BIOINFORMATICS

Vol. 20 no. 8 2004, pages 1222–1232
DOI: 10.1093/bioinformatics/bth068

# Bayesian mixture model based clustering of replicated microarray data

M. Medvedovic[1,*], K.Y. Yeung[2] and R.E. Bumgarner[2]

[1]Department of Environmental Health, Center for Genome Information, University of Cincinnati Medical Center, 3223 Eden Avenue ML 56, Cincinnati, OH 45267-0056, USA and [2]Department of Microbiology, Box 358070, University of Washington, Seattle, WA 98195, USA

Received on July 14, 2003; revised on November 4, 2003; accepted on November 5, 2003
Advance Access publication February 10, 2004

**ABSTRACT**

**Motivation:** Identifying patterns of co-expression in microarray data by cluster analysis has been a productive approach to uncovering molecular mechanisms underlying biological processes under investigation. Using experimental replicates can generally improve the precision of the cluster analysis by reducing the experimental variability of measurements. In such situations, Bayesian mixtures allow for an efficient use of information by precisely modeling between-replicates variability.

**Results:** We developed different variants of Bayesian mixture based clustering procedures for clustering gene expression data with experimental replicates. In this approach, the statistical distribution of microarray data is described by a Bayesian mixture model. Clusters of co-expressed genes are created from the posterior distribution of clusterings, which is estimated by a Gibbs sampler. We define infinite and finite Bayesian mixture models with different between-replicates variance structures and investigate their utility by analyzing synthetic and the real-world datasets. Results of our analyses demonstrate that (1) improvements in precision achieved by performing only two experimental replicates can be dramatic when the between-replicates variability is high, (2) precise modeling of intra-gene variability is important for accurate identification of co-expressed genes and (3) the infinite mixture model with the 'elliptical' between-replicates variance structure performed overall better than any other method tested. We also introduce a heuristic modification to the Gibbs sampler based on the 'reverse annealing' principle. This modification effectively overcomes the tendency of the Gibbs sampler to converge to different modes of the posterior distribution when started from different initial positions. Finally, we demonstrate that the Bayesian infinite mixture model with 'elliptical' variance structure is capable of identifying the underlying structure of the data without knowing the 'correct' number of clusters.

**Availability:** The MS Windows™ based program named Gaussian Infinite Mixture Modeling (GIMM) implementing the Gibbs sampler and corresponding C++ code are available at http://homepages.uc.edu/~medvedm/GIMM.htm

**Contact:** Mario.Medvedovic@uc.edu

**Supplemental information:** http://expression.microslu.washington.edu/expression/kayee/medvedovic2003/medvedovic_bioinf2003.html

*To whom correspondence should be addressed.

## 1 INTRODUCTION

Identifying patterns of gene co-expression in microarray data by cluster analysis has been a productive approach to uncovering molecular mechanisms underlying biological processes under investigation. The goal of the cluster analysis is to identify groups of genes with similar patterns of expression across multiple experimental conditions. Such a multiple-gene at a time analysis approach utilizes the inherent parallelism in measuring gene expression by microarray technology. The biological significance of cluster analyses results has been demonstrated in numerous studies. The reproducibility and the biological validity of conclusions drawn from results of a cluster analysis will depend directly on the reproducibility of patterns of expressions and groups of genes associated with these patterns, which are both identified in the cluster analysis.

The utility of a cluster analysis, as well as the utility of any other analytic approach, is dependent on the quality of the data that is analyzed. The traditional experimental approach to improving precision of inherently noisy microarray data is by performing experimental replicates. In the context of cluster analysis, the increased power of detecting existing patterns in the data that is achieved by performing experimental replicates has been discussed by Dougherty *et al.* (2002). Replicated observations also allow us to quantify precisely the experimental noise in measurements for each gene at each experimental condition. When the level of experimental variability varies between different genes and between different experimental conditions, experimental replicates

are necessary for assessing the reproducibility of observed patterns.

Virtually all classical clustering algorithms (Eisen *et al*., 1998; Tavazoie *et al*., 1999; Tamayo *et al*., 1999; Yeung *et al*., 2001), as well as a multitude of brand new procedures (Herrero *et al*., 2001), have been applied in the context of clustering microarray data. However, the majority of the currently used approaches are not able to accommodate appropriately replicated microarray data. Previously, we demonstrated that clustering approaches that make use of the information about the between-replicates variability generally perform better than approaches that do not (Yeung *et al*., 2003). We also showed that model-based clustering approaches generally outperform heuristic methods in this context. Generally, algorithms based on the finite mixture model (FMM; Yeung *et al*., 2001; Fraley and Raftery, 2002) applied to averaged profiles performed better than heuristic algorithms. On the other hand, the Bayesian infinite mixture model (IMM; Medvedovic and Sivaganesan, 2002), capable of capturing gene- and experiment-specific variability, performed better than finite mixtures algorithms that do not have built-in error models for replicated measurements. In addition to precisely modeling intra- and inter-gene variabilities in expression measurements, the IMM approach is unique in its ability to incorporate uncertainties related to the choice of the 'correct' number of clusters in the analysis. In our experiments with simulated and real world datasets, Bayesian infinite mixtures outperformed all alternative approaches.

In this paper, we describe different variants of the Bayesian FMM- and IMM-based algorithms for clustering replicated microarray data and investigate their performance on simulated and real world datasets. The new models not previously described are the IMM model with the 'elliptical' variance structure and Bayesian FMMs with 'spherical' and 'elliptical' variance structures. The performance of new procedures is compared with alternative approaches for clustering replicated data based on heuristic and model-based clustering methods. We investigate the consequences of incorporating uncertainties related to the choice of the number of clusters by comparing otherwise equivalent FMMs and IMMs. The need to incorporate information on between-replicates variation is addressed by comparing equivalent models with different variance structures. Effects of experimental replicates are demonstrated by analyzing simulated and real world datasets with different numbers of replicates. A particularly striking observation is that the 'elliptical' IMM with an automatic algorithm for creating clusters performed as well as or better than the majority of other clustering approaches that utilized the information about the correct number of clusters in the data. Finally, we demonstrate that the previously described Gibbs sampler has problems with multimodal posterior distributions that are induced by some datasets and describe heuristic modifications that effectively circumvent this problem.

## 2 METHODS

### 2.1 Bayesian mixtures for replicated microarray data

Suppose that $T$ gene expression profiles were observed across $M$ experimental conditions and that the experiment was replicated $G$ times. If $y_{ijg}$ represents the $g$th replicate of the expression measurement for the $i$th gene under $j$th experimental condition, then $\boldsymbol{y}_{ig} = (y_{i1g}, y_{i2g}, \ldots, y_{iMg})$ represents the expression profile in the $g$th replicate for the $i$th gene. Suppose that $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM})$ is the mean expression profile for the $i$th gene, where $x_{ij}$ is the mean expression level of the $i$th gene at the $j$th experimental condition. In our hierarchical model, each gene mean expression profile is viewed as being generated by one out of $Q$ different underlying expression patterns. Expression profiles generated by the same pattern form a cluster of similar expression profiles. If $c_i$ is the classification variable indicating the pattern that generates the $i$th mean expression profile ($c_i = q$ means that the $i$th expression profile was generated by the $q$th pattern), then a 'clustering' is defined by a set of classification variables for all genes, $\mathbf{C} = (c_1, c_2, \ldots, c_T)$. Underlying patterns generating clusters of expression profiles are represented by multivariate Gaussian random variables. Profiles clustering together are assumed to be a random sample from the same multivariate Gaussian distribution.

The hierarchical structure of the model is described in terms of a directed acyclic network in Figure 1. Nodes (squares) in this diagram represent random variables, and directed arcs (arrows) specify conditional dependences between variables in terms of the directed Markov property, which states that a variable is conditionally independent of its non-descendants, given its parents in the model. $\mathbf{M} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_Q)$ and $\boldsymbol{\Sigma} = (\sigma_1^2\mathbf{I}, \ldots, \sigma_Q^2\mathbf{I})$ denote means and variance-covariance matrices of multivariate Gaussian random variables defining $Q$ underlying patterns, respectively ($\mathbf{I}$ denotes the identity matrix). $\psi_{ij}$ represents the between-replicates variance for the $i$th gene at the $j$th experimental condition. $\boldsymbol{\psi}_i = (\psi_{i1}, \ldots, \psi_{iM})$ is the diagonal of the between-replicates variance–covariance matrix for the $i$th gene, and off-diagonal elements are assumed to be zero. This represents the assumption that experimental replicates under different experimental conditions are obtained by independent experiments, which is commonly the case in practice. We call the model 'spherical' when the between-replicates variance of a single gene is assumed to be homogeneous across all experimental conditions (i.e. $\psi_{i1} = \psi_{i2} = \cdots = \psi_{iM}$ for all $i = 1, \ldots, T$). Otherwise, we say that the model is 'elliptical' with respect to between-replicates covariance structure. Variables $(\boldsymbol{\lambda}, \tau)$, $(\beta, \phi)$ and $\alpha$ are hyper-parameters in prior distributions of model parameters $\mathbf{M}$, $\boldsymbol{\Sigma}$ and $\mathbf{C}$, respectively. The specification of the prior distribution for classification variables ($\mathbf{C}$) determines whether the model represents finite or infinite mixtures. Such a model for strictly Bayesian infinite mixtures

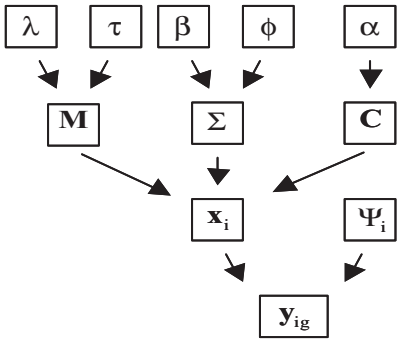


**Fig. 1.** Bayesian mixture model.

with the spherical variance structure has been described by Medvedovic and Sivaganesan (2002). Here, we describe spherical and elliptical models for both finite and infinite mixtures.

*2.1.1 Conditional distributions for finite and infinite mixture models* Distribution of the data for gene $i$, given the mean expression profile $\mathbf{x}_i$ and the between-replicates variance vector, $\boldsymbol{\psi}_i$, is

$$p(\mathbf{y}_{ig} \mid \mathbf{x}_i, \psi_i) = f_N(\mathbf{y}_{ig} \mid \mathbf{x}_i, \mathbf{I}_{\psi_i^2}), \quad g = 1, \ldots, G, \quad i = 1, \ldots, T,$$

where $\mathbf{I}_{\psi_i^2}$ is a diagonal matrix with $\psi_i^2$ on the diagonal and $f_N(\cdot \mid m, v)$ denoting the probability density function (p.d.f) of a Gaussian random variable with the mean vector $m$ and the variance–covariance matrix $v$. Mean expression profiles from the $q$th cluster are distributed as

$$p(\mathbf{x}_i \mid \mathrm{c}_i = \mathrm{q}, \mathbf{M}, \Sigma) = f_N(\mathbf{x}_i \mid \boldsymbol{\mu}_q, \sigma_q^2), \quad i = 1, \ldots, T.$$

In the FMM with the number of components fixed at $Q$, the conditional prior distribution of $c_i$, given $\alpha$ and all other classification variables $\mathbf{C}_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_T)$, is

$$p(c_i = q \mid \mathbf{C}_{-i}, \alpha) = \frac{n_{-i,q} + \alpha/Q}{T - 1 + \alpha}, \quad i = 1, \ldots, T; \quad q = 1, \ldots, Q,$$

where $n_{-i,q}$ is the number of profiles in the cluster $q$ without the $i$th profile (defined by $\mathbf{C}_{-i}$) (Rasmussen, 2000; Neal, 2000). In the case of infinite mixtures, $Q$ is let to go to infinity, which results in following prior probabilities of the $i$th profile being generated by an already existing component $q$,

$$p(c_i = q \mid \mathbf{C}_{-i}, \alpha) = \frac{n_{-i,q}}{T - 1 + \alpha},$$

and the probability that a new component should be created,

$$p(c_i \neq c_j, j \neq i \mid \mathbf{C}_{-i}, \alpha) = \frac{\alpha}{T - 1 + \alpha}.$$

Differences between the prior distributions of classification variables are propagated into the corresponding conditional posterior distributions, given data and other parameters. Finite mixtures posterior classification probabilities are

$$p(c_i = q \mid \mathbf{C}_{-i}, \mathbf{y}_{i1}, \ldots, \mathbf{y}_{iG}, \boldsymbol{\mu}_q, \sigma_q^2, \boldsymbol{\psi}_i) = b \frac{n_{-i,q} + \alpha/Q}{T - 1 + \alpha} f_N\left(\bar{\mathbf{y}}_{i\bullet} \mid \boldsymbol{\mu}_q, \mathbf{I}_{\sigma_q^2 + \psi_i^2/G}\right) \quad i = 1, \ldots, T, \ q = 1, \ldots, Q;$$

$$\bar{\mathbf{y}}_{i\bullet} = \frac{\sum_g \mathbf{y}_{ig}}{G}.$$

On the other hand, infinite mixtures posterior classification probabilities also describe the probability of creating a new component/cluster:

$$p(c_i = q \mid \mathbf{C}_{-i}, \mathbf{y}_{i1}, \ldots, \mathbf{y}_{iG}, \boldsymbol{\mu}_q, \sigma_q^2, \psi_i) = b \frac{n_{-i,q}}{T - 1 + \alpha} f_N\left(\bar{\mathbf{y}}_{i\bullet} \mid \boldsymbol{\mu}_q, I_{\sigma_q^2 + \psi_i^2/G}\right) \quad i = 1, \ldots, T, \ q = 1, \ldots, Q,$$

$$p(c_i = c_j, j \neq i \mid C_{-i}, y_{i1}, \ldots, y_{iG}, \boldsymbol{\psi}_i, \alpha) = b \frac{\alpha}{T - 1 + \alpha} \int f_N\left(\bar{y}_{i\bullet} \mid \mu_q, I_{\sigma_q^2 + \psi_i^2/G}\right) \times p(\mu_q, \sigma_q^2 \mid \lambda, \tau, \beta, \varphi)\, d\mu_q\, d\sigma_q^2,$$

where $b$ is a normalizing constant assuring that all probabilities for a single profile add to 1 in both finite and infinite models. In the current implementation, the integral above is approximated by

$$f_N\left(\bar{\mathbf{y}}_{i\bullet} \mid \boldsymbol{\mu}_p, \mathbf{I}_{\sigma_p^2 + \psi_i^2/G}\right),$$

where $\boldsymbol{\mu}_p$ and $\sigma_p$ are sampled from their prior distributions.

The differences between the spherical and elliptical intra-gene variance models is in the composition of vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_T$. In the elliptical model, each component of each vector is estimated separately, while in the spherical model, all data within the same gene are pooled to estimate the single intra-gene variance. The prior distributions for the elements of $\psi$s are identical in both situations.

Instead of specifying parameter $\alpha$, which describes the prior belief about the number of clusters in the data, as we did previously (Medvedovic and Sivaganesan, 2002), in the infinite mixture model, we treat $\alpha$ as a random variable with a vague gamma prior and sample new $\alpha$ in each cycle of the Gibbs sampler (Rasmussen, 2000). In the case of the FMM, $\alpha$ is set to 1. Prior and posterior distributions for all parameters in both models are given in the Web supplement.

## 2.2 The Gibbs sampler

The Gibbs sampler (Gelfand and Smith, 1990) is a general procedure for sampling observations from a multivariate distribution. In short, a Gibbs sampler proceeds by iteratively drawing observations from complete posterior conditional

distributions of all components. In the limit, such a sequence describes observations from the joint multivariate distribution. In our case, the joint posterior distribution of interest is the posterior distribution of all parameters, given data. Since in the limit this process generates the sample from the distribution of interest, it can be assumed that the empirical distribution of generated clusterings, $\mathbf{C}^B, \mathbf{C}^{B+1}, \ldots$, after $B$ 'burn-in' samples approximates the true posterior distribution of clusterings.

### 2.3 Cluster formation and inference

Given the sequence of clusterings $(\mathbf{C}^B, \mathbf{C}^{B+1}, \ldots, \mathbf{C}^S)$ generated by the Gibbs sampler after $B$ 'burn-in' cycles, pair-wise probabilities for two genes to be generated by the same pattern are estimated as

$$P_{ij} = \frac{\text{No. of samples after 'burn-in' for which } c_i = \mathrm{c}_j}{S - B}.$$

Using these probabilities as similarity measures or equivalently using $D_{ij} = 1 - P_{ij}$ as the distance measure, clusters of similar expression profiles are created by applying one of the traditional linkage principles. We used the complete linkage with the distance of 1 to create clusters in a completely unsupervised fashion. This has an intuitive justification by defining clusters as groups of genes for which there exist at least one pair of genes such that the probability of them being co-expressed is equal to 0. In the analysis section, we refer to this method as 'Auto' IMM clustering since the method chooses the number of clusters automatically. When the number of clusters is known for any reason, and in the case of the FMM models, we used the average linkage principle to create the pre-specified number of clusters.

In the context of Bayesian inference, such posterior pair-wise probabilities of co-expression carry more meaning than any other traditionally used similarity measures because it actually describes the degree of belief or confidence in the statement that two profiles are generated by the same underlying expression pattern. Such posterior probabilities incorporate all the information about various sources of noise in the data and, in the case of infinite mixtures, the uncertainties with respect to the correct number of clusters in the data. In contrast, traditional pair-wise measures of similarity or distance are only meaningful in the context of the estimated null-distribution that describes the probability of observing a specific value purely by chance due to random fluctuations in the data. Constructing a null distribution that would take into account all sources of variability and uncertainties seems to be a difficult problem, and we are not aware of any existing solution. Furthermore, posterior pair-wise probabilities of co-expression incorporate information from the whole dataset, while the traditional pair-wise distance/similarity measures use only the data specific to two profiles at a time. Resulting improvements in precision have been demonstrated by Medvedovic and Sivaganesan (2002).

### 2.4 Convergence of the Gibbs sampler

Two aspects of the Gibbs sampler convergence that generally need to be assessed are the appropriateness of the 'burn-in' period after which a Gibbs sampler has attained its stationary distribution, and the mixing of the sampler, which describes how well a finite sample obtained by the Gibbs sampler approximates the target distribution. In situations when the posterior distribution being approximated is multi-modal, the Gibbs sampler often has difficulties in switching between different areas with high posterior probabilities, which can result in 'poor mixing'. That is, the sampler will be unable to describe the whole posterior distribution in a computationally feasible number of steps. This can result in either the sampler getting trapped in a sub-optimal mode of the posterior distribution resulting in sub-optimal clustering results; or, because the sampler fails to visit all areas with significant posterior probabilities, confidence estimates in the generated clustering will be biased. Mixing problems related to multi-modality of the target distribution can often be identified by running multiple independent samplers from different initial positions and comparing generated samples.

We devised a heuristic procedure for identifying mixing problems with our sampler and the heuristic solution to the potential problem of it being trapped in a sub-optimal mode based on the idea of 'reverse annealing' (Medvedovic, 2000). If $\pi(\cdot)$ is the target posterior distribution, 'reverse annealing' refers to 'flattening' the posterior distribution using the transformation

$$\pi^{(\xi)}(x) = \frac{\pi^{\xi}(x)}{K(\xi)}, \quad \xi < 1,$$

where $K(\xi)$ is the normalizing constant. Based on this general idea, if $p(c_i = j \mid C_{-i}, \Theta)$ is the conditional posterior probability of placing the $i$th profile into the $j$th cluster, then 'flattened probabilities' are defined as

$$p(c_i = j \mid \mathbf{C}_{-i}, \Theta)^{\xi} = \frac{p(c_i = j \mid \mathbf{C}_{-i}, \Theta)^{\xi}}{K(\xi)}, \quad \xi < 1.$$

We use such modified probabilities during the 'burn-in' period with the 'cooling' sequence that ensures sampling from the original posterior distribution after the 'burn-in'. Suppose that the sampler is run for a total of 20 000 cycles, with the first 10 000 being discarded as the 'burn-in'. Then we define the 'cooling' sequence in such a way that $\xi_0 = 0.01$ represents an almost completely flattened distribution, $\xi_{10\,000} = 0.99$ represents an almost non-modified distribution and $\xi_n \to 1$ as $n \to \infty$, where $n$ is the number of Gibbs sampler iterations. In the analysis of simulated data described in this paper, we use the linear logistic function $\xi_n = 1/[1 + \exp(4.6 - 0.0009n)]$ to define the cooling sequence. As in the general 'annealing' optimization approach, allowing the sampler to identify the highest mode of the distribution while it is flattened and mixing is easy, and slowly transitioning to the unmodified distribution facilitates a high probability of staying in the highest

**Table 1.** Summary of all clustering methods that were compared in the analysis.

| Variants of Bayesian mixtures | Averaged over replicated measurements | SD-weighted similarity |
|---|---|---|
| IMM (elliptical, spherical and averaged) | Hierarchical complete linkage (correlation or distance) | Hierarchical complete linkage (correlation or distance) |
| IMM auto (elliptical, spherical and averaged) | Hierarchical average linkage (correlation or distance) | Hierarchical average linkage (correlation or distance) |
| FMM (elliptical, spherical and averaged) | $k$-means (correlation or distance) | $k$-means (correlation or distance) |
| | Hierarchical finite mixture algorithms (MCLUST-HC) | FITSS (e.g. MCLUST-HC) |

mode. Results of our analysis seem to support this heuristic argument. In the supplemental material, we demonstrate in detail effects of the annealing in the analysis of one simulated dataset (Figure S1 and accompanying text).

In the analysis of the galactose dataset, such an annealing strategy did not solve the problem of the Gibbs sampler converging to a sub-optimal mode (Figure S2 and accompanying text in the Web supplement). This is probably due to the existence of multiple modes with the basin of attraction of similar posterior probability. In such situations, it is crucial to average over all such areas of high posterior probabilities. In this case, we apply an additional heuristic modification of the algorithm by stopping the annealing process before $\xi$ reaches 1. That is, we allow the cooling parameter to converge only to $\xi_{\max} < 1$. We choose the stopping point, $\xi_{\max}$, based on three criteria: (1) we look for the maximum correlation between pair-wise posterior probabilities generated by four independent samplers; (2) among all stopping points with the approximately maximal correlations, we choose the one that had the 'best mixing properties', meaning that clustering labels for each gene changed at least once after the burn-in period and (3) if there is more than one stopping point that satisfies the first two criteria, we choose the largest one. In our analyses, such a heuristic modification of the Gibbs sampler performed very well in all situations. However, since such an analysis is computationally intensive, we used this approach only for the analysis of galactose data.

## 3 RESULTS

We examined the performance of various Bayesian mixture models on synthetic and the real-world datasets. Synthetic datasets were designed to reflect noise and artifacts commonly seen in the microarray data. For all situations, we established benefits of performing replicated microarray experiments as well as benefits of using Bayesian mixtures over heuristic procedures and the traditional finite mixture models that do not take into account gene-specific between-replicates variability.

The ability of different clustering procedures to re-create the known underlying structure of the data was measured in terms of the adjusted RAND index. The adjusted RAND index (Hubert and Arabie, 1985) is a metric designed to assess the degree of agreement between two partitions. The RAND index itself is defined as the number of pairs of objects 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. The adjusted RAND index adjusts the score so that its expected value in the case of random partitions is 0. A high-adjusted RAND index indicates a high level of agreement between the true partition of the data and the partition (i.e. clustering) generated by a clustering procedure.

### 3.1 Alternative clustering procedures

Performance of Bayesian mixtures-based methods was compared with commonly used alternative clustering procedures (Table 1). Detailed descriptions of all algorithms and error-weighted correlation and Euclidean distances can be found in Yeung *et al*. (2003) and in the Web supplement.

### 3.2 Synthetic sine wave data

Each dataset consists of 400 data points (genes), 20 attributes (experiments) and six classes, each defined by a distinct underlying pattern. Four of the six classes follow the periodic sine function $x_{ij} = \sin(2\pi j/10 - \pi q/4)$, $j = 1, \ldots, 20$, for $c_i = q$, for $q = 1, 2, 3, 4$, and the remaining two classes follow the non-periodic linear function $x_{ij} = j/20$ for $c_i = 5$ and $x_{ij} = -j/20$ for $c_i = 6$. The $G$ replicates ($G = 1, \ldots, 4$) for the $i$th gene were generated as $y_{ijg} = x_{ij} + \varepsilon_{ijg}$ where $\varepsilon_{ijg}$ were generated as independent observations from the normal distribution with mean 0 and SD $\psi_{ij}$ which were randomly sampled from SD observed in the data described by Hughes *et al*. (2000). In the 'spherical' scenario, $\psi_{ij}$s were sampled once per gene, resulting in $\psi_{i1} = \psi_{i2} = \cdots = \psi_{iM}$, while in the 'elliptical' scenario a new $\psi_{ij}$ was sampled for each gene for each experimental condition resulting in an 'elliptical' variance structure. In the 'high-noise' scenario, randomly sampled $\varepsilon_{ijg}$s were multiplied by a factor of 6.

Several conclusions can be reached based on results shown in Tables 2 and 3. First, as the number of experimental replicates increases, the ability of all clustering methods to re-create the underlying structure of the data improved. Regardless of the true variance structure of the data, all methods that take into account between-replicates variance show significant

**Table 2.** (A) Average adjusted RAND indices for the 'elliptical' simulated data and (B) average number of clusters induced by different IMM algorithms with Auto cluster creation

| Method | Number of replicates | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Low Noise | | | | High Noise | | | |
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| *A* | | | | | | | | |
| Heuristic methods | | | | | | | | |
| Average link; correlation | 0.76 | 0.76 | 0.80 | 0.90 | 0.20 | 0.34 | 0.32 | 0.43 |
| Average link; distance | 0.86 | **1.00** | 0.96 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 |
| Complete link; correlation | 0.70 | 0.71 | 0.75 | 0.82 | 0.26 | 0.39 | 0.44 | 0.52 |
| Complete link; distance | 0.96 | 0.93 | 0.96 | 0.96 | 0.07 | 0.27 | 0.27 | 0.53 |
| *k*-means; correlation | 0.76 | 0.76 | 0.80 | 0.90 | 0.31 | 0.52 | 0.57 | 0.62 |
| *k*-means; distance | 0.86 | **1.00** | 0.96 | **1.00** | 0.00 | 0.21 | 0.16 | 0.48 |
| SD-adjusted heuristic methods | | | | | | | | |
| Average link; correlation | NA | 0.32 | 0.59 | 0.61 | NA | 0.18 | 0.48 | 0.59 |
| Average link; distance | NA | 0.96 | **1.00** | **1.00** | NA | 0.13 | 0.47 | 0.67 |
| Complete link; correlation | NA | 0.29 | 0.46 | 0.63 | NA | 0.18 | 0.48 | 0.57 |
| Complete link; distance | NA | 0.96 | **1.00** | **1.00** | NA | 0.24 | 0.67 | 0.83 |
| *k*-means; correlation | NA | 0.47 | 0.59 | 0.71 | NA | 0.22 | 0.44 | 0.51 |
| *k*-means; distance | NA | 0.95 | **1.00** | **1.00** | NA | 0.49 | 0.78 | 0.77 |
| Model-based methods | | | | | | | | |
| IMM—elliptical | 0.93 | **1.00** | **1.00** | **1.00** | **0.41** | 0.98 | **1.00** | **1.00** |
| IMM—elliptical Auto | 0.60 | **1.00** | **1.00** | **1.00** | 0.34 | **0.99** | **1.00** | **1.00** |
| IMM—spherical | 0.93 | **1.00** | 0.96 | **1.00** | **0.41** | 0.48 | 0.66 | 0.75 |
| IMM—spherical Auto | 0.60 | 0.98 | 0.99 | **1.00** | 0.34 | 0.82 | 0.85 | 0.92 |
| IMM—averaged | 0.93 | 0.93 | 0.99 | **1.00** | **0.41** | 0.47 | 0.67 | 0.77 |
| IMM—averaged Auto | 0.60 | 0.61 | 0.62 | 0.64 | 0.34 | 0.45 | 0.46 | 0.50 |
| FMM—elliptical | **0.98** | **1.00** | **1.00** | **1.00** | 0.30 | 0.98 | **1.00** | **1.00** |
| FMM—spherical | **0.98** | **1.00** | **1.00** | **1.00** | 0.30 | 0.86 | 0.89 | 0.95 |
| FMM—averaged | **0.98** | 0.80 | 0.95 | **1.00** | 0.30 | 0.44 | 0.50 | 0.57 |
| MCLUST-HC | 0.97 | 0.97 | 0.99 | **1.00** | 0.36 | 0.44 | 0.57 | 0.51 |
| MCLUST-FITTS | 0.97 | 0.96 | 0.97 | **1.00** | 0.36 | 0.38 | 0.43 | 0.50 |
| *B* | | | | | | | | |
| IMM—elliptical Auto | 16.2 | 6.0 | 6.0 | 6.0 | 15.6 | 6.0 | 6.0 | 17.6 |
| IMM—spherical Auto | 16.2 | 7.6 | 7.8 | 7.0 | 15.6 | 10.4 | 10.2 | 6.0 |
| IMM—averaged Auto | 16.2 | 16.2 | 14.2 | 14.2 | 15.6 | 17.2 | 16.8 | 10.0 |

'IMM elliptical Auto', 'IMM spherical Auto' and 'IMM averaged Auto' refer to the corresponding IMM algorithm with the automatic selection of the number of clusters. Highest average indices for each situation are in boldface and underlined.

improvement over corresponding methods that use only average profiles. Second, Bayesian mixtures showed significantly better performance than heuristic methods. This was the case when only averaged profiles were used in the analysis and much more so when between-replicates variability was used in the analysis. Third, the 'elliptical' variance model performed significantly better than the 'spherical' model when the true variance structure was 'elliptical', and it performed only slightly worse when the actual variance structure was 'spherical'. Finally, when the correct variance structure was specified, the automatic IMM clustering performed as well as IMM and FMM approaches with the specified correct number of clusters. Not only that, in these situations the automatic clustering produced clusters that closely corresponded to the correct classification, but it also produced exactly the correct number of clusters.

### 3.3 Synthetic sporadic-genes data

Real-world microarray data generally contain a certain portion of sporadic genes that do not belong to any particular pattern in the data. To investigate the effects of such genes on the performance of different clustering methods, we modified our synthetic data by replacing 10% of genes with 'sporadic genes' whose expression measurements at all experimental conditions were generated by drawing independent observations from the uniform random variable on the interval $[-1, 1]$. Such sporadic genes were designated to the seventh class in the calculation of RAND indices.

The relative performance of the tested clustering procedures in this situation resembled the results when there were no sporadic genes (Fig. 2), in the sense that methods that utilized the information about between-replicates variability performed better than the methods that ignored this

**Table 3.** (A) Average adjusted RAND indices for the 'spherical' simulated data (B) average number of clusters induced by different IMM algorithms with Auto cluster creation

| Method | Number of replicates | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Low Noise | | | | High Noise | | | |
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| *A* | | | | | | | | |
| Heuristic methods | | | | | | | | |
| Average link; correlation | 0.61 | 0.67 | 0.67 | 0.82 | 0.25 | 0.33 | 0.42 | 0.40 |
| Average link; distance | 0.28 | 0.70 | 0.73 | 0.90 | 0.00 | 0.00 | 0.00 | 0.00 |
| Complete link; correlation | 0.69 | 0.72 | 0.73 | 0.79 | 0.33 | 0.46 | 0.49 | 0.53 |
| Complete link; distance | 0.62 | 0.76 | 0.80 | 0.96 | 0.00 | 0.00 | 0.02 | 0.09 |
| *k*-means; correlation | 0.70 | 0.67 | 0.67 | 0.82 | **0.38** | 0.51 | 0.56 | 0.57 |
| *k*-means; distance | 0.44 | 0.70 | 0.73 | 0.90 | 0.00 | 0.00 | 0.00 | 0.00 |
| SD-adjusted heuristic methods | | | | | | | | |
| Average link; correlation | NA | 0.35 | 0.64 | 0.73 | NA | 0.14 | 0.35 | 0.34 |
| Average link; distance | NA | 0.77 | 0.75 | 0.90 | NA | 0.00 | 0.00 | 0.00 |
| Complete link; correlation | NA | 0.33 | 0.70 | 0.77 | NA | 0.11 | 0.35 | 0.49 |
| Complete link; distance | NA | 0.77 | 0.85 | 0.96 | NA | 0.09 | 0.14 | 0.23 |
| *k*-means; correlation | NA | 0.52 | 0.75 | 0.75 | NA | 0.18 | 0.53 | 0.55 |
| *k*-means; distance | NA | 0.83 | 0.75 | 0.90 | NA | 0.14 | 0.17 | 0.28 |
| Model-based methods | | | | | | | | |
| IMM—elliptical | **0.90** | 0.85 | 0.91 | 0.96 | 0.23 | 0.80 | 0.87 | 0.90 |
| IMM—elliptical Auto | 0.52 | **1.00** | **1.00** | **1.00** | 0.28 | 0.77 | 0.87 | 0.90 |
| IMM—spherical | **0.90** | **1.00** | **1.00** | **1.00** | 0.23 | **0.84** | **0.90** | **0.93** |
| IMM—spherical Auto | 0.52 | **1.00** | **1.00** | **1.00** | 0.28 | **0.84** | 0.89 | **0.93** |
| IMM—averaged | **0.90** | 0.94 | 0.91 | 0.92 | 0.23 | 0.42 | 0.46 | 0.64 |
| IMM—averaged Auto | 0.52 | 0.51 | 0.55 | 0.56 | 0.28 | 0.36 | 0.40 | 0.45 |
| FMM—elliptical | 0.71 | **1.00** | **1.00** | **1.00** | 0.23 | 0.78 | 0.87 | 0.90 |
| FMM—spherical | 0.71 | **1.00** | **1.00** | **1.00** | 0.23 | **0.84** | **0.90** | 0.92 |
| FMM—averaged | 0.71 | 0.89 | 0.93 | 0.96 | 0.23 | 0.32 | 0.39 | 0.48 |
| MCLUST-HC | 0.77 | 0.80 | 0.85 | 0.85 | 0.26 | 0.24 | 0.29 | 0.31 |
| MCLUST-FITTS | 0.77 | 0.90 | 0.95 | 0.95 | 0.26 | 0.27 | 0.38 | 0.35 |
| *B* | | | | | | | | |
| IMM—elliptical Auto | 19.0 | 6.8 | 6.4 | 6.2 | 15.4 | 5.8 | 6.0 | 6.0 |
| IMM—spherical Auto | 19.0 | 6.0 | 6.0 | 6.0 | 15.4 | 6.0 | 6.0 | 6.0 |
| IMM—averaged Auto | 19.0 | 19.4 | 18.4 | 17.4 | 15.4 | 16.2 | 17.0 | 16.2 |

Highest average indices for each situation are in boldface and underlined.

information, and Bayesian mixture-based models performed better than heuristic methods. The automatic clustering procedure based on the IMM with the elliptical variance structure again performed as well as any other method with a specified number of clusters (Table 4). The number of clusters created by this approach for the high-error situation was $(7, 7, 7, 22, 22)$ which reflected the uncertainty about whether sporadic genes should be placed in a single noisy cluster or they all belong to individual clusters. However, the method consistently identified the six distinct clusters in the data.

**Table 4.** Automatic IMM clustering results for datasets with outliers

| Noise level | Variance | RAND | No. of clusters |
|---|---|---|---|
| Low | Averaged | 0.78 | 13.4 |
| Low | Elliptical | 0.99 | 7.6 |
| Low | Spherical | 0.99 | 8 |
| High | Averaged | 0.51 | 16.6 |
| High | Elliptical | 0.97 | 13 |
| High | Spherical | 0.89 | 8.8 |

## 3.4 Yeast galactose data

We used the same subset of the Ideker *et al*. (2001) galactose dataset as described in Yeung *et al*. (2003). This set consists of 205 genes whose expression patterns reflect four functional categories in the Gene Ontology Consortium (Ashburner *et al*., 2000). From this dataset, we constructed subsets with $g = 1$, 2 and 3 replicates by systematically subsetting $g$ out of four replicated runs of the experiments. This has resulted in four datasets with a single observation for each experimental condition, six datasets with two experimental replicates per experimental condition and four

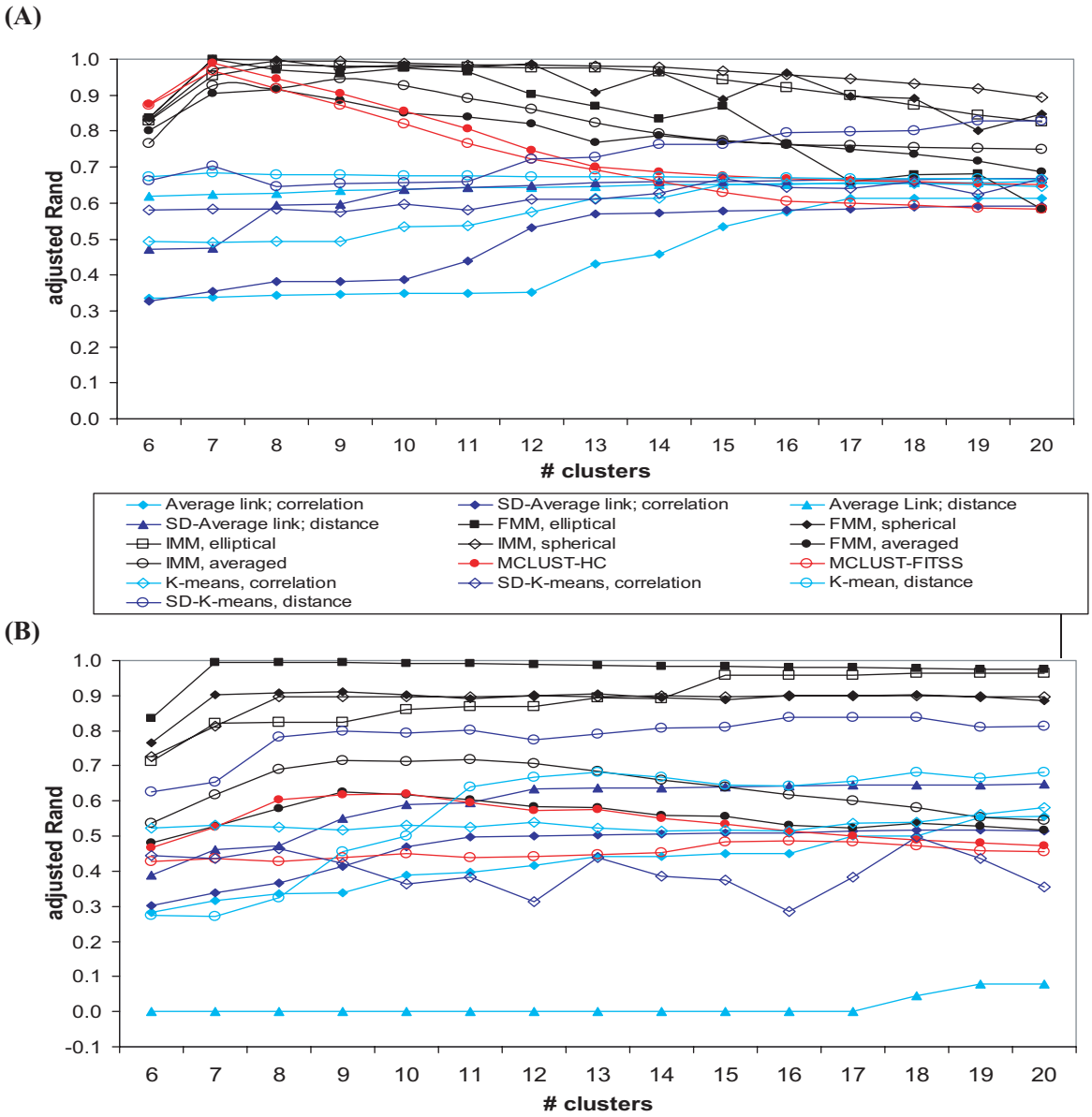


**Fig. 2.** Clustering data with sporadic genes. (**A**) Adjusted RAND indices under different number of clusters for low-noise simulated data and (**B**) high-noise simulated data.

datasets with three experimental replicates per experimental condition.

For each subset of the data, we ran four independent Gibbs samplers over a series of 'stopping points' and identified the optimal $\xi_{max}$ by following the algorithm described in the previous section. Since the analysis of simulated data indicated that the elliptical model is generally applicable even if the true covariance structure of experimental replicates is 'spherical', and due to a high computational cost of identifying optimal stopping points, we did not analyze this data under the 'spherical' model.

The IMM model for both 'elliptical' and averaged profiles variants have shown again the most consistent and precise results (Table 5). For each number of replicates, these two models performed as well as or better than any other clustering approach. As the number of replicates increased, the precision of these two approaches increased as well. Other methods that showed the same trend of consistently increasing precision for the increasing number of replicates were our versions of FMM models, MCLUST-based finite mixtures and distance-based complete linkage and variance adjusted hierarchical methods. However, all these methods showed

**Table 5.** Galactose dataset analysis

| Method | Number of replicates 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| *A* | | | | |
| Heuristic methods | | | | |
| Average link; correlation | **0.88** | 0.87 | 0.91 | 0.87 |
| Average link; distance | 0.67 | 0.90 | 0.92 | 0.86 |
| Complete link; correlation | 0.69 | 0.78 | 0.70 | 0.68 |
| Complete link; distance | 0.62 | 0.85 | 0.88 | 0.96 |
| K-means; correlation | 0.75 | 0.77 | 0.77 | 0.87 |
| K-means; distance | 0.86 | 0.92 | 0.91 | 0.86 |
| SD-adjusted heuristic methods | | | | |
| Average link; correlation | NA | 0.68 | 0.82 | 0.82 |
| Average link; distance | NA | 0.84 | 0.86 | 0.86 |
| Complete link; correlation | NA | 0.52 | 0.67 | 0.72 |
| Complete link; distance | NA | 0.83 | 0.96 | 0.97 |
| K-means; correlation | NA | 0.63 | 0.75 | 0.64 |
| K-means; distance | NA | 0.79 | 0.85 | 0.86 |
| Model-based methods | | | | |
| IMM—elliptical | 0.85 | 0.92 | **0.97** | **0.97** |
| IMM—elliptical Auto | 0.78 | **0.94** | 0.94 | 0.96 |
| IMM—averaged | 0.85 | 0.92 | 0.94 | **0.97** |
| IMM—averaged Auto | 0.78 | 0.93 | 0.96 | 0.96 |
| FMM—elliptical | 0.71 | 0.75 | 0.93 | **0.97** |
| IMM—averaged | 0.71 | 0.85 | 0.91 | **0.97** |
| MCLUST-HC | 0.68 | 0.90 | 0.96 | **0.97** |
| MCLUST-FITTS | 0.68 | 0.86 | 0.89 | **0.97** |
| *B* | | | | |
| IMM—elliptical Auto | 2.8 | 4.2 | 4.8 | 5.0 |
| IMM—averaged Auto | 2.8 | 4.7 | 4.5 | 5.0 |

(A) Average adjusted RAND indices. Highest average indices for the specific number of replicates are in boldface and underlined. (B) Number of clusters for IMM models with the automatic cluster creation.

significantly lower precision than IMM models in at least some scenarios. The IMM model with the automatic choice of the number of clusters in the data performed as well as any other method with the specified number of clusters. Interestingly, this model indicates that there are actually five clusters in the data. Potential biological consequences of such an implication need to be further investigated.

### 3.5 Correlated data, large datasets and different cluster sizes

We simulated additional datasets to assess the performance of Bayesian mixtures based clustering procedures on large datasets (10 000 'genes'), data with complex between-replicates covariance structure and datasets with highly unbalanced cluster sizes. Descriptions of the datasets and results are given in the Web supplement. In these tests, the elliptical IMM model with the automatic selection of the numbers of clusters performed better than any other clustering method we tested.

### 3.6 Computational complexity

Fitting mixture models via the Gibbs sampler is computationally expensive in terms of the CPU time. For a fixed number of profiles to be clustered, the computational complexity of the algorithm is approximately linear in the number of clusters. For example, on a 3 GHz Pentium workstation, for the 20-dimensional observation vectors, four replicates per experiment, each Gibbs sampler's cycle takes approximately 0.0015 s per gene in a 100-clusters FMM elliptical model and about 0.00013 s per gene in a five-clusters model. This results in about 100 min run-time to fit the 100-clusters model to the galactose dataset and about 9 min to fit the five-clusters model. For the run-times and memory requirements on various datasets and various models, see the Web supplement.

## 4 DISCUSSION

### 4.1 Choosing the right clustering procedure

After comparing various approaches on a series of simulated and real-world datasets, it is our impression that the 'elliptical' IMM is the best candidate for a universally recommended approach for clustering replicated microarray data. When compared with the corresponding FMM model and other clustering approaches, the IMM model performed equally well or better on a wide range of simulated data and the galactose dataset. Furthermore, IMM has a big advantage due to its ability to automatically choose the appropriate number of clusters. In many situations, the IMM with the automatic selection of the number of clusters outperformed FMM models and all heuristic methods that used the information about the correct number of clusters in the data. This was somewhat surprising, but it can be explained by the fact that although there exists an underlying model with the correct number of clusters, the data might be best explained with a different number of clusters, and forcing the pre-specified number of clusters can distort resulting clusters. Furthermore, in a 'fuzzy' situation when more than one model offers a reasonable fit to data, averaging over different models might improve the overall result. When compared with the 'averaged' and the 'spherical' variance structure, the 'elliptical' model generally performed at least as well or better than any other model. The only exception was the simulated 'spherical' wave data, in which the spherical model performed slightly better. Altogether, when in doubt about the correct covariance structure, the 'elliptical' model seems to be a natural choice among currently available models.

### 4.2 Importance of replicating experiments

Results of our analysis clearly demonstrate the importance of experimental replicates in the cluster analysis of microarray data. Even two experimental replicates can significantly improve the precision of clustering results, both in terms of adjusted RAND indices as well as in terms of the number of clusters inferred by the statistical model. The Bayesian

mixture models with 'elliptical' variance structure were particularly efficient in using additional information obtained by replicating the experiment. Improvements in the precision were the most dramatic in the 'high-noise' simulated data. Generally, when comparing the performance of the models with the 'elliptical' error model with the simple approach of clustering averaged expression profiles, more dramatic differences have been seen in the simulation study than in the analysis of galactose data. The reason for such differences probably lies in the fact that our simulated data represented the extreme situation when all variability in the data is induced by the between-replicates variations, which is likely to emphasize the importance of modeling explicitly this variability instead of assuming a uniform variance as the analysis of averaged profiles implicitly assumes. This also suggests that the importance of experimental replicates as well as the importance of the precise modeling of between-replicates variability will be more important in situations where this variability is expected to be high, such as when biological replicates are used (Hatfield *et al*., 2003).

### 4.3 The Gibbs sampler, statistical significance and future work

One of the conceptually appealing features of the IMM approach is that it facilitates assessment of the statistical significance of various clustering features after incorporating uncertainties related to choosing the correct number of clusters. However, for such assessments to be unbiased, the Gibbs sampler needs to have good 'mixing' properties. That is, it needs to be capable of generating samples from different regions of the probability space in proportion to their posterior probabilities within the specified sample size. When the posterior probability distribution is multi-modal, as was the case in many of our datasets, this will generally not be the case. While our heuristic modifications did manage to correct the mixing properties of the sampler with respect to obtaining the optimal clustering, the distribution approximated by the sampler is not any more described by our hierarchical model. Improving mixing properties of the sampler remains one of the priorities in the development of IMM models.

Finally, summarizing the posterior distribution generated by the Gibbs sampler is not trivial in the context of Bayesian mixtures mainly due to the 'label switching' problem (Celeux *et al*., 2000). Reducing the full posterior distribution of clusterings to pair-wise probabilities of co-expression as described here effectively circumvents the issue of 'label switching'. 'Similarity matrices' consisting of pair-wise posterior probabilities can be used to both create clusters and compare posterior clustering distributions from different runs of the Gibbs sampler. Still, reducing the full posterior distribution of clusterings to pair-wise probabilities of co-expression will result in loss of information, and the applicability of the re-labeling approaches (Celeux *et al*., 2000) in the context of infinite mixtures seem to be worth exploring.

## ACKNOWLEDGEMENTS

We would like to thank Vesteinn Thorsson for providing us with the yeast galactose data and to thank Siva Sivaganesan for valuable discussions relating to mixing properties of the Gibbs sampler. We would also like to acknowledge the following publicly available software packages: MCLUST (Fraley and Raftery, 2002) and KNNimpute (Troyanskaya *et al*., 2001). R.E.B. and K.Y are supported by NIH-NIDDK grant 5U24DK058813-02. R.E.B. is also supported by NIH-NIAID grants 5P01 AI052106-02, 1R21AI052028-01 and 1U54AI057141-01, NIH-NIEHA grant 1U19ES011387-02, NIH-NIDA grant 1 P30 DA015625-01, NIH-NHLBI grants 5R01HL072370-02 and 1P50HL073996-01. M.M is supported by the NHGRI grant 1R21HG002849-01 and NIEHS grants 2P30 ES06096-11 and ES04908-12.

## REFERENCES

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

Celeux,G., Hurn,M. and Robert,C.P. (2000) Computational and inferential difficulties with mixture posterior distributions. *JASA*, **95**, 957–970.

Dougherty,E.R., Barrera,J., Brun,M., Kim,S., Cesar,R.M., Chen,Y., Bittner,M. and Trent,J.M. (2002) Inference from clustering with application to gene-expression microarrays. *J. Comput. Biol.*, **9**, 105–126.

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

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

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

Hatfield,G.W., Hung,S.P. and Baldi,P. (2003) Differential analysis of DNA microarray gene expression data. *Mol. Microbiol.*, **47**, 871–877.

Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.

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

Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ideker,T., Thorsson,V., Ranish,J.A., Christmas,R., Buhler,J., Eng,J.K., Bumgarner,R., Goodlett,D.R., Aebersold,R. and Hood,L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Medvedovic,M. (2000) Clustering multinomial observations via finite and infinite mixture models and MCMC algorithms. *Proceedings of the Joint Statistical Meeting 2000: Statistical Computing Section*, August 13–17, Indianapolis, IN, pp. 48–51.

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

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

Rasmussen,C.A. (2000) The infinite Gaussian mixture model. *Adv. Neural Inform. Process. Syst.*, **12**, 554–560.

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. *Nat. Genet.*, **22**, 281–285.

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

Yeung,K.Y., Fraley,C., Murua,A., Raftery,A.E. and Ruzzo,W.L. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Yeung,K.Y., Medvedovic,M. and Bumgarner,R.E. (2003) Clustering gene expression data with repeated measurements. *Genome Biol.*, **4**, R34.