Supporting Text

Clustering microarray gene expression data using weighted Chinese restaurant process

Notation:
N – number of genes,
M – number of experiments,
K – number of clusters, $K = |E|$,
$X=\{x_{ij},\ i=1,...,N,\ j=1,...,M\}$: expression profile,
$E=\{E(i),\ i=1,...,N\}$: indicator of cluster membership.

Statistical model:
We assumed that genes fall into the same cluster share the same set of normal distributions:

$$x_{ij} \sim N(\beta_{kj},\sigma_{kj}^2)\ ,$$

where gene $i \in$ cluster k , $k=1,...,K$ and $j=1,...,M$.

The likelihood function is:

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

To gain efficiency in computation, we apply the predictive updating technique, integrating out parameters β_{kj} and σ_{kj}^2 analytically to obtain the marginal likelihood (Halpern 973, Elliott and Little 2000):

$$P(X\mid E)=\prod_{k=1}^{|E|}\prod_{j=1}^{M}\iint\prod_{E(i)=k}P\left(x_{ij}\mid\beta_{kj},\sigma_{kj}^2\right)p\left(\beta_{kj}\mid\beta_0,\sigma_{kj}^2\right)p\left(\sigma_{kj}^2\right)d\beta_{kj}d\sigma_{kj}^2$$

$$=\prod_{k=1}^{|E|}\prod_{j=1}^{M}\iint\left(\frac{1}{2\pi\sigma_{kj}^2}\right)^{-\frac{n_k}{2}}e^{-\frac{1}{2\sigma_{kj}^2}\sum_{E(i)=k}\left(x_{ij}-\beta_{kj}\right)^2}\left(\frac{1}{2\pi\sigma_{kj}^2}\right)^{-\frac{1}{2}}e^{-\frac{1}{2\sigma_{kj}^2}\left(\beta_{kj}-\beta_0\right)^2}\frac{b^a}{\Gamma(a)}\left(\sigma_{kj}^2\right)^{-(a+1)}e^{-\frac{b}{\sigma_{kj}^2}}d\beta_{kj}d\sigma_{kj}^2$$

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

The likelihood ratio for a gene belonging to a particular cluster *k* is the follows:

$$\frac{P(X\,|\,E(1),\ldots,E(i-1),E(i)=k,E(i+1),\ldots,E(N))}{P(X\,|\,E(1),\ldots,E(i-1),E(i)=0,E(i+1),\ldots,E(N))}=$$

$$=\frac{\Gamma(a)\sqrt{2n}}{b^{a}\sqrt{n+1}}\frac{\Gamma(a+n/2)\left(b+\frac{1}{2}\left(\sum_{i=2}^{n}x_i^2+\beta_0^2-\frac{\left(\sum_{i+2}^{n}x_i+\beta_0\right)^2}{n}\right)\right)^{a+(n-1)/2}\left(b+\frac{\left(x_1-\beta_0\right)^2}{2}\right)^{a+1/2}}{\Gamma(a+(n-1)/2)\Gamma(a+1/2)\left(b+\frac{1}{2}\left(\sum_{i=1}^{n}x_i^2+\beta_0^2-\frac{\left(\sum_{i=1}^{n}x_i+\beta_0\right)^2}{n+1}\right)\right)^{a+n/2}}$$

Prior distributions:

For each cluster and each experiment, the same conjugate prior distributions were assumed for the parameters β_{kj} and σ_{kj}^{2} $k=1,\ldots,K$ and $j=1,\ldots,M$:

$$P(\beta_{kj}\,|\,\sigma_{kj}^2)\sim N(\beta_0,\sigma_{kj}^2),$$

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

The parameters in the prior distributions, β_0 , *a* and *b*, are all assumed known. For simplicity, β_0 is taken to be the overall mean of the expression profiles: $\overline{x}=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}x_{ij}$, and let *a* and *b* to be 1 and $2s^2(x)$, where $s^2(x)$ represents the sample variance of x_{ij}'s $s^2(x)=\frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(x_{ij}-\overline{x})^2$. It is admitted that strictly speaking, these prior distributions are data dependent. However, expression profiles for an experiment within a cluster only represent a small fraction of the overall data, hence the use of overall mean and variance to help determine the prior distributions for an experiment within a cluster seems reasonable.

The prior distribution for cluster indicator *E*(*i*) is defined conditional on *E*(-*i*) (the indicator variables for all genes except *i*), which is a Dirichlet process:

$$P(E(i)=j\,|\,E(-i))=\begin{cases}\dfrac{n_{-i,k}}{N-1+\alpha} & \text{for } j=0,\\[2ex] \dfrac{\alpha}{N-1+\alpha} & \text{for } j>0.\end{cases}$$

$n_{-i,k}$ is the size of the *k*th cluster for the $N-1$ genes excluding *i*. α can be treated as a tuning parameter specified *a priori*. α is fixed at 1 throughout this study. This prior distribution is widely used in Bayesian nonparametric models and mixture models.

Implementation:

A C++ program named CRC (Chinese restaurant cluster) has been developed based on the algorithm proposed. The program and a detailed manual can be downloaded from the website: http://www.sph.umich.edu/csg/qin/CRC/.

By default, CRC runs 10 Markov chains in parallel, each with 20 cycles of iterations performed on all genes. At the beginning of each iteration, the initial number of clusters is specified as $\sqrt{N}$, which corresponds to the expected number of clusters formed by N objects following the Chinese restaurant process. Genes were randomly assigned to these clusters. During iteration, cluster assignment for each gene is updated conditional on all other genes' assignment. Each assignment can be regarded as a sample from a multinomial distribution. Marginal likelihood is calculated after each cluster assignment update. The one with the highest likelihood is retained as the final answer. At the end of the iteration, the posterior probability of the final cluster assignment for each gene is calculated, and cluster strengths are calculated to indicate the significance of each cluster.

Measuring inconsistencies between different clustering results

This part is adapted and summarized from the document *Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data"* (Yeung and Ruzzo, 2001). For details, please refer to their original document, which can be found at http://faculty.washington.edu/kayee/pca/supp.pdf.

There are numerous ways to compare the agreement between two different partitions among objects in clustering analysis. Milligan and Cooper recommended ARI among many different indices after extensive empirical comparisons and evaluations with different cluster sizes (Milligan and Cooper 1986). This is also the measurement of choice in Yeung et al. (2001, 2003) and Medvedovic et al. (2004).

Assume we are interested in comparing two different partitions of N genes: $P=\left(p_1,p_2,...,p_{K_p}\right)$ and $Q=\left(q_1,q_2,...,q_{K_q}\right)$. K_p and K_q are numbers of clusters formed by these two partitions. The agreement can be measured by the Adjusted Rand Index (ARI), which evaluates the degree of overlap between these two partitions. Let n_{ij} represents the common elements in the ith cluster of partition 1 and the jth cluster in partition 2. The two partitions can be summarized by the following table:

Clusters	1	2	…	q	Sums
1	n_{11}	n_{12}	…	n_{1q}	$n_{1.}$
2	n_{21}	n_{22}	…	n_{2q}	$n_{2.}$
⋮	⋮	⋮	⋱	⋮	⋮
p	n_{p1}	n_{p2}	…	n_{pq}	$n_{p.}$
Sums	$n_{.1}$	$n_{.2}$	…	$n_{.q}$	$n_{..} = N$

The ARI is calculated by the following formula:

$$\frac{\sum_{i,j}\binom{n_{ij}}{2}-\left[\sum_{i}\binom{n_{i.}}{2}\sum_{j}\binom{n_{.j}}{2}\right]\Big/\binom{N}{2}}{\frac{1}{2}\left[\sum_{i}\binom{n_{i.}}{2}+\sum_{j}\binom{n_{.j}}{2}\right]-\left[\sum_{i}\binom{n_{i.}}{2}\sum_{j}\binom{n_{.j}}{2}\right]\Big/\binom{N}{2}}.$$

The ARI ranges from 0 to 1. Higher index value indicates better agreement between two partitions.

Reference:

1 Elliott MR, Little, R.J.A.: Model-based Alternatives to Trimming Survey Weights. *journal of Official Statistics* 2000; **16**: 191-209.

2 Halpern EF: Polynomial regression from a Bayesian approach. *Journal of the American Statistical Association* 1973; **68**: 137-143.

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

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

5 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.

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

Supplementary Figures:

Figure S1. 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.

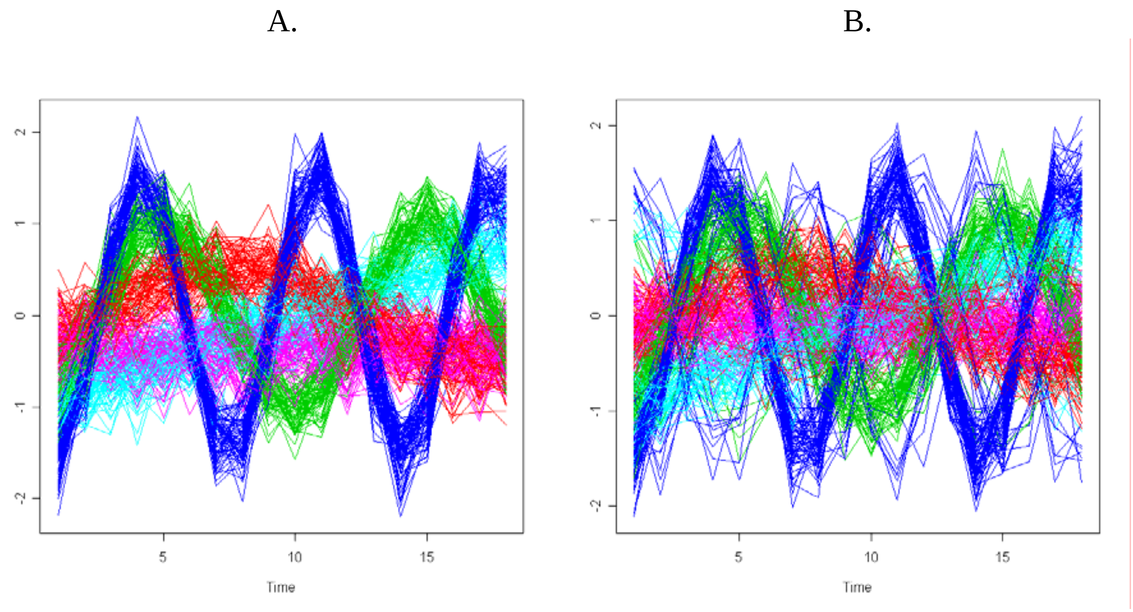


Figure S2. Trace plots of the real galactose data adapted from Ideker et al. (2001). There are 205 genes that belong to four different GO functional categories. Data were collected from 20 different perturbation experiments conducted on the GAL pathway. (A) expression profiles of all 205 genes; (B) a sample output from the CRC program, each subgraph represents a single cluster identified.

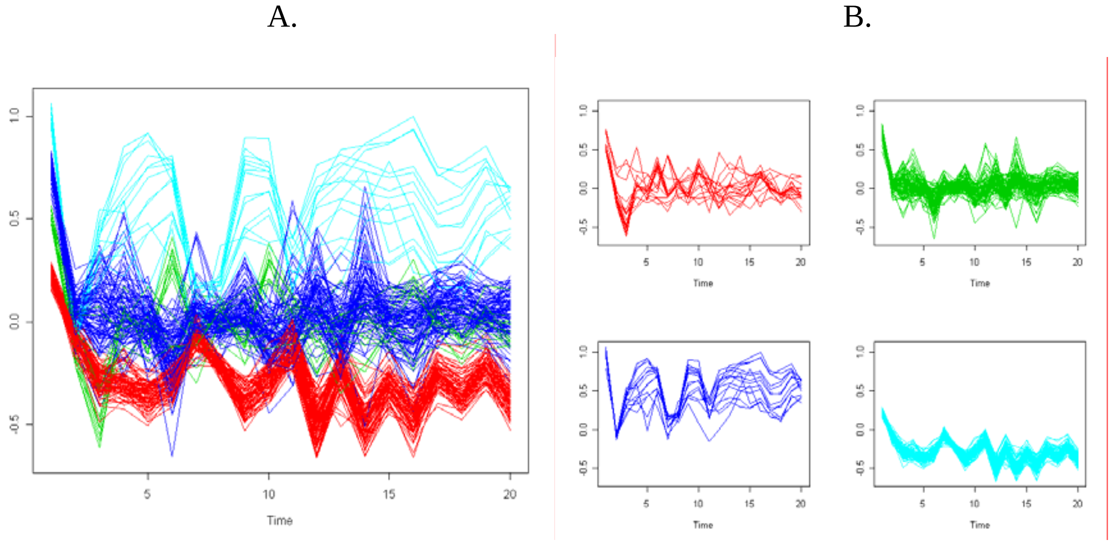
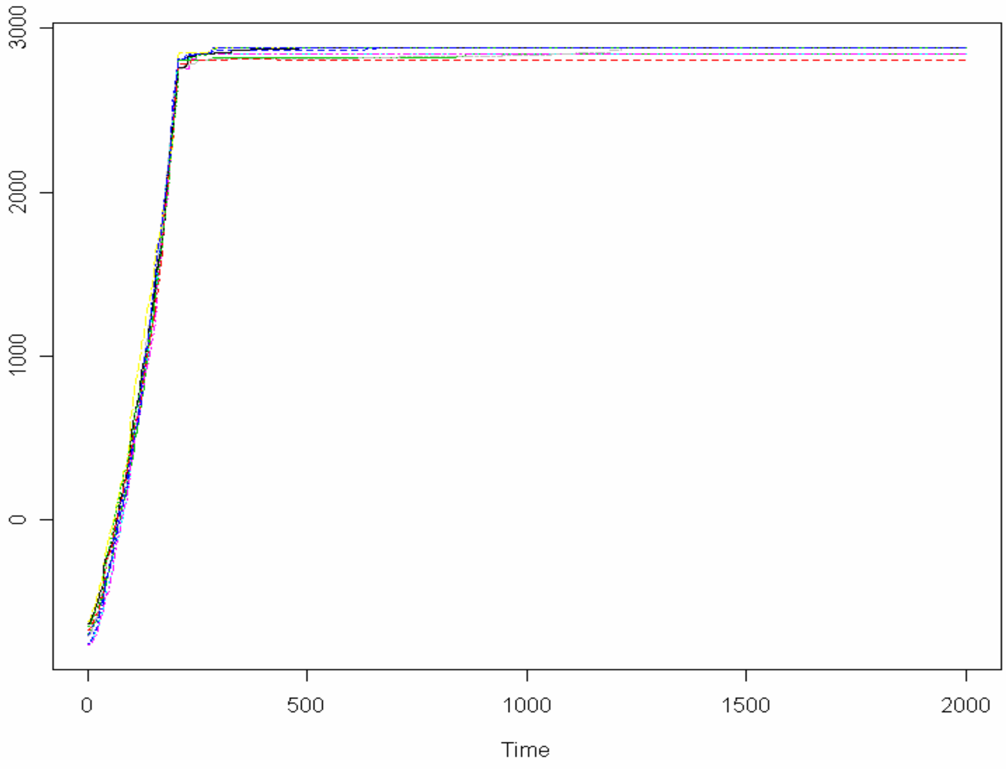


Figure S3. Marginal likelihood trace plots of 10 independent Markov chains, each with 10 cycles of iterations.


Time