

Identifying statistically significant patterns of expression via Bayesian Infinite Mixture Models

Mario Medvedovic, PhD, University of Cincinnati Medical Center, 3223 Eden Av ML56, Cincinnati, OH 45267-0056, email: medvedm@email.uc.edu

Expression data generated by DNA arrays incorporates different sources of variability present in the process of obtaining fluorescence intensity measurements. When examining expression profiles of thousands of genes at once, certain groups of genes will exhibit some level of similarity purely due to chance. Such spurious results are almost inevitable unless a proper statistical model is applied to assess the statistical significance of the observed patterns. Assessing statistical significance of observed expression patterns means determining the level of similarity that is unlikely to be the result of random fluctuations in observed data.

Suppose that T gene expression profiles were observed across M experimental conditions. If x_{ki} is the differential expression of the i^{th} gene for the k^{th} experimental condition, then $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{Mi})$ denotes the complete expression profile for the i^{th} gene. All gene expression profiles can be viewed as being generated by $Q < T$ 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 cluster to which the i^{th} expression profile belongs ($c_i = k$ means that the i^{th} expression profile belongs to the k^{th} cluster), then a "clustering" is defined by a set of classification variables for all expression profiles $\mathbf{c} = (c_1, c_2, \dots, c_T)$. The values of classification variables are meaningful only to the extent that all observed expression profiles having the same value for their classification variable form a cluster. In our probabilistic model, expression profiles that cluster together are assumed to be generated by a single Multivariate Normal random variable. Parameters of this random variable describe a "pattern" that generates corresponding cluster of expression profiles. We developed a statistical procedure based on the Bayesian Infinite Mixture model in which conclusions about the probability of a set of profiles being generated by the same pattern are based on the posterior probability distribution of clusterings given the data $p(\mathbf{c} | \mathbf{x}_1, \dots, \mathbf{x}_T)$.

The following hierarchical model defines the stochastic procedure that is assumed to generate gene expression profiles. This model implicitly defines the posterior distribution of the classification set \mathbf{c} and consequently of the number of clusters (patterns) in the data Q .

$$p(\mathbf{x}_i | c_i = j, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_Q, \sigma_j^2) = f_N(\mathbf{x}_i | \boldsymbol{\mu}_j, \sigma_j^2 \mathbf{I}) \quad (\text{I})$$

$$p(c_i) = \prod_{j=1}^Q \pi_j^{I(c_i=j)} \quad (\text{II})$$

$$p(\boldsymbol{\mu}_j) = f_N(\boldsymbol{\mu}_j | \boldsymbol{\mu}_x, \sigma_x^2 \mathbf{I}) \quad (\text{III})$$

$$p(\sigma_j^{-2}) = f_G(\sigma_j^{-2} | 1/2, \frac{\sigma_x^2}{2}) \quad (\text{IV})$$

$$p(\pi_1, \dots, \pi_Q) = f_D(\pi_1, \dots, \pi_Q | \alpha/Q, \dots, \alpha/Q) \quad (\text{V})$$

In equations above, $p(x | \boldsymbol{\theta})$ denotes the marginal probability distribution function of x given parameters $\boldsymbol{\theta}$; f_N , f_G and f_D represent probability density functions of Multivariate Normal, Gamma, and Dirichlet random variables respectively and $I(c_i=j)=1$ whenever $c_i=j$ and it is 0 otherwise. This model is a multivariate extension of the previously described univariate Bayesian Infinite Mixture model (1) and it represents the starting point in this project. Currently we are experimenting with different generalizations and extensions of this model including a more general covariance structure, adding

additional levels of variability, putting a prior distribution on the α parameter, etc. As Q approaches infinity, this model is a special case of the general Dirichlet Process Prior Mixture model (2).

The goal of the statistical analysis based on this model is to approximate the joint posterior distribution of classification vectors given data, $p(\mathbf{c} | \mathbf{x}_1, \dots, \mathbf{x}_T)$, which is implicitly specified by this hierarchical model but can not written in the closed form. However, it can be shown (2) that the posterior marginal distribution of classification variables when Q approaches infinity is fully specified by following two equations:

$$p(c_i = j | \mathbf{c}_{-i}, \mathbf{x}_i, \boldsymbol{\mu}_j, \sigma_j^2) = b \frac{n_{-i,j}}{T-1+\alpha} f_N(\mathbf{x}_i | \boldsymbol{\mu}_j, \sigma_j^2 \mathbf{I}) \quad (\text{VI})$$

$$p(c_i \neq c_j, j \neq i | \mathbf{c}_{-i}, \mathbf{x}_i, \boldsymbol{\mu}_x, \sigma_x^2) = b \frac{\alpha}{T-1+\alpha} f_N(\mathbf{x}_i | \boldsymbol{\mu}_j, \sigma_j^2 \mathbf{I}) p(\boldsymbol{\mu}_j, \sigma_j^2 | \boldsymbol{\mu}_x, \sigma_x^2) d\boldsymbol{\mu}_j d\sigma_j^2 \quad (\text{VII})$$

where, $n_{-i,c}$ is the number of expression profiles classified in c , not counting the i^{th} profile, \mathbf{c}_{-i} is the classification vector for all except the i^{th} profile.

Posterior marginal distributions for other model parameters are given in the following two equations

$$f(\boldsymbol{\mu}_j | \mathbf{c}, \boldsymbol{\mu}_x, \sigma_j^2, \sigma_x^2, \mathbf{x}_1, \dots, \mathbf{x}_T) = f_N\left(\boldsymbol{\mu}_j \mid \frac{\sigma_x^2 n_j \bar{\mathbf{x}}_j + \sigma_j^2 \boldsymbol{\mu}_x}{\sigma_x^2 n_j + \sigma_j^2}, \frac{\sigma_x^2 \sigma_j^2}{\sigma_x^2 n_j + \sigma_j^2} \mathbf{I}\right) \quad (\text{VIII})$$

$$f(\sigma_j^{-2} | \mathbf{c}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_Q, \boldsymbol{\mu}_x, \sigma_x^2, \mathbf{x}_1, \dots, \mathbf{x}_T) = f_G\left(\sigma_j^{-2} \mid \frac{M\Gamma+1}{2}, \frac{s_j^2 + \sigma_x^2}{2}\right) \quad (\text{XIX})$$

Gibbs sampler (3) is a general procedure for sampling observations from multivariate distributions. It proceeds by iteratively drawing observations from marginal distributions of all components. Under mild condition, the distribution of generated multivariate observations converges to the target distribution. The Gibbs sampler for generating sequence of clusterings $\mathbf{c}^1, \mathbf{c}^2, \mathbf{c}^3, \dots, \mathbf{c}^G$ proceeds as follows

Initialization phase: The algorithm is started by assuming that all profiles are clustered together. That is

\mathbf{c}^0 is initialized as:

$$c_1^0 = c_2^0 = \dots = c_T^0 = 1$$

Consequently, Q_0 is set to one. Corresponding pattern parameters $\boldsymbol{\mu}_1$ and σ_1^2 are generated as random samples from their prior distributions (III) and (IV) respectively.

Iterations: Given parameters after the k^{th} step ($\mathbf{c}^k, Q_k, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{Q_k}, \sigma^2$), the $k+1^{\text{st}}$ set of parameters is generated by first updating classification variables, that is drawing \mathbf{c}^{k+1} according to (VI) and (VII).

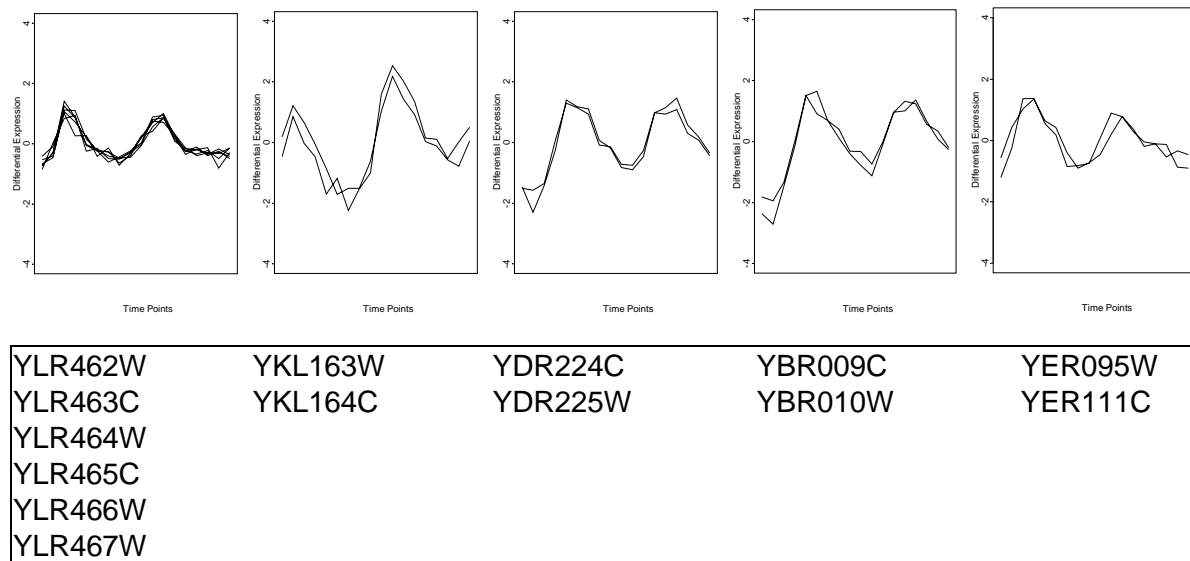
Given that, new $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{Q_k}$ and σ^2 are generated according to (VIII) and (XIX). Whenever the number of profiles in a cluster falls to zero, the cluster is removed from the list. A new cluster is created whenever a $c_i \neq c_j$ for all $i \neq j$ is selected.

It can be shown (2) that this algorithm, in the limit, generates clusterings from the desired posterior distribution of clusterings. Therefore it can be assumed that the empirical distribution of generated clusterings $\mathbf{c}^G, \mathbf{c}^{G+1}, \dots$, after G "burn-in" samples, approximates the true posterior distribution of clusterings. Groups of genes that had common assignments in a large proportion of generated clusterings are likely to have been generated by the same underlying pattern. That is, the proportion of clusterings in which a group of genes had common assignments approximates the probability that they are generated by the same underlying pattern.

Preliminary results

The power of this procedure to detect statistically significant patterns is illustrated by the analysis of α factor-based synchronization cell-cycle data (a subset of the whole cell-cycle data set). Only profiles of genes that were two-fold induced or repressed in for at least one time point, and had valid observations at all time points were used in this preliminary analysis. Expression profiles of 433 genes satisfied both of these conditions. In this analysis we looked for highly specific patterns of expression even if represented by two genes. A cluster of expression profiles was considered to be generated by the same pattern if its elements clustered together in at least 70% of 90,000 clusterings generated by the Gibbs sampler after 10,000 burn-in cycles. Several clusters implicated at this level of significance are shown in Figure 1. A striking feature of most of the identified clusters is the genomic proximity of ORF's represented. It is possible that some of these similarities are due to the fact that spotted clones represent the same gene. Currently, we are assessing the significance of observed similarities and analyzing the complete cell-cycle data set.

Figure 1



The flexibility of this approach to clustering is enormous. By using a specific covariance structure it is possible to assess more subtle relationships between profiles within the same clusters as well as between different clusters. Finally, the Gibbs sampler described here can be easily modified to handle incomplete profiles (profiles missing some data points) by imputing missing data.

Reference List

- (1) Rasmussen CA. The Infinite Gaussian Mixture Model. *Advances in Neural Information Processing Systems* 12 2000;554-560.
- (2) Neal RM. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* 2000; 9:249-265.
- (3) Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall, 1996.