Clustering microarray gene expression data using weighted Chinese restaurant process

Zhaohui Qin

Center for Statistical Genetics, Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029;

Running title: *Chinese restaurant cluster for gene expression data.*

Dr. 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, Tel: (734) 763-5965, Fax: (734) 763-2215, E-mail: qin@umich.edu.

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 inverted and time-shifted to be clustered together. Other notable features including automatic handling of missing data, quantitative measures of cluster strength and assignment confidence. Synthetic and real microarray gene expression datasets are 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. Due 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. 1999).

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 is 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 (McLachlan et al. 2002; Yeung et al. 2001). 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 (2002) 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, she will be seated according to the current seating arrangement of all previous customers. In this method, cluster number determination 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 manuscript, we devise a different model-based clustering algorithm based on weighted CRP (Lo 2005) in which customers tend to be attracted to tables that contain “similar customers”. The predictive updating technique is applied to integrate out nuisance parameters, which greatly improves the efficiency of the computation-intensive Gibbs sampler procedure. The marginal likelihood is calculated during the iteration. The cluster assignments 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, due 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 inverted and/or time-shifted. See Figure 1 for illustrations of these relationships. These non-standard relationships will be missed by current clustering tools. It is of great interest if we can identify genes showing diverse 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 biological 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 a model 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_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, M$. Although these experiments may be related (for example, 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. This assumption can be dropped by adopting a multivariate normal distribution with non-zero covariance. Let $E = (E(1), \ldots, E(N))$ be the cluster indicator variable, $E(i) = k, \quad 1 \le k \le K$ denotes that the *i*th gene was assigned to the *k*th cluster, $1 \le i \le N$. We use $| E |$ 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 the follows:

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

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

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

Here β_0 , a and b are all assumed known. Cluster indicator variable E is assumed to follow a Dirichlet process. The detailed information about these prior distributions can be found in the online supplementary text.

The clustering procedure is modeled after a weighted Chinese restaurant seating scheme. For each customer (gene), we need to determine which table (cluster) fits her best. Assume others have already been seated, we calculate the likelihood of fitting this customer to each of the existing table. Then table assignment can be regarded as sampling from a multinomial distribution with probabilities proportional to the conditional posterior probabilities:

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

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 $\left(E(1),...,E(N),\left(\beta_{1j},\sigma_{1j}^2\right)_{j=1}^M,...,\left(\beta_{Kj},\sigma_{Kj}^2\right)_{j=1}^M\right)$ 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. The predictive updating technique (Liu 1994; Chen and Liu 1996) can be applied to improve the efficiency of our algorithm. That is, integrate out unwanted nuisance parameters analytically from the likelihood function. For each cluster, after the incorporation of prior distributions, we have:

$$\iint \prod_{E(i)=k} p\left(x_{ij} \mid \beta_{kj},\sigma_{kj}^2\right) p\left(\beta_{kj} \mid \beta_0,\sigma_{kj}^2\right) p\left(\sigma_{kj}^2\right) d\beta_{kj} d\sigma_{kj}^2$$

$$=\frac{b^a}{\Gamma(a)}\frac{(2\pi)^{-\frac{n_k}{2}}}{\sqrt{n_k+1}}\frac{\Gamma\left(\frac{n_k}{2}+a\right)}{\left(b+\frac{1}{2}\left(\sum_{E(i)=k}x_{ij}^2+\beta_0^2-\frac{\left(\sum_{E(i)=k}x_{ij}+\beta_0\right)^2}{n_k+1}\right)\right)^{\frac{n_k}{2}+a}}.$$

This formula will be used to calculate the likelihood ratio and the assignment probabilities Q_l. Details can be found in the online supplementary text.

Algorithm

A Gibbs sampler is implemented to carry out the iterative weighted CRP, with predictive updating scheme to improve its efficiency. The clustering procedure can be summarized as follows:

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

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:

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

    $K = \mid E \mid$ and $l = 0$ indicates that this gene is standing alone by itself, $l > 0$ means assignment to an existing cluster.

    b. Assign gene *i* to the *K* + 1 possible clusters according to probabilities $Q_l, l = 0,1,\ldots,K$. Update indicator variable *E*(*i*) based on the assignment, as well as total number of cluster, *K*, if a new cluster is formed or an existing cluster with gene *i* as the single member is removed.

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

The details of implementing this algorithm can be found in the online supplementary text.

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 simple positive and simultaneous correlation relationship, like most of 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 inverted or time-shifted. 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 inverted or time-shifted version of those in a cluster, then transforming the original profile by inversion or time-shifting will produce profiles that can be clustered into their right clusters using the aforementioned clustering algorithm. For that reason, we added an extra model 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 cluster to find the best fit (please see Figure 2 for illustration). For example, if we choose *s* = 3, then profiles $(x_{i1}, x_{i2}, \cdots, x_{iM-2})$, $(x_{i2}, x_{i3}, \cdots, x_{iM-1})$ and $(x_{i3}, x_{i4}, \cdots, x_{iM})$ as well as $(-x_{i1}, -x_{i2}, \cdots, -x_{iM-2})$, $(-x_{i2}, -x_{i3}, \cdots, -x_{iM-1})$ and $(-x_{i3}, -x_{i4}, \cdots, -x_{iM})$ will be compared to 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 2*sK* possible outcomes.

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. 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 is 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 supports 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.

Posterior probability measurement of each cluster assignments

Due 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, at the end of the iteration, we can calculate the posterior probability of each gene belonging to its assigned cluster (Q_l 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.

Cluster strength

After clustering analysis has been performed on a particular dataset, it is of great interest to determine which clusters are more statistical significant than others since such information may lead to further biological insights. In a model-based setting, the significance can be evaluated by calculating a so called Bayes ratio for each cluster to indicate how close the members of this cluster are. 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 the first n_1 genes belong to the same cluster. Incorporating priors, the Bayes ratio can be described as follows:

$$\frac{\prod_{j=1}^{M}\int P(x_{1j},x_{2j},...,x_{n_1j}\mid\mu_j,\sigma_j^2)P(\mu_j,\sigma_j^2)d\mu_j d\sigma_j^2}{\prod_{j=1}^{M}\prod_{i=1}^{n_1}\int P(x_{ij}\mid\mu_{ij},\sigma_{ij}^2)P(\mu_{ij},\sigma_{ij}^2)d\mu_{ij}d\sigma_{ij}^2}.$$

The final strength 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, and therefore more likely to be biologically related. Based on these quantities, we are able to rank the clusters generated, and help investigators to triage them based on this confidence measure. We can also set a threshold; and retain only clusters that meet certain significance level. Similar measures have been proposed in studies on clustering regulatory motifs using model-based clustering approaches BMC (Qin et al. 2003) and PHYLOCLUS (Jensen et al. 2005). Such ranking is not available in most of the current clustering approaches devised for microarray gene expression data.

Results

Clustering accuracy

To evaluate 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 Arable 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 text.

Datasets

To test the performance of our new model-based clustering algorithm, we tested it on two different datasets: a synthetic dataset and the galactose metabolism dataset (Idekar et al. 2001).

Synthetic dataset

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 $x_{ij} = \left(k/2\right)\sin(\pi jk/10 - \pi k/4) + \varepsilon$,

$j = 1, 2, ..., 20$ and $k = 1, 2, 3$. One cluster display 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: $a \sim Uniform(-1,1)$. A random perturbation term ε ($\varepsilon \sim N(0, 0.5^2)$) is added to each data point x_{ij} to account for the noise associated with gene expression levels observed. 100 such datasets were generated. The trace plot of a sample simulated dataset can be found in Figure 3 and S1. Clustering results obtained from running CRC are summarized in Table 1. For simple datasets like this, it is not surprising that CRC performs almost perfectly in terms of both clustering accuracy and cluster number estimate.

The presence of complex correlation relationships among expression profiles such as inverted and/or time-shifted complicated the clustering problem. We mimic such situations to investigate how well does our algorithm 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 the two types of complex expression profiles are 10% each and they may overlap. 100 such datasets were generated, trace plots of a sample dataset is shown in Figure 3 and S1. The clustering results using CRC were summarized in Table 1. Where 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, we also run CRC on the same datasets with the model selection option turned off (such that it won't be able to identify genes that displaying non-synexpression correlations with a cluster), which resulted in much higher clustering error (the average ARI reduce to 0.852). For hierarchical clustering (*hclust* function in R), even with the correct number of clusters specified, the average ARI is only 0.574. The superior performance of CRC in these datasets suggests that both the weighted Chinese restaurant seating scheme and the added model 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 set consists of 205 genes whose expression patterns reflect four functional categories in the gene Ontology Consortium (Ashburner et al. 2000). 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 was downloaded from Dr. Yeung's website http://expression.microslu.washington.edu/expression/kayee/medvedovic2003/medvedovic_bioinf2003.html. The original dataset contains about 8% of missing data, these missing values has been imputed using KNN impute (Troyanskaya et al. 2001) in the version we downloaded. The average expression profiles are illustrated in Figure 4 and S2.

This dataset contain results from four replicated runs, CRC was tested with and without replicates to assess whether its performance improves in the presence of replicates. For test with single replicate, we run CRC 25 times on each set of the four replicates. For test with replicates, we simply take average of the four replicates and perform clustering 100 times on the averaged expression profiles using CRC. The final clustering accuracies in both scenarios were obtained by taking average of the performance measures from the 100 total runs. The results are summarized in Table 2 and a cluster-specific trace plot based on a sample result is shown in Figure 4 and S2. Overall, CRC performs well on this dataset. Replicates did significantly improve the clustering accuracy. The average ARI is 0.968 when all four replicates were considered and 0.785 when single replicate was used. In contrast, the best performance achieved by hierarchical clustering is 0.863 using all replicates and 0.692 using single replicate. Another encouraging result is that, the cluster number estimates provided by CRC are quite accurate for this real dataset.

Since GIMM performed favorably comparing to other clustering approaches in a previous study (Yeung et al. 2003), we compared the performance of CRC with that of GIMM. Default settings of GIMM are used on this dataset with 10,000 iterations. Since this method *per se* does not suggest the “right” number of clusters to use (M. Medvedovic, personal communication),we evaluated using several different cluster sizes. When all replicates are considered, the clustering accuracy is 0.847 when the correct cluster size—4 was used. And its clustering accuracy improves to 0.953 and 0.950 when the cluster size is specified at 5 and 6 respectively, but is still slightly lower than what CRC produced where cluster size is automatically inferred from the data. When there is no replicate, the best performance GIMM achieved under different cluster sizes ranged from 0.56 to 0.67 for the four sets of replicates.

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. When there is single replicate, CRC was run on each of the four replicate datasets 25 times and the average ARI from the 100 runs was reported. On average, there are 146 (71.2%) genes contain missing data in at least one of the 20 experiments. The average missing proportions in these four sets are 7.8%. The clustering results shown in Table 2 indicate that there is no significant difference in terms of clustering accuracy when there is moderate amount of missing data present. When there are replicates, one 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%. A more strict rule require all four replicates to be present otherwise it will be called missing. There are 198 (96.6%) genes contain missing data under this rule, and the overall proportion of missing data is

24%. This is clearly a less sensible way to summarize replicated data. The sole purpose of testing this scheme is to assess the performance of CRC on datasets with large proportion of missing data. From the results shown in Table 2, there is very little performance difference in terms of clustering accuracy using the first imputation strategy. The clustering accuracy did experiencing a moderate drop when using the second strategy, but keep in mind that the proportion of missing data using the second strategy is much higher than before thus their results are not directly comparable. What the results suggest is that moderate proportion of missing data has little impact on clustering performance of CRC.

Discussion

In summary, we implemented a model-based clustering strategy that is based on the weighted Chinese restaurant process for clustering microarray gene expression data. This algorithm is able to clustering genes and inferring 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. In addition, this algorithm is able to recognize genes that display complex correlation relationships such as time-delayed 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. 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 model selection step during the iterative process is especially attractive since it enables us to identify complicated correlation relationships other than synexpression. Since genes participated in the same regulatory network may display complicated relationships and will be missed by most of the currently available clustering tools. By adding this step, we will be able to achieve better understanding of the complex underlying biological processes, and to generate more sensible and accurate hypotheses.

Another contribution is that we demonstrated that the marginal likelihood can be a good indicator of the performance of our clustering procedure based on our studies using simulated data and real data. 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 about 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 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. The current algorithm can be easily extended to take the correlation structure into account which can potentially further improve the performance of this clustering algorithm when experiments conditions are correlated or replicated experiments were performed. We plan to give users options to choose between these two models.

Microarray technology is not precise; the expression profile from a single gene is often not informative due to experimental noise. 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. Similar reason is behind the favorable performance of model-based algorithm for clustering putative DNA binding motifs (Qin et al. 2003). We believe that improvement in model-based clustering techniques will enhance our ability to reveal more knowledge buried underneath massive amount of gene expression data.

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, the Gibbs sampler converged fairly rapidly (please see Figure S3 for a likelihood trace plot). For the galactose dataset with 205 genes and 20 experiments, it only takes CRC about 15 seconds to run on a SUN opteron server under the default settings, whereas the same dataset takes GIMM about 4 minutes to finish.

Acknowledgement

We thank Drs. Michael Elliott, Debashis Ghosh for fruitful discussion and insightful comments on an earlier draft of this manuscript. We thank Dr. Mario Medvedovic for helping with the GIMM software, and Dr. Ka Yee Liang for sharing the yeast galactose data.

References

1 Aldous D: Exchangeability and related topics. Springer-Verlag, 1985.

2 Ashburner M, Ball CA, Blake JA *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**: 25-29.

3 Banfield JD, Raftery A. E.: Model-based Gaussian and non-Gaussian clustering. *Biometrics* 1993; **49**: 803-821.

4 Chen R, Liu JS: Predictive Updating Methods With Application to Bayesian Classification. *Journal of the Royal Statistical Society, Series B* 1996; **58**: 397-415.

5 Cho RJ, Campbell MJ, Winzeler EA *et al*: A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 1998; **2**: 65-73.

6 Dempster AP, Laird NM, Rubin DB: Maximum Likelihood From Incomplete Data Via EM Algorithm. *Journal of the Royal Statistical Society, Series B* 1977; **39**: 1-38.

7 Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998; **95**: 14863-14868.

8 Ferguson TS: A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1973; **1**: 209-230.

9 Fraley C, Raftery, AE: Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 2002; **97**.

10 Gelfand AE, Smith AFM: Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**: 398-409.

11 Gelman A, Carlin JB, Stern HS, Rubin DB: Bayesian data analysis. London, Chapman \& Hall, 1995.

12 Hubert L AP: Comparing partitions. *Journal of Classification* 1985; **2**: 193-218.

13 Hughes JD, Estep PW, Tavazoie S, Church GM: Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* 2000; **296**: 1205-1214.

14 Ideker T, Thorsson V, Ranish JA *et al*: Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 2001; **292**: 929-934.

15 Jensen ST, Shen L, Liu JS: Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics* 2005; **21**: 3832-3839.

16 Liu J: Monte Carlo Strategies in scientific computing. New York, Springer-Verlag, 2001.

17 Liu JS, Wong WH, Kong A: Covariance Structure of the Gibbs Sampler With Applications to the Comparisons of Estimators and Augmentation Schemes. *Biometrika* 1994; **81**: 27-40.

18 Lo AY: Weighted Chinese restaurant processes. *COSMOS* 2005; **1**: 59-63.

19 Lockhart DJ, Dong H, Byrne MC *et al*: Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996; **14**: 1675-1680.

20 McLachlan GJ, Basford, K. E.: Mixture models: inference and applications to clustering. New York, Marcel Dekker, 1988.

21 McLachlan GJ, Bean RW, Peel D: A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 2002; **18**: 413-422.

22 Medvedovic M, Sivaganesan S: Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* 2002; **18**: 1194-1206.

23 Medvedovic M, Yeung KY, Bumgarner RE: Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics* 2004; **20**: 1222-1232.

24 Milligan GW, Cooper, M. C.: A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavior Research* 1986; **21**: 441-458.

25 Neal R: Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 2000; **9**: 249-265.

26 Pitman J: Some developments of the Blackwell-MacQueen urn scheme. Hayward, California, IMS, 1996.

27 Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M: Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression

profiles identifies new, biologically relevant interactions. *J Mol Biol* 2001; **314**: 1053-1066.

28 Qin ZS, McCue LA, Thompson W, Mayerhofer L, Lawrence CE, Liu JS: Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol* 2003; **21**: 435-439.

29 Rand WM: Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 1971; **66**.

30 Rusmussen C. The infinite gaussian mixture model.: Advances in Neural information Processing Systems, 2000, vol 12, pp 554-560.

31 Schena M, Shalon D, Davis RW, Brown PO: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995; **270**: 467-470.

32 schwarz G: Estimating the dimension of a model. *The Annals of Statistics* 1978; **6**: 461-464.

33 Spellman PT, Sherlock G, Zhang MQ *et al*: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 1998; **9**: 3273-3297.

34 Tamayo P, Slonim D, Mesirov J *et al*: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* 1999; **96**: 2907-2912.

35 Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: Systematic determination of genetic network architecture. *Nat Genet* 1999; **22**: 281-285.

36 Troyanskaya O, Cantor M, Sherlock G *et al*: Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; **17**: 520-525.

37 Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: Model-based clustering and data transformations for gene expression data. *Bioinformatics* 2001; **17**: 977-987.

38 Yeung KY, Medvedovic M, Bumgarner RE: Clustering gene-expression data with repeated measurements. *Genome Biol* 2003; **4**: R34.

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

Figure 2. Illustration of the model 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.

Figure 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 inverted and/or time-shifted relationships. A different version of these two plots which display the five clusters using different colors is shown in Figure S1 in the online supplementary text.

Figure 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 Figure S2 in the online supplementary text.

Table 1. Performance of CRC on synthetic datasets.

	Algorithm	Clustering accuracy		Cluster# estimate (true =5)	
Simple dataset	CRC simple	0.994	0.036	4.97	0.17
	CRC full	0.991	0.043	4.96	0.20
Complex dataset	CRC simple	0.852	0.027	8.77	0.72
	CRC full	0.972	0.072	4.87	0.34

Table 2. Performance of CRC on yeast galactose datasets.

	Datasets	Clustering accuracy		Cluster# estimate (true = 4)	
Single replicate	No missing	0.785	0.033	5.49	0.50
	Missing	0.788	0.044	5.35	0.56
All data (average)	No missing	0.967	0.004	4.00	0.00
	Missing (low)	0.955	0.000	4.38	0.48
	Missing (high)	0.851	0.049	3.94	0.47

Figure 1.

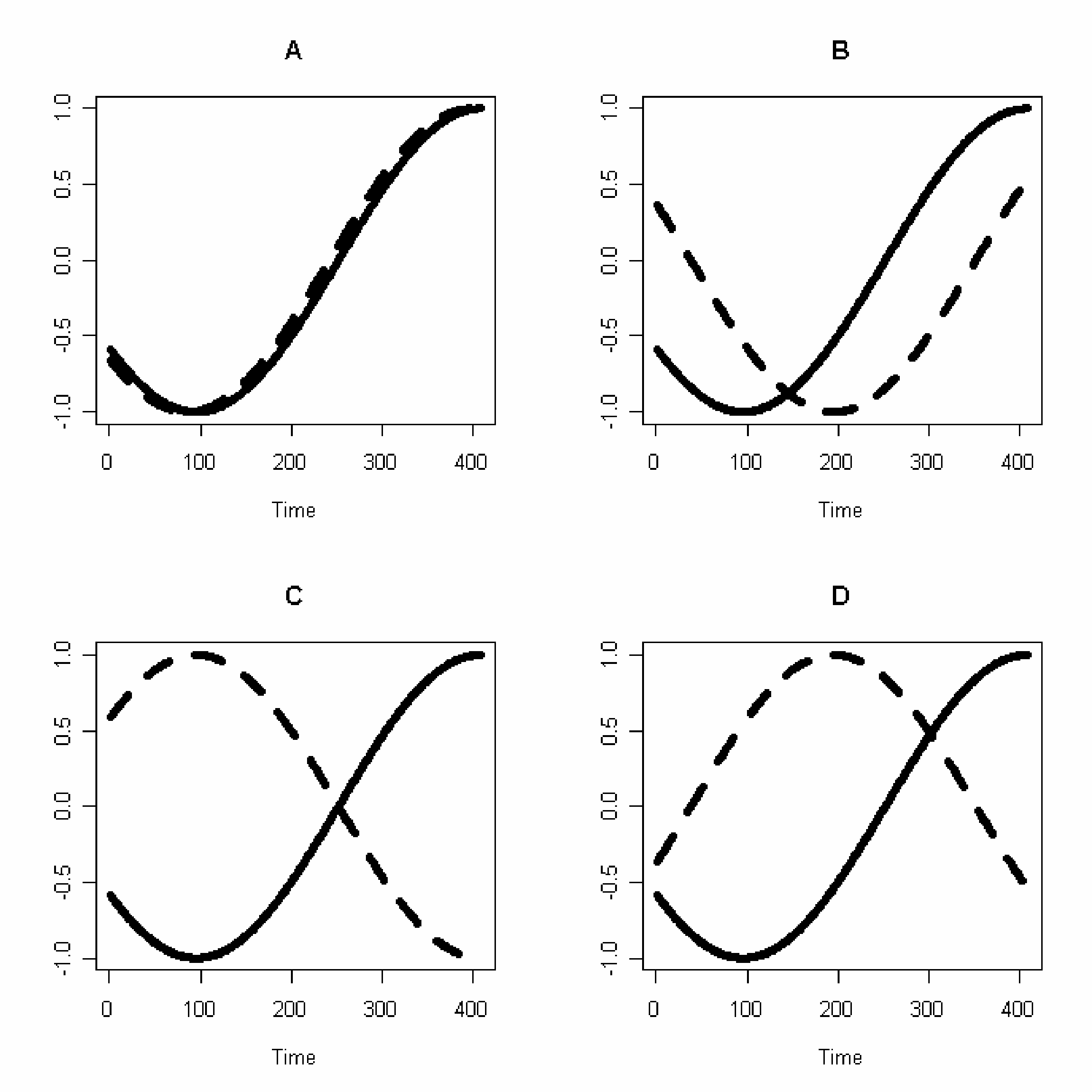
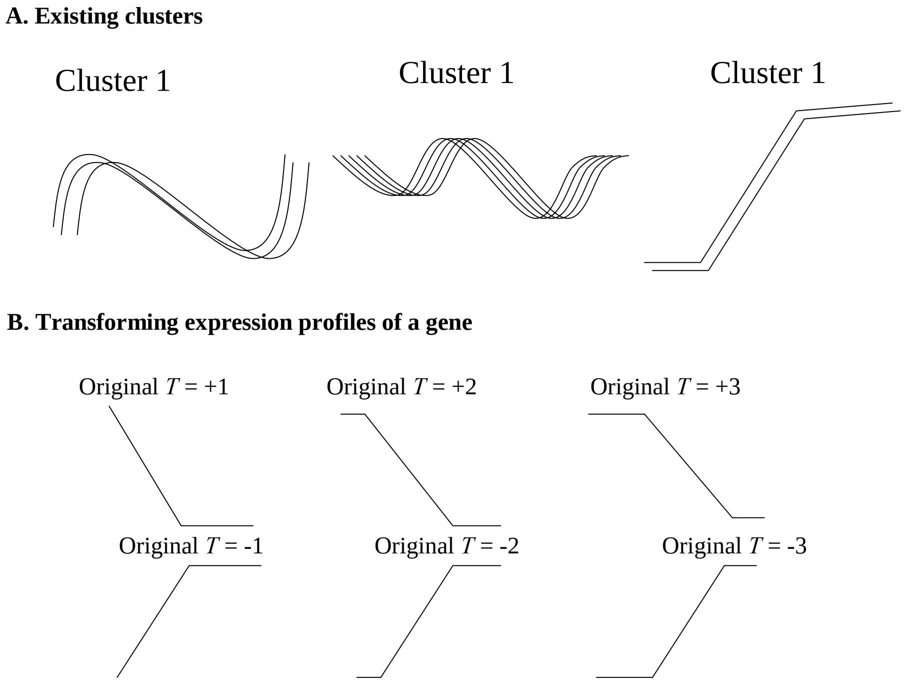


Figure 2.



A. Existing clusters

B. Transforming expression profiles of a gene

Figure 3.

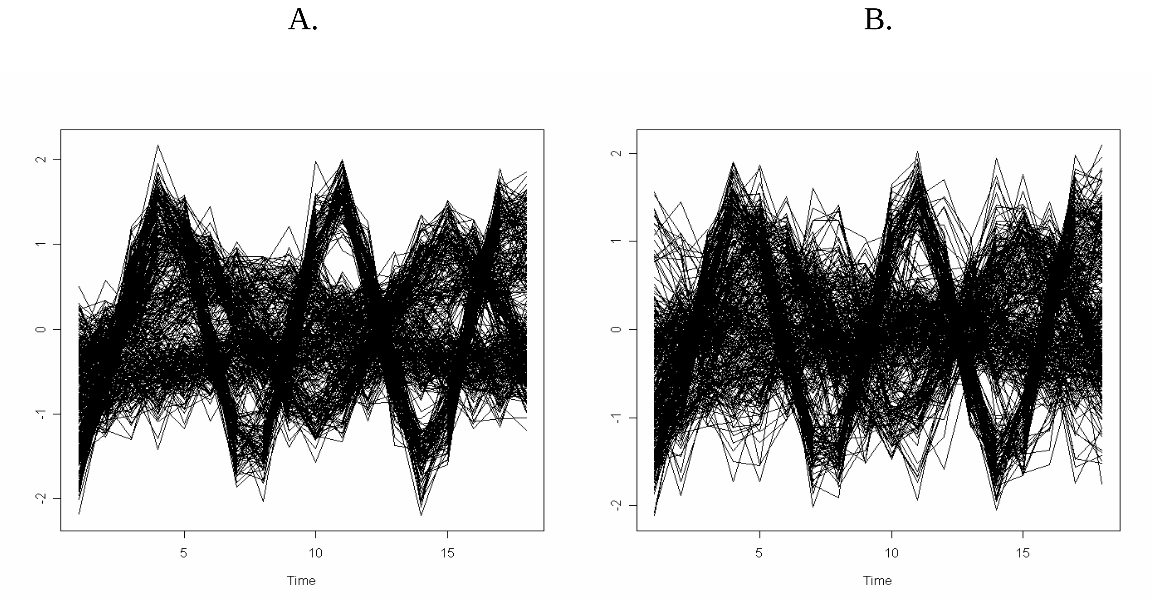
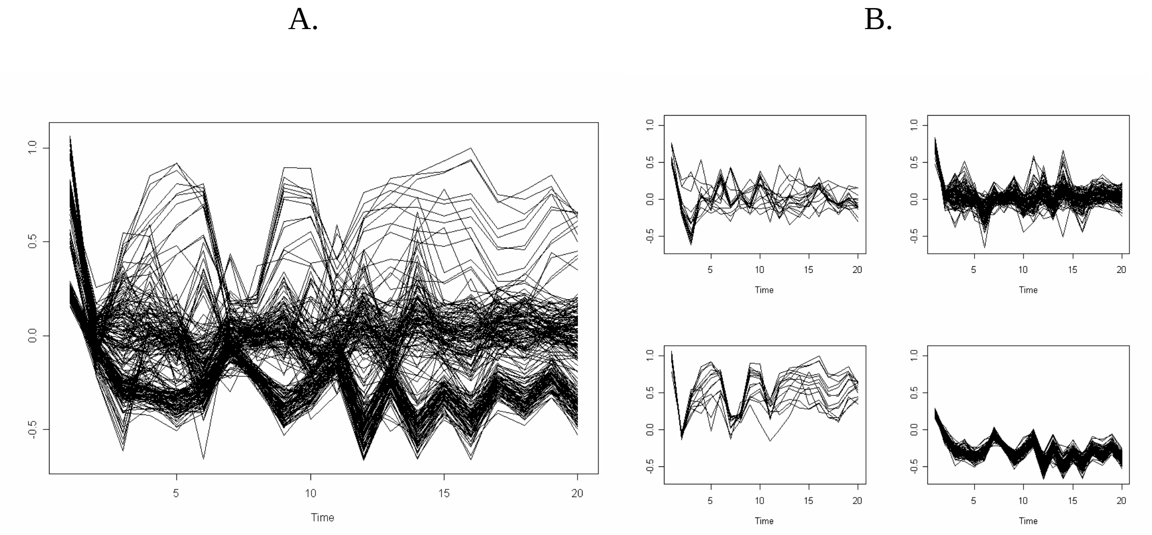


Figure 4.


A.
B.
Time