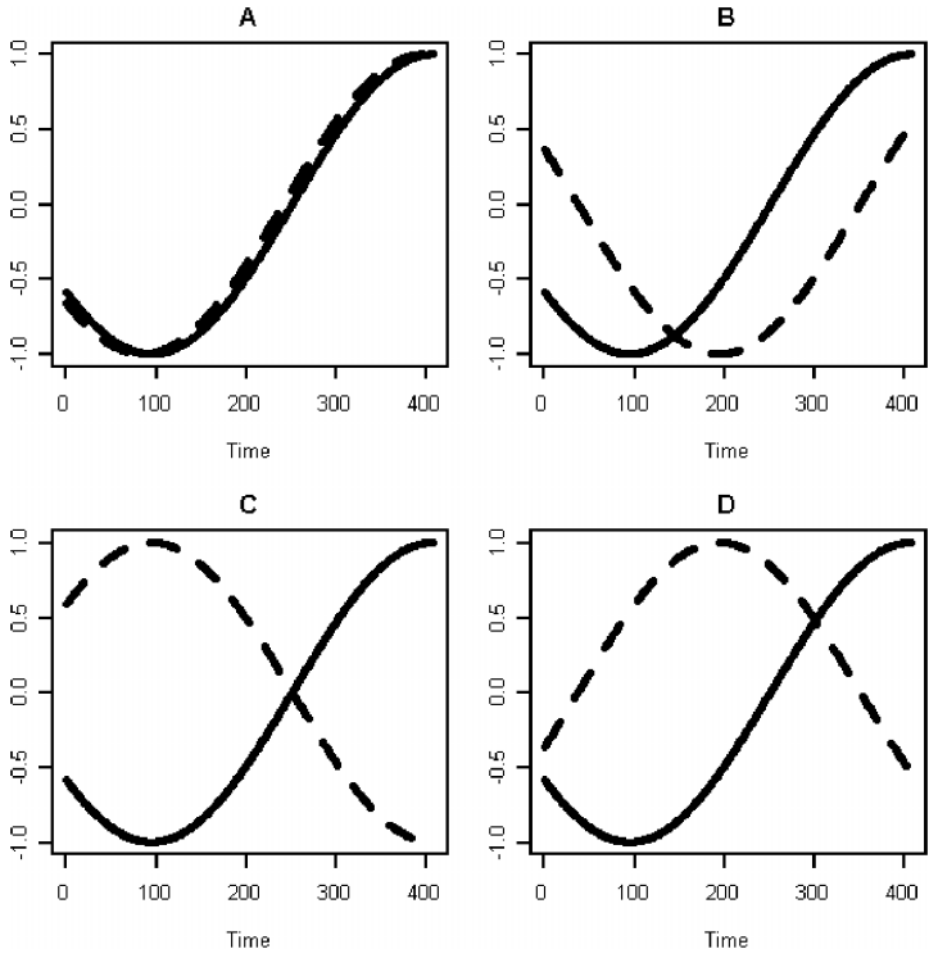

Fig. 1. Illustration of diverse correlation relationships among gene expression profiles: (**A**) synexpression; (**B**) time-shifted; (**C**) inverted and (**D**) time-shifted and inverted.

she will be seated according to the current seating arrangement of all previous customers. In this method, cluster number inference and mixture model parameter estimation are unified and computed simultaneously in an iterative procedure.

One of such models, Gaussian infinite mixture model (GIMM), has recently been applied to clustering microarray gene expression data (Medvedovic and Sivaganesan, 2002, Medvedovic *et al*., 2004). The authors built a Bayesian hierarchical model for this problem, and applied the Gibbs sampler (Gelfand and Smith, 1990, Liu, 2001) to obtain posterior samples for all parameters. The final result is obtained by averaging posterior samples in a post-processing step, where a distance measure is defined for each gene pair based on their co-occurrence frequencies during the iterative process. Subsequently, hierarchical clustering with complete linkage was applied to create the final clusters. Similar approaches related to CRP have also been applied to clustering putative transcription factor binding sites (Qin *et al*., 2003, Jensen *et al*., 2005).

In this article, we devise a modified model-based clustering algorithm based on CRP. The predictive updating technique is applied to integrate out nuisance parameters, which greatly improves the efficiency of the Gibbs sampler procedure. The marginal likelihood is calculated during the iteration. The cluster assignment that produces the highest likelihood is retained as the final result.

A key feature is added to this new clustering approach to allow identifying and assigning genes that have strong yet complicated correlation into the same cluster. Most of the current clustering approaches focus on identifying genes that show identical expression profiles, i.e. genes whose expression levels go up and down simultaneously in all experiments. However, for experiments performed over time, owing to diverse and different regulation mechanisms, such as repressor, feedback loops and regulation cascade in regulatory pathways, groups of genes may display diverse correlation relationships such as time-shifted and/or inverted. See Figure 1 for illustrations of these diverse correlation patterns between expression profiles. These non-standard relationships will be missed by current clustering tools. It is of great interest if we can identify genes showing diverse correlation relationships and put them into the same cluster. Qian and colleagues proposed a novel local clustering technique, which is capable of identifying relationships beyond commonly used 'synexpression' relationships (positive and simultaneous) (Qian *et al*., 2001). They showed that their method is able to uncover new and biologically relevant interactions. Their method is analogous to a local sequence alignment

algorithm such as Smith–Waterman (Smith and Waterman, 1981). One caveat is that, like alignment algorithms, only pairwise relationships are explored in this approach, an additional step is needed to put genes into clusters. By introducing an additional pattern selection step in our clustering algorithm, our approach is able to put genes which display non-synexpression correlation relationships into the same cluster. This property is highly desirable since we will be able to reduce the chance of missing important genes participating in the same biological process, and may be able to reveal a more detailed and comprehensive picture of the underlying biological pathways and regulatory mechanisms under investigation.

METHOD

Statistics model

In model-based clustering, it is assumed that the expression profiles of genes in one cluster are random samples generated from the same distribution and this distribution is different from that of another cluster. As in GIMM, we choose normal distribution to model the expression profiles of these clusters. Suppose that the expression levels of N genes from M experiments are collected. The expression data can be denoted as $X = \{X_{ij},\ i = 1, \ldots, N,\ j = 1, \ldots, M\}$;. Although these experiments may be related (e.g. conducted over time during a cell cycle), for simplicity we assume expression levels from different experiments are independent such that likelihood for each experiment can be multiplied together. It is possible to extend this model to a more complicated one, where multivariate normal distribution with non-zero covariance is used to model the entire expression vector of each gene. It would be much more complex and time-consuming though.

Let $E = (E(1), \ldots E(N))$ be the cluster indicator variable, $E(i) = k, 1 \le k \le K$ denotes that the i-th gene was assigned to the k-th cluster, $1 \le i \le N$. We use $\mid E \mid$ to denote the number of clusters present. Assume that $|E| = K$ (K is unknown). We have $X_{ij} \sim N(\beta_{kj}, \sigma_{kj}^2)$ if $E(i) = k$, $i = 1, \ldots, N$, $j = 1, \ldots, M$ and $k = 1, \ldots, K$. The complete likelihood is as follows:

$$P(X \mid E, \beta, \sigma^2) \propto \prod_{k=1}^{|E|} \prod_{E(i)=k} \prod_{j=1}^{M} \left((\sigma_{kj}^2)^{-1/2} \mathrm{e}^{(-1/2\sigma_{kj}^2)(x_{ij}-\beta_{kj})^2} \right).$$

To ensure proper posterior distributions, we adopt the standard conjugate priors (Gelman *et al*., 1995) for parameters β_{kj} and σ_{kj}^2:

$$P(\beta_{kj} \mid \sigma_{kj}^2) \sim N(\beta_0, \sigma_{kj}^2),$$
$$P(\sigma_{kj}^2) \sim \text{Inv Gamma}(a, b).$$

Here β_0, a and b are all assumed known. The prior distribution for the cluster indicator variable E is assumed to be a Dirichlet process. The detailed information about these prior distributions can be found in the online Supplementary information.

The clustering procedure is modeled after a weighted Chinese restaurant seating scheme (Lo, 2005). The prior for cluster assignment is a CRP, where the probability of joining each existing table is proportional to the size of that table. The seating probability is the product of the prior and the likelihood, which compare the properties of the new customer and those of the people at the table. Therefore the seating process follows a weighted CRP. The 'weight' is the likelihood ratio of the customer joining that table versus she starts a new one. Then table assignment can be regarded as sampling from a multinomial distribution with probabilities proportional to the conditional posterior probabilities Q_k:

$$Q_k = p(E(i) = k \mid E(-i), X), \quad k = 1, \ldots, K.$$

Each assignment can be viewed as an update of one component of the indicator vector E conditional on all the other components. Therefore, the whole process can be naturally fit into a Gibbs sampler framework, such that the memberships can be updated iteratively until convergence.

An appealing feature of CRP-based clustering approaches lies in its natural way of modifying the number of components. This number can increase or decrease naturally within this modeling framework, obviate the need to resort to complicated and time-consuming computation techniques to accommodate the changes in parameter space.

Predictive updating

The complete parameter vector of this model is $\left(E(1), \ldots, E(N), \left(\beta_{1j}, \sigma_{1j}^2\right)_{j=1}^{M}, \ldots, \left(\beta_{Kj}, \sigma_{Kj}^2\right)_{j=1}^{M}\right)$ which contains many parameters. Among them, only $E(i)$s are parameters that are of interest. The rest can be regarded as nuisance parameters. Including these parameters will slow down the Gibbs sampler procedure. The predictive updating technique (Liu, 1994; Chen and Liu, 1996) can be applied to improve the efficiency of our algorithm. We integrate out unwanted nuisance parameters, $(\beta_{kj}, \sigma_{kj}^2)$—mean and variance of normal distributions of each cluster analytically from the likelihood function. For each cluster, after the incorporation of prior distributions, we have

$$\iint \prod_{E(i)=k} p\left(x_{ij} \mid \beta_{kj}, \sigma_{kj}^2\right)\, p\left(\beta_{kj} \mid \beta_0, \sigma_{kj}^2\right)\, p\left(\sigma_{kj}^2\right) \mathrm{d}\beta_{kj} \mathrm{d}\sigma_{kj}^2$$
$$= \frac{b^a}{\Gamma(a)} \frac{(2\pi)^{-n/2}}{\sqrt{n_k + 1}} \frac{\Gamma\left(\frac{n_k}{2} + a\right)}{\left(b + \frac{1}{2} \left(\sum_{E(i)=k} x_{ij}^2 + \beta_0^2 - \frac{\left(\sum_{E(i)=k} x_{ij} + \beta_0\right)^2}{n_k + 1} \right) \right)^{(n_k/2)+a}}.$$

This formula will be used to calculate the likelihood ratio and the assignment probabilities Q_k. Details can be found in the online Supplementary information.

Algorithm

A Gibbs sampler is implemented to carry out the clustering scheme, using the predictive updating technique to improve its efficiency. The clustering procedure can be summarized as follows. The details of implementing this algorithm can be found in the online Supplementary information.

(1) Initialization: randomly assign genes into an arbitrary number of K_0 clusters $1 < K_0 \le N$.

(2) For each gene i, perform the following reassignment:

(a) Remove gene i from its current cluster. Conditional on the current assignment of all the other genes, calculate the probability of this gene joining each of the existing clusters as well as being alone in its own cluster:

$$Q_k = P(E(i) = k \mid E(-i), X), \quad k = 0, \ldots, K.$$

$k = 0$ indicates that this gene is standing alone by itself, $k > 0$ means assignment to an existing cluster.

(b) Assign gene i to one of the $K + 1$ possible clusters according to probabilities Q_k, $k = 0, 1, \ldots, K$. Update indicator variable $E(i)$ based on the assignment. Also update total number of clusters K, if gene i, starts a new cluster or leaves a singleton cluster.

(c) Repeat the above two steps for every gene, and repeat for a large number of rounds until convergence.

Marginal likelihood is monitored during the iteration process, only the result corresponding to the highest likelihood will be reported at the end. Such strategy allows us to avoid the difficult label switching problem commonly seen in model-based clustering. To avoid getting trapped in local modes, we use 10 parallel Markov chains, and report the best result from all the runs.

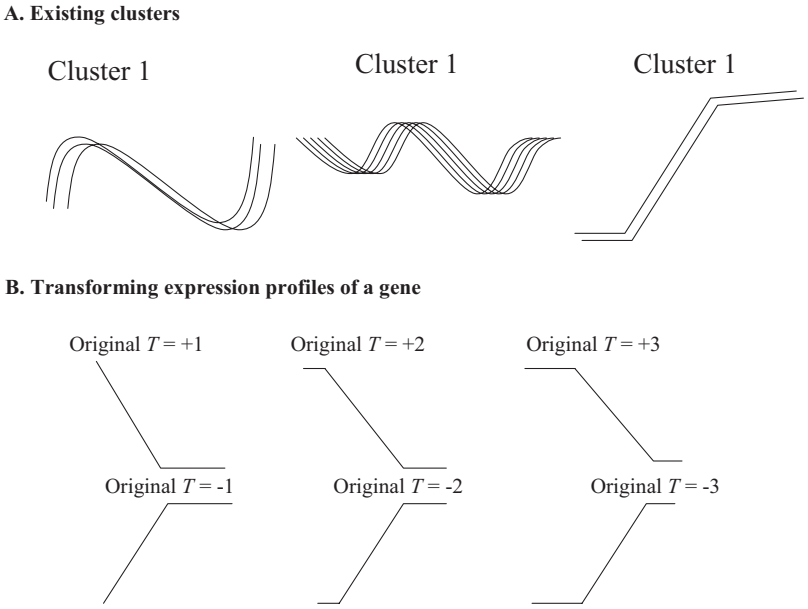

Fig. 2. Illustration of the pattern selection step during cluster assignment. (**A**) illustration of currently existing clusters; (**B**) transforming the expression profile of a gene. Assume $s = 3$. Although the original gene expression profile does not fit any of the three clusters. A transformed one ($T = -3$) does fit cluster 3, and will be assigned to it.

Complicated time-dependent correlation relationships

By requiring expression levels to follow the same set of normal distributions within each cluster, we only cluster together genes displaying positive and simultaneous correlation relationship, like most of the current existing clustering approaches are capable of doing. However biological systems are complex, which result in a great variety of relationships among genes such as time-shifted and/or inverted. It will be highly desirable if a clustering algorithm can allow such genes to be clustered together. Note that if the expression profile of a gene can be viewed as an time-shifted and/or inverted version of the consensus pattern of that cluster, then transforming the original profile by inversion or time-shifting will produce profiles that can be clustered into the right cluster using the aforementioned clustering algorithm. For this reason, we added an extra pattern selection step. That is, when assigning gene i, we first transform its original profile by inverting and shifting (up to s units), then compare both the original and the transformed expression profiles to each of the existing clusters to find the best fit (please see Fig. 2 for illustration). For example, if we choose $s = 3$, then profiles $(x_{i1}, x_{i2}, \ldots, x_{iM-2})$, $(x_{i2}, x_{i3}, \ldots, x_{iM-1})$ and $(x_{i3}, x_{i4}, \ldots, x_{iM})$ as well as $(-x_{i1}, -x_{i2}, \ldots, -x_{iM-2})$, $(-x_{i2}, -x_{i3}, \ldots, -x_{iM-1})$ and $(-x_{i3}, -x_{i4}, \ldots, -x_{iM})$ will be compared with each of the existing clusters to see if there is a cluster that fits one of the profiles well. The indicator variables $E(i)$, is expanded to $[E(i), T(i)]$, where $T(i)$ takes values $\pm 1, \ldots, \pm s$, to specify which transformation the gene had gone through in order for it to be clustered in its current cluster. The same Gibbs sampler procedure can be applied to sample the new augmented $E(i)$, which follows a multinomial distribution with $2sK + 1$ possible outcomes.

Note that genes shown time-shifted and/or inverted correlation pattern in time course experiments may in fact belong to different clusters in the sense of biological functions, especially when the study involves only few time points and the waiting time between experiments are long. Because in such situations unrelated genes may show time-shifted and/or inverted correlation pattern just by chance, not due to the pre- or post-coregulation effect we try to detect. Therefore, we recommend that the optional inversion and time-shifted switches of CRC be turned off when analyzing such datasets.

Missing data

Missing data are ubiquitous in microarray gene expression datasets. There are many reasons contributing to their occurrences: experimental artifacts and mishaps such as insufficient resolution, image corruption or quality control considerations (Troyanskaya *et al*., 2001). Most of the existing clustering procedures such as hierarchical clustering, *K*-means, are unable to handle missing data. Pre-processing step is needed to either remove or impute back those missing data. On the other hand, sporadic random missing data are hardly an issue for model-based approaches. When expression level for the i-th gene at the j-th experiment is missing, we simply do not have any information to judge whether this unobserved data support the clustering decision one way or the other. So based on this experiment alone, the probabilities for this gene to join each of the existing clusters are all equal. The clustering decision for this gene has to be placed on information collected from other experiments. To be specific, if a gene contains missing expression data, we only use the observed partial expression profile of that gene, and likewise only use the corresponding partial expression profile of each cluster to calculate the likelihood and Bayes ratio for this gene joining all clusters.

Posterior probability measurement of each cluster assignments

Owing to the noisy nature of microarray experiments, uncertainty needs to be considered in statistical procedures such as clustering. Traditional approaches such as Hierarchical or *K*-means clustering do not directly grant uncertainty measures. On the other hand, model-based clustering approaches naturally provide such information. Under our Bayesian scheme, conditional on the final clustering result, we can calculate the posterior probability of each gene belonging to its assigned cluster [Q_k as in the (2b) step of the aforementioned algorithm]. Using these probabilities, the user has the option of only keeping those genes with posterior probability greater than a certain threshold specified *a priori*, and remove those genes that are only weakly associated with a cluster.

A potential problem is that, when a gene contains missing data, we may inflate the variability of its posterior assignment probability since only fraction of the data is used. Our solution is to mark such posterior probabilities when reporting final results to warn the users that caution need to be exercised when interpreting them.

Cluster strength measures

After clustering analysis has been performed on a particular dataset, it is of great interest to determine which clusters are more statistically significant than others, since such information may lead to further biological insights. In a model-based setting, the significance can be evaluated by calculating the so called Bayes ratio for each cluster to indicate how close the members of this cluster are. We refer to this measure as the tightness of the cluster. To be specific, for a particular cluster, we calculate two different likelihoods, one is under the assumption that all the genes in this cluster follow the same set of normal distributions across experiments, hence by our definition, they belong to the same cluster; the other is under the alternative assumption that each of these genes follow its own unique set of normal distributions. Assume that the first n_1 genes belong to the same cluster. Incorporating priors, the Bayes ratio can be described as follows

$$\frac{\prod_{j=1}^{M} \int P(x_{1j}, x_{2j}, \ldots, x_{n_1 j} \mid \mu_j, \sigma_j^2) P(\mu_j, \sigma_j^2) \mathrm{d}\mu_j \mathrm{d}\sigma_j^2}{\prod_{j=1}^{M} \prod_{i=1}^{n_1} \int P(x_{ij} \mid \mu_{ij}, \sigma_{ij}^2) P(\mu_{ij}, \sigma_{ij}^2) \mathrm{d}\mu_{ij} \mathrm{d}\sigma_{ij}^2}.$$

The final tightness measure is the log Bayes ratio normalized by the number of genes in that cluster. Essentially, this statistic reflects the level of homogeneity among genes in this cluster. The higher the value, the more likely that these genes are indeed generated from the same distribution. Another important measure is the cluster stability, which we derive from the similarity measure used in GIMM (Medvedovic and Sivaganesan, 2002). Medvedovic and Sivaganesan defined the similarity measure between a pair of genes as the proportion of times during iteration that these two genes were assigned to the same cluster together (Medvedovic and Sivaganesan, 2002). We define the stability measure of a cluster as the average similarity measure of all gene pairs in this cluster. The higher the value, the more likely that these genes are closely related. The tightness and stability measures offer quantitative assessment of the clusters from different perspectives. Based on these quantities, the investigators will be able to triage all the clusters generated and focus their attention on the most promising ones.

RESULTS

The aforementioned algorithm with all the features described has been implemented in a C++ program called CRC (Chinese restaurant cluster). To test its performance, we applied it to three different datasets: synthetic datasets, the yeast galactose metabolism dataset (Idekar *et al*., 2001) and the *Bacillus anthracis* sporulation dataset (Liu *et al*., 2004). In addition to CRC, we also tested two well-established model-based clustering algorithms: MCLUST (Fraley and Raftery, 1999, http://www.stat.washington.edu/mclust) and GIMM (Medvedovic and Sivaganesan, 2002, Medvedovic *et al*., 2004, http://eh3.uc.edu/gimm/). The default setting of these programs was used.

Clustering accuracy

To evaluate the performance of different clustering approaches, we need a statistic that is able to measure the agreement between different clustering results. There are many statistics that have been proposed. We adopted the adjusted Rand index (ARI) (Hubert and Arabie, 1985) as the measure in our study. ARI has also been used by Yeung *et al*. (2001, 2003) and Medvedovic *et al*. (2004) in their studies. Its values lie between 0 and 1, and a higher value indicates a higher level of agreement. Milligan and Cooper recommended it as the measure of agreement based on extensive empirical studies (Milligan and Cooper, 1986). ARI is derived from the Rand index (Rand, 1971), which is defined as the number of pairs that are either in the same groups in both partitions or in different groups in both partitions, divided by the total number of pairs of objects. ARI adjusts the score so that its expected value in the case of random partitions is 0. The detailed formula on how to calculate ARI can be found in the online Supplementary information.

Synthetic datasets

Each simulated dataset contains 400 rows (genes) and 20 columns (experiments). The expression profiles were generated from five different clusters. Three of them formed by genes displaying the periodic sine function $x_{ij} = (k/2)\sin(\pi jk/10 - \pi k/4) + \varepsilon$, $j = 1, 2, \ldots, 20$ and $k = 1, 2, 3$. One cluster displays a monotone increasing or decreasing profile: $x_{ij} = -1 \pm j/10 + \varepsilon$, The other cluster corresponds to a constant expression profile $x_{ij} = a + \varepsilon$, where a is an uniform random variate between -1 and 1: $a \sim \text{Uniform}(-1,1)$. A random perturbation term ε ($\varepsilon \sim N(0, 0.5)$) is added to each data point x_{ij} to account for the noise associated with gene expression levels observed. A total of 100 such datasets were generated. The trace plot of a sample simulated dataset can be found in Figure 3(A) and Supplementary Figure S1(A). For this type of data, CRC, MCLUST and GIMM all perform almost perfectly.

The presence of complex correlation relationships among expression profiles such as time-shifted and/or inverted complicated the clustering problem. We mimic such situations to investigate how well these algorithms perform. In the original simulated dataset, expression profiles of some genes were replaced by the ones that show non-synexpression relationships with others. Except for the constant expression profile cluster, all remaining clusters contain such 'complex' profiles. The proportions of time-shifted and inverted expression profiles are 10% each and they may overlap. A total of 100 such datasets were generated, trace plots of a sample dataset is shown in Figure 3(B) and Supplementary Figure S1(B). The clustering results using CRC and MCLUST (GIMM produce almost the same results as MCLUST) are summarized in Table 1. It is evident that CRC performed well even in the presence of genes displaying complex relationships other than synexpression, the average ARI is 0.972. Furthermore, not only does it assign genes into the right clusters, it also provides accurate estimate of the true cluster numbers. As a comparison, MCLUST produces higher clustering error (the average ARI reduce to 0.852) since it is unable to identify genes that display non-synexpression correlations with a cluster. The superior performance of CRC in these datasets suggests that both the weighted Chinese restaurant seating scheme and the added pattern selection step work well in the context of microarray gene expression data analysis.

Yeast galactose dataset

This dataset originally came from the study conducted by Ideker *et al*. (2001), and later used by Yeung *et al*. (2003) and Medvedovic *et al*. (2004) for performance comparison of various clustering algorithms. This dataset consists of 205 genes whose expression patterns reflect four functional categories in the gene Ontology Consortium (Ashburner *et al*., 2000). This implies that the true

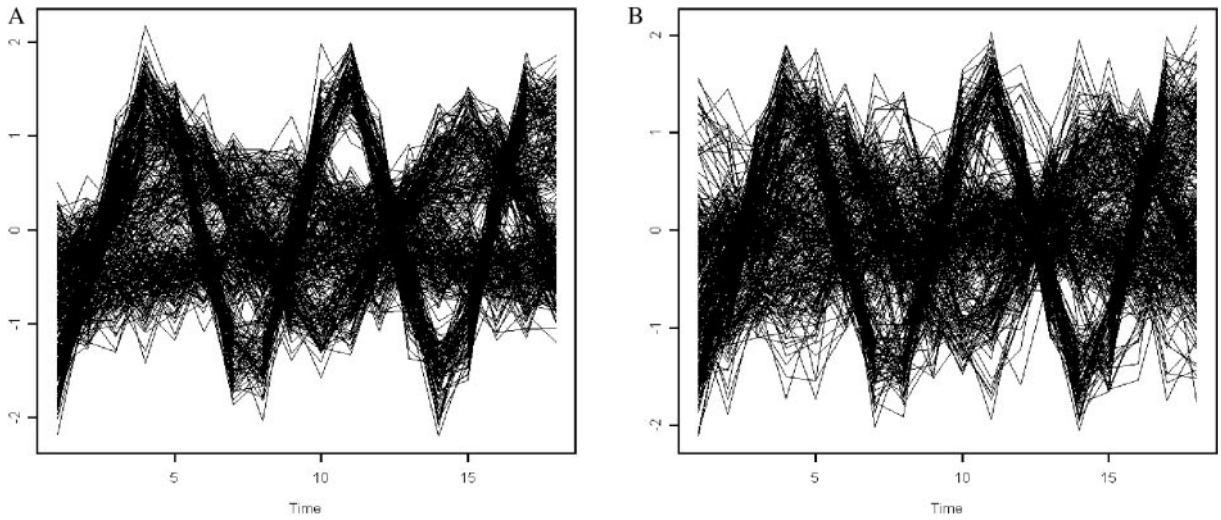

Fig. 3. Trace plots of simulated gene expression profiles over time. Two different sample datasets were shown here. (**A**) Simulated with only positive and simultaneous relationships; (**B**) contains genes displaying time-shifted and/or inverted relationships. A different version of these two plots which display the five clusters using different colors is shown in Supplementary Figure S1.

Table 1. Performance of CRC and MCLUST on synthetic datasets containing non-synexpression correlated genes

Algorithm	Clustering accuracy		Cluster# estimate (true =5)	
CRC	0.972	0.072	4.87	0.34
MCLUST	0.855	0.025	9.00	0.00

number of clusters is probably four. In the original experiment, Microarrays were used to measure the mRNA expression profiles of yeast growing under 20 different perturbations to the GAL pathway. Four replicate hybridizations were performed for each condition. Data were downloaded from Dr Yeung's website http://expression.microslu.washington.edu/expression/kayee/medvedovic2003/medvedovic_bioinf2003.html. The original dataset contains ~8% of missing data, these missing values have been imputed using KNNimpute (Troyanskaya *et al*., 2001) in the version we downloaded. The average expression profiles are illustrated in Figure 4 and Supplementary Figure S2.

This dataset contains results from four replicated runs, to assess performance of various algorithms under different scenarios, we tested with and without replicates. GIMM is able to handle replicated data, so the original data were used, for CRC and MCLUST, we use the average of the four replicates as the expression level. Since CRC is a stochastic procedure, 100 runs were performed on each dataset and the average performance on these runs was taken as the final results. The results are summarized in Table 2 and a cluster-specific trace plot based on a sample result of CRC is shown in Figure 4 and Supplementary Figure S2.

Overall, CRC performs the best on these datasets. It achieves the highest ARI for the replicate dataset as well as all four single replicate datasets. For GIMM, since this method *per se* does not suggest the 'right' number of clusters to use (M. Medvedovic, personal communication), we tried several different cluster sizes around the 'true' cluster size four, and reported the best result as the final performance measure of GIMM. With this given advantage, GIMM performs better than MCLUST overall.

Another encouraging result is that, the cluster number estimates provided by CRC, which is inferred during the clustering procedure, are quite accurate for this real dataset. MCLUST has an internal function that is able to select the model with the optimal number of clusters and variance structure using BIC. However, for this dataset, these estimates are not as accurate. We acknowledge that here we assume that the 'true' number of clusters is four, which is the number of GO categories these genes belong to. However, caution has to be exercised since it is not guaranteed that the genes in the same GO functional category should be co-expressed.

The original dataset contains missing data. Since CRC is able to accommodate missing data intrinsically, no pre-processing step is needed to impute them. To evaluate its performance in the presence of missing data, we applied CRC to the original dataset, where no imputation has been performed on the missing data. Again, we distinguish single replicate and multiple replicate cases. On average, there are 146 (71.2%) genes contain missing data in at least one of the 20 experiments. The average missing proportion in these four single replicate datasets is 7.8%. The clustering results shown in Table 2 indicate that there is no significant difference in terms of clustering accuracy when moderate amount of missing data are present. When there are replicates, a natural strategy is to take average of all non-missing expression levels for each gene at each experiment. So missing data only occurs when all four replicates are missing. Adopting this strategy, there are only 11 (5.4%) genes containing missing data; the overall proportion of missing data is 0.27%. From the results shown in Table 2, there is little performance difference in terms of clustering accuracy using this summarizing strategy (0.967 versus 0.955).

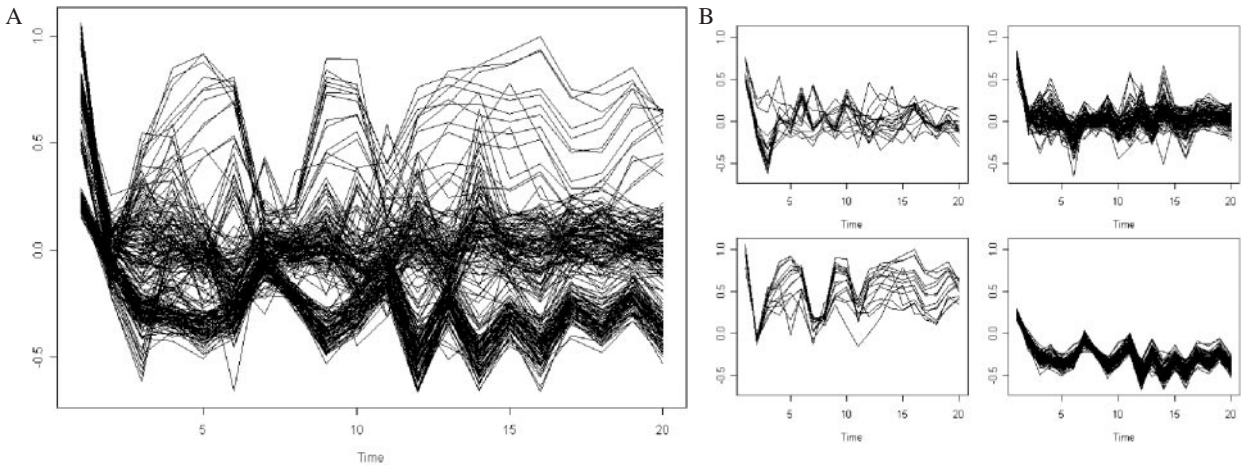

Fig. 4. Trace plots of the real galactose data adapted from Ideker *et al*. (2001). There are 205 genes that belong to four different GO functional categories. Data were collected from 20 different perturbation experiments conducted on the GAL pathway. (**A**) Expression profiles of all 205 genes; (**B**) a sample output from the CRC program, each subgraph represents a single cluster identified. A different version of these plots which display the four clusters using different colors is shown in Supplementary Figure S2.

Table 2. Performance of CRC, MCLUST and GIMM on yeast galactose datasets

	All data	Replicate 1	Replicate 2	Replicate 3	Replicate 4
	Clustering accuracy				
MCLUST	0.724	0.544	0.370	0.686	0.724
GIMM	0.953	0.660	0.571	0.561	0.667
CRC	0.967	0.833	0.794	0.762	0.765
CRC (missing)	0.955	0.833	0.783	0.752	0.773
	Cluster number estimate (true = 4)				
MCLUST	2	2	9	2	2
CRC	4.00	5.65	5.02	6.00	5.28
CRC (missing)	4.38	4.84	5.03	6.11	5.51

B.anthracis sporulation data

In a recent study (Liu *et al*., 2004), a global gene expression microarray experiment was conducted to study the synchronized temporal pattern changes in gene expression during *B.anthracis* sporulation process. The log-transformed and normalized data were kindly provided by Dr Nicholas Bergman. Samples were collected every 15 min over 5 h. The original dataset contains 5594 genes across 20 different time points. We apply a variation filter as described in Tamayo *et al*. (1999) to eliminate genes that did not change significantly across the time course, with both relative change and absolute change thresholds set at 1 U. As a result, 1314 genes pass this filter. There are 20 gene expression levels that are missing (0.08%) among the remaining genes. CRC was applied to these genes and generated 25 clusters. Using a stringent cut-off value, we only retain significant genes which is defined as those with posterior probability of belonging to the cluster >0.9. We display the nine clusters that show clear trend over time in Figure 5 (information about all 25 clusters and all significant genes in these clusters can be found on our website). From the figure, we see that complex relationships are quite common. In fact, among all the genes shown in this figure, there are ∼70% genes that show positive and simultaneous correlation, ∼12% genes show positive and time-shifted correlation, ∼14% show negative and simultaneous correlation and 4% show negative and time-shifted correlation. Using most of the current clustering algorithms, those genes displaying such complex relationships with the consensus pattern of the cluster would have been put into different clusters. Using CRC, new hypothesis stem from genes displaying non-synexpression correlations may be generated.

We also compared the performance of CRC with MCLUST and GIMM using this dataset. We applied KNNimpute (Troyanskaya *et al*., 2001) to fill in missing data such that all three programs could use the same dataset. The best clustering result reported by MCLUST only contains two clusters, the bigger cluster contains 1038 (79%) genes. We also ran GIMM, and cut the tree to obtain 25 clusters, number of clusters reported by CRC. The biggest cluster contains 697 (53%) genes, and the smallest one contains only two genes. Whereas the sizes of clusters generated by CRC distributed more evenly, range from 16 to 96. This observation seems to suggest CRC produces more informative result for this dataset.

DISCUSSION

In summary, we implemented a model-based clustering strategy based on CRP for clustering microarray gene expression data. This algorithm is able to cluster genes and infer the number of clusters simultaneously and with high accuracy. Predictive updating technique was applied during the iterative assignment process to improve the efficiency of the Gibbs sampler. A unique feature of CRC is that it is able to recognizing genes that display complex correlation relationships such as time-shifted and/or inverted with others and put them into the right cluster. Another benefit is that the new algorithm is able to accommodate missing data seamlessly such that separate missing data imputation step can be avoided. In addition, CRC provides multiple strength measurements for

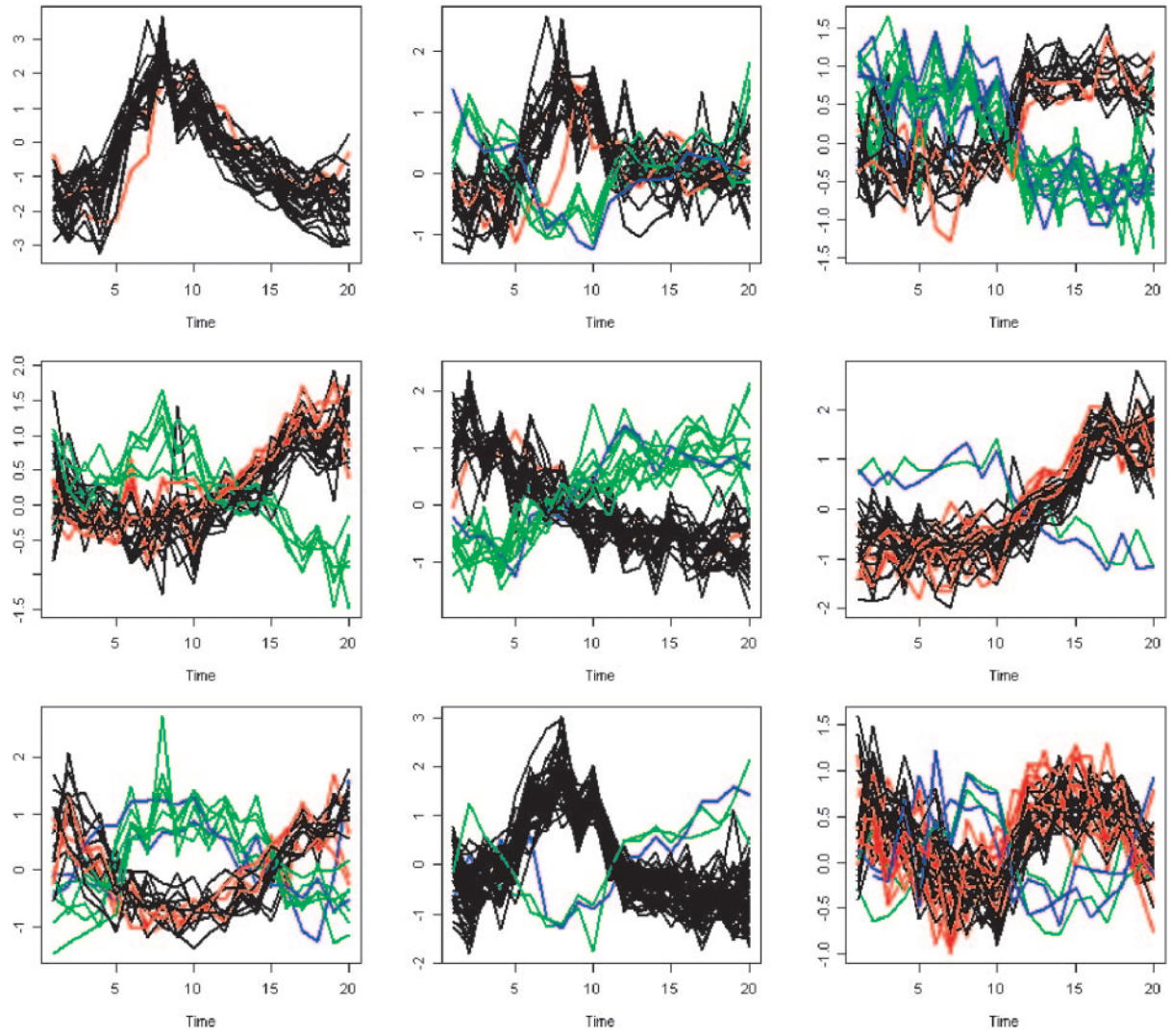

Fig. 5. Trace plots of significant genes belonging to clusters identified from the *B.anthracis* sporulation experiment (Liu *et al*., 2004). There are 1314 genes being clustered and 302 significant ones (with posterior probability >0.9) were shown in this figure. Data were collected from 20 experiments conducted every 15 min over 5 h. Black lines represent majority of the genes in that cluster; red line indicates genes showing time-shifted correlation relationship with the majority of the genes in that cluster; green line indicates genes showing inverted correlation relationship with the majority of the genes in that cluster; blue line indicates genes showing time-shifted and inverted correlation relationship with the majority of the genes in that cluster.

each cluster produced, including tightness measure and stability measure. Such information can help investigators to identify significant clusters to generate high-quality hypotheses and perform follow up studies. Tests conducted on simulated as well as real datasets indicate that the new algorithm works well, even with the presence of missing data and complicated correlations among genes.

The newly proposed pattern selection step during the iterative procedure is especially attractive since it enables the identification of complicated correlation relationships other than synexpression. Using the *B.anthracis* dataset, we found that as much as 30% of significant genes clustered in the same group shown non-synexpression relationships with the consensus pattern of their cluster. This demonstrated that by adding this step, we will be able to achieve better understanding of the complex underlying biological processes, and generate new, more sensible and accurate hypotheses.

Another contribution is that we demonstrated that the marginal likelihood can be a good performance indicator of our clustering procedure. We rely on it to determine the final clustering result. In other model-based clustering approach such as GIMM, the final result is obtained by taking average of the posterior samples. A dissimilarity measure is defined between a pair of genes to determine whether they belong to the same cluster together. One drawback is that when the number of objects is large, a huge number of pairwise distances need to be computed, the complexity is about $O(n^2)$. For example, to cluster 20 000 genes, the number of pairwise distance is $\sim$200 000 000, which can be cumbersome to manipulate. For model-based clustering, gene–gene comparison was replaced by gene–cluster comparison, the complexity of

computing is $O(n\log n)$ [the expected number of cluster under the Dirichlet process is $\log(n)$], which greatly reduces the computation cost.

Although CRC performed well in datasets we have tested, there is still ample room for further improvement. In the present algorithm, we did not consider the correlation among experiments. A naïve Bayes scheme was used assuming expression levels from different experiments are independent. This maybe true for experiments performed under different conditions, but for experiments conducted over time, such as in the cell cycle study, the expression levels from adjacent time points are expected to be correlated. Studies have shown that better performance can be achieved when correlation between experiments is considered (Yeung *et al*., 2003, Medvedovic *et al*., 2004). Another reason for considering correlation is experiment replicates. Strong correlation among replicates is expected and should be accounted for. We plan to extent our model to take correlation structure into account, and give users the options to choose between these two models.

Microarray technology is not precise; the expression profile from a single gene is often not informative owing to experimental noises. Clustering techniques are able to combine information and borrow strength from each other. Among all the clustering techniques proposed, it has been shown that model-based clustering algorithms, in general, outperform traditional distance-based approaches (Yeung *et al*., 2004; Medvedovic *et al*., 2004). This can be explained by its explicit modeling of uncertainty involved and ability to average out noise. A similar reason is behind the favorable performance of model-based algorithm for clustering putative DNA binding motifs (Qin *et al*., 2003). Admittedly, clustering large-scale, multiple assay microarray gene expression data is still very challenging. The recently developed resampling-based methods such as tight clustering (Tseng and Wong, 2005) and consensus clustering (Swift *et al*., 2004) presented encouraging results. We believe that such techniques may help CRC further to obtain stable and meaningful clusters when applied to larger and more complex datasets.

CRP-based clustering approaches are known to be computation-intensive and therefore time-consuming. However, with a relative simple model we assumed and the predictive updating techniques, our Gibbs sampler runs converged fairly rapidly (please see Supplementary Figure S3 for a likelihood trace plot). For the yeast galactose dataset with 205 genes and 20 experiments, it only takes CRC $\sim$15 s to run on a SUN opteron server under the default setting, whereas the same dataset takes GIMM $\sim$4 min to finish. We also tested a much larger synthetic dataset that contains 3000 genes and 50 time points to mimic the real life cases. Under the default setting, CRC took $\sim$36 min and GIMM took $\sim$5.4 h to finish.

ACKNOWLEDGEMENTS

We thank Drs Michael Elliott, Debashis Ghosh for fruitful discussion and their insightful comments on an earlier draft of this manuscript. We thank Dr Mario Medvedovic for helping with the GIMM software, Dr Ka Yee Yeung for sharing the yeast galactose data and Dr Nicholas Bergman for providing the *B.anthracis* sporulation data. We are also grateful to the two anonymous reviewers for their constructive suggestions and critiques.

Conflict of Interest: none declared.

REFERENCES

Aldous,D. (1985) *Exchangeability and Related Topics*. New York: Springer-Verlag, Vol. 1117.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Banfield,J.D. and Raftery,A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Chen,R. and Liu,J.S. (1996) Predictive updating methods with application to bayesian classification. *J. R. Stat. Soc. B*, **58**, 397–415.

Cho,R.J. *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*., **2**, 65–73.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via EM algorithm. *J. R. Stat. Soc. B*, **39**, 1–38.

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

Ferguson,T.S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Stat*., **1**, 209–230.

Fraley,C. and Raftery,A.E. (1999) MCLUST: Software for model-based cluster analysis. *J. Classif*., **16**, 297–306.

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

Gelfand,A.E. and Smith,A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc*., **85**, 398–409.

Gelman,A., Carlin,J.B., Stern,H.S. and Rubin,D.B. (1995) *Bayesian Data Analysis*. Reprinted 1997 edn. Chapman & Hall, London.

Ghosh,D. and Chinnaiyan,A.M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275–286.

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

Hughes,J.D. *et al*. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol*., **296**, 1205–14.

Ideker,T. *et al*. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–34.

Jensen,S.T. *et al*. (2005) Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics*, **21**, 3832–3839.

Liu,J. (2001) *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York.

Liu,J.S. *et al*. (1994) Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27–40.

Lo,A.Y. (2005) Weighted Chinese restaurant processes. *Cosmos*, **1**, 59–63.

Lockhart,D.J. *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol*., **14**, 1675–1680.

McLachlan,G.J. and Basford,K.E. (1988) *Mixture Models: Inference and Applications To Clustering*. Marcel Dekker, New York.

McLachlan,G.J. *et al*. (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**, 1194–206.

Medvedovic,M. *et al*. (2004) Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20**, 1222–32.

Milligan,G.W. and Cooper,M.C. (1986) A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behav. Res*., **21**, 441–458.

Neal,R. (2000) Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat*., **9**, 249–265.

Pitman,J. (1996) *Some Developments of the Blackwell-MacQueen Urn Scheme*. IMS, Hayward, California.

Qian,J. *et al*. (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Biol*., **314**, 1053–1066.

Qin,Z.S. *et al*. (2003) Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat. Biotechnol*., **21**, 435–439.

Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc*., **66**, 846–850.

Rusmussen,C. (2000) The infinite Gaussian mixture model. *Advances in Neural information Processing Systems*, MIT Press, 2000, **12**, 554–560.

Schena,M. *et al*. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat*., **6**, 461–464.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Swift,S. *et al*. (2004) Consensus clustering and functional interpretation of gene-expression data. *Genome Biol*., **5**, R94.

Tamayo,P. *et al*. (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. *et al*. (1999) Systematic determination of genetic network architecture. *Nat. Genet*., **22**, 281–285.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

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.

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

Yeung,K.Y. *et al*. (2003) Clustering gene-expression data with repeated measurements. *Genome Biol*., **4**, R34.