Clustering Short Gene Expression Profiles

Ling Wang[1], Marco Ramoni[2], and Paola Sebastiani[1]

[1] Department of Biostatistics, Boston University School of Public Health,
Boston, MA, 02118, USA
{wangling, sebas}@bu.edu
http://www.bu.edu/dbin/sph/departments/biostatistics
[2] Children's Hospital Informatics Program, Harvard Medical School,
Boston, MA, 02115, USA
marco_ramoni@harvard.edu

Abstract. The unsupervised clustering analysis of data from temporal or dose-response experiments is one of the most important and challenging tasks of microarray data anlysis. Here we present an extension of CAGED (Cluster Analysis of Gene Expression Dynamics, one of the most commonly used programs) to identify similar gene expression patterns measured in either short time-course or dose-response microarray experiments. Compared to the initial version of CAGED, in which gene expression temporal profiles are modeled by autoregressive equations, this new method uses polynomial models to incorporate time/dosage information into the model, and objective priors to include information about background noise in gene expression data. In its current formulation, CAGED results may change according to the parametrization. In this new formulation, we make the results invariant to reparametrization by using proper prior distributions on the model parameters. We compare the results obtained by our approach with those generated by STEM to show that our method can identify the correct number of clusters and allocate gene expression profiles to the correct clusters in simulated data, and produce more meaningful Gene Ontology enriched clusters in data from real microarray experiments.

1 Introduction

Since the original development of microarray technology, unsupervised machine learning methods, clustering methods in particular, have provided a data analytical paradigm and played a central role in the discovery of functionally related genes. Different unsupervised methods have been used to analyze microarray data in order to portray various gene functional behaviors. Correlation-based hierarchical clustering [2] is today one of the most popular analytical methods to characterize gene expression profiles. In [9], we introduced a Bayesian model-based clustering method that takes into account the dependency and dynamic nature of gene expression data measured in temporal experiments. This algorithm, implemented in CAGED (Clustering Analysis of Gene Expression Dynamics), models gene expression temporal profiles by autoregressive equations

A. Apostolico et al. (Eds.): RECOMB 2006, LNBI 3909, pp. 60–68, 2006.

and uses improper prior distributions on the model parameters. As a general framework, CAGED can be used to represent a variety of cross-correlated gene expression data beyond standard temporal experiment, such as dose response data.

It has been recently shown in [3] that the model based formulation implemented in CAGED is more appropriate to cluster long temporal gene expression profiles, possibly measured at regularly spaced time points. There are many scenarios in which experiments are conducted either over a short number of time points, or at a small number of different dosages of drugs. Due to biological considerations, intervals between consecutive time points may not be the same, and the variations in dosages may not be constant. Motivated by these situations, we present an algorithm that uses polynomials models of time or dosage to capture the dynamics of gene expression profiles. The use of polynomial models however requires the specification of proper prior distributions for the regression parameters, so that to ensure the model search algorithm is invariant to reparameterization of time or dosages [7]. A further advantage of the use of proper priors on the model parameters is to include information about background noise of gene expression measured at low intensity, with the effect of making the algorithm more robust to noise and less prone to false positives.

Compared to autoregressive models, polynomial models incorporate information about time/dosage in the design matrix. Therefore, they do not require that the temporal profiles are stationary and appear to be particularly suitable to describe short expression profiles, possibly sampled at irregularly spaced points. By using the same heuristic search strategy in [9], our algorithm can automatically cluster the gene expression data into groups of genes whose profiles are generated by the same process. Furthermore, the Bayesian model-based formulation of the algorithm provides us a principled way to automatically choose the number of clusters with the maximum posterior probability. By properly specifying the prior distribution of the parameters, the clustering model is invariant to linear transformations of time/dosage. In this paper we first describe the Bayesian clustering model in Section 2. In Section 3, we evaluated the accuracy of the results obtained using this method on three simulated datasets and on the immune response data from [4]. We found that compared to STEM, our method is able to reconstruct the generating processes with higher accuracy in simulated data, and produce more Gene Ontology enriched clusters for data from real microarray experiment.

2 Model Formulation

A short time-course/dosage experiment exploring the behavior of J genes usually consists of a set of n microarrays, each measuring the gene expression level x_{jt_i} at a time point/dosage t_i, $i = 1, 2, ..., n$. For each gene, we denote the fold changes of expression levels relative to the first sample (normalized), transformed in natural logarithmic scale, by $S_j = \{x_{jt_1}, x_{jt_2}, ..., x_{jt_n}\}$, $j = 1, 2, ..., J$. These J genes are believed to be generated from an unknown number of processes, and our goal is to group these J genes into clusters by merging genes with similar expression patterns.

The clustering method currently implemented in CAGED is based on a novel concept of similarity for time series from which we derive a model-based description of a set of clusters. We assume that two gene expression profiles are similar when they are generated by the same stochastic process represented by the same parametric model. Under this definition of similarity, the clustering method groups gene expression profiles that are similar into the same cluster. To achieve this objective, CAGED has three components:

1. A model describing the dynamics of gene expression temporal profiles;
2. A probabilistic metric to score different clustering models based on the posterior probability of each clustering model;
3. A heuristics to make the search for the best clustering model feasible. The heuristic was introduced in [8] and adapted to the specific task of clustering gene expression temporal profiles in [9].

In the current implementation, CAGED uses autoregressive models to represent temporal cross-correlation. Here, we replace these models with polynomial models to describe normalized temporal patterns of gene expression data from short temporal/dose-response microarray experiments. The polynomial model describing the temporal pattern of expression for a gene j can be written as

$$x_{jt_i}|\beta_j, \epsilon_{jt} = \mu_j + \beta_{j1}t_i + ... + \beta_{jp}t_i^p + \epsilon_{jt_i}$$

where $\beta_j = (\mu_j, \beta_{j1}, ..., \beta_{jp})^T$ is the vector of regression coefficients that are assumed to be random variables, and ϵ_{jt_i} is random error. Using a matrix notation, we have

$$x_j = F\beta_j + \epsilon_j \tag{1}$$

where $x_j = (x_{jt_1}, x_{jt_2}, ..., x_{jt_n})^T$, F is the $n \times (p+1)$ design matrix with the i^{th} row being $(1, t_i, t_i^2 ..., t_i^p)$, $\epsilon_j = (\epsilon_{jt_1}, \epsilon_{jt_2}, ..., \epsilon_{jt_n})^T$ is the vector of uncorrelated errors that we assume to be normally distributed, with $E(\epsilon_{jt_i}) = 0$ and $V(\epsilon_{jt_i}) = 1/\tau_j$, and the value p is the polynomial order.

We assume a proper normal-gamma prior density on the parameters β_j and τ_j. Therefore, the marginal distribution of τ_j and the distribution of the regression parameters β_j, conditional on τ_j, are

$$\tau_j \sim \text{Gamma}(\alpha_1, \alpha_2)$$
$$\beta_j|\tau_j \sim N(\beta_0, (\tau_j R_0)^{-1})$$

where R_0 is the identity matrix. The prior hyper-parameters $\alpha_1, \alpha_2, \beta_0$ are identical across genes. One of the advantages offered by this novel parametrization is the possibility to include information about background noise and, in so doing, enables the clustering algorithm to properly handle it. We will show next a method to define the hyper-parameters so that to incorporate information about background noise.

Given the data S_j — a set of observed expression values for gene j — we can then estimate the model parameters β_j and τ_j by updating their prior distribution into the posterior distribution using Bayes' Theorem:

$$f(\beta_j, \tau_j | x_j, p) = \frac{f(x_j|\beta_j, \tau_j, p) f(\beta_j, \tau_j)}{f(x_j|p)}.$$

Standard conjugate analysis leads to compute the marginal likelihood of the data

$$f(x_j|p) = \frac{1}{(2\pi)^{n/2}} \frac{(\det R_0)^{1/2}}{(\det R_{jn})^{1/2}} \frac{\Gamma(\alpha_{j1n})}{\Gamma(\alpha_1)} \frac{\alpha_{j2n}^{\alpha_{j1n}}}{\alpha_2^{\alpha_1}} \tag{2}$$

and hence a closed form solution of the posterior distribution of the model parameters τ_j and β_j [1]

$$\begin{aligned} \tau_j | x_j &\sim \text{Gamma}(\alpha_{j1n}, \alpha_{j2n}) \\ \beta_j | x_j, \tau_j &\sim \text{N}(\beta_{jn}, (\tau_j R_{jn})^{-1}) \end{aligned}$$

where

$$\begin{aligned} \alpha_{j1n} &= \alpha_1 + \frac{n}{2} \\ 1/\alpha_{j2n} &= \frac{-\beta_{jn}^T R_{jn} \beta_{jn} + x_j^T x_j + \beta_0^T R_0 \beta_0}{2} + \frac{1}{\alpha_2} \\ R_{jn} &= R_0 + F^T F \\ \beta_{jn} &= R_{jn}^{-1} (R_0 \beta_0 + F^T x_j) \end{aligned}$$

Specification of the hyper-parameters of the prior distribution is an important component of the analysis and we take the approach to define *objective* hierarchical prior distributions on the parameters β_j and τ_j. The main intuition is to use the expression values of genes that are not used in further analysis to model the baseline hyper-variability of gene expression measured with microarrays. Several statistical software for low-level preprocessing of gene expression data score the intensities that represent relative expressions. For example the statistical software implemented in MAS 5.0 and GCOS to process expression data measured with Affymetrix arrays uses a non-parametric statistical method to label gene expression as "absent", "marginally present" or "present". These calls are based on significance tests of differences between intensities of matched probe pairs [10]. Absent calls may denote either technical errors, non-detectable expression or non-expression of the gene in the target, so that investigators are recommended not to use genes that are labelled as absent in the majority of the Affymetrix microarray samples. The more recent Illumina system for microarray data [5] assigns a quality control score to each expression summary and recommends users not to consider genes that have a score lower than 0.99. In both systems, between 25–50% of the total number of genes/probes in the arrays are usually disregarded from further analysis when they are labelled as absent or scored too low. These data however contain information about the variability of non expressed genes and therefore we use them to build our prior distributions.

We assume that disregarded genes do not exhibit any specific patterns, so after normalization and log transformation, they are expected to simply represent noise around zero. Therefore, assuming that $\beta_0 = 0$, then we only need to consider the precision parameters τ_j. We further assume that all absent gene have the same precision. Now let x_{at_i} be the normalized and log-transformed expression of one of these genes at time $t_i, i = 1, ..., n$, then $x_{at_i}|\tau \sim N(0, 1/\tau)$, where τ is the precision parameter whose prior distribution is $\tau \sim \text{Gamma}(\alpha_1, \alpha_2)$. From the properties of conditional mean and conditional variance, it is easy to show that the marginal variance of the data is functionally related to the hyper-parameters:

$$\alpha_2 = \frac{1}{(\alpha_1 - 1)\sigma_a^2}$$

where σ_a^2 is the sample variance of the disregarded expression data. So here, with $\alpha_1 = 2$, we can easily specify the hyper-parameter α_2.

3 Evaluation

We evaluate our algorithm by simulation study and analysis of the data from the microarray experiment on immune response to Helicobacter pylori infection in [4], and compare it to the program STEM recently introduced in [3]. Section 3.1 reports the results from three simulation studies, and section 3.2 presents the analysis of real data from [4]. All the analysis were done with our clustering algorithm and STEM.

3.1 Simulation Study

We simulated three sets of 5,000 gene expression profiles measured over 5 different time points: 0, 1, 2, 3, 4. All the profiles were generated assuming the gene expressions were normalized and transformed into natural logarithmic scale. The first 5,000 profiles were simply noise, and were generated from a normal distribution with mean 0 and a variance representing the average variability of noisy patterns that we inferred from the analysis of previous real microarray experiments. For this dataset, we generated another 1,000 noise profiles to be the data from genes with low intensities and we used these to specify the hyperparameters of the model. The second 5,000 profiles had 4 different baseline patterns and some background noise. Data for each gene expression profile were generated by adding random noise to one of the four baseline patterns (Figure 1 left panel), and the gene expression profiles of the background noise were generated from a normal distribution with mean 0 and a variance inferred from previous analysis of temporal microarray experiments. The number of genes representing each of the four patterns and the background noise was randomly chosen from 11 to 5,000. For this dataset, we simulated another set of 5,000 noise profiles with low intensities to specify the hyperparameters. The third 5,000 expression profiles had 6 different baseline patterns (Figure 1 right panel) that are more difficult to discriminate, plus some background noise. The data were generated using the

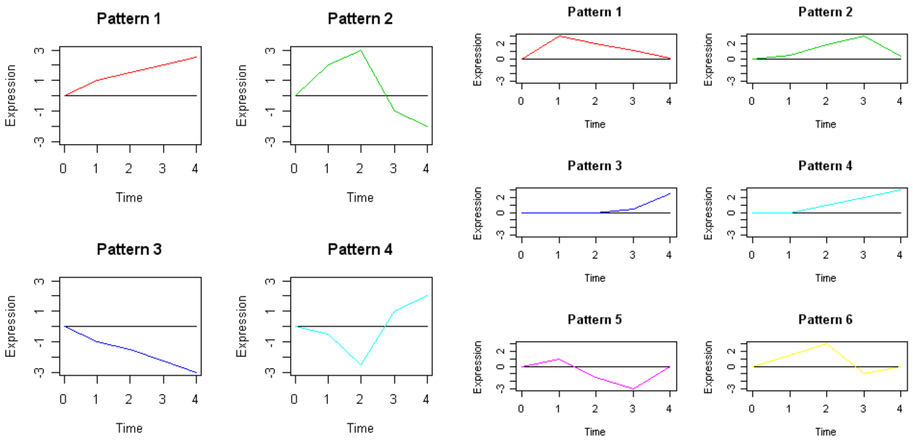


Fig. 1. Left: The 4 distinct baseline patterns of the simulated data. Right: The 6 indistinct baseline patterns of the simulated data.

Table 1. Clustering results of simulated datasets from our program and STEM

Simulated dataset	number of true profiles	# of profiles our program found	# of significant profiles STEM found
noise	0	0	17
4 patterns with noise	4	4	4
6 patterns with noise	6	6	11

same strategy as the second dataset. For the third dataset another 5,000 noise profiles with low intensities were simulated for the specification of hyperparameters. Note that in the last two datasets with planted patterns, the range of variability of the simulated patterns was within the range of variability of the noisy patterns.

Each of the three datasets was analyzed using our clustering algorithm, with polynomial orders 0, 4 and 4 respectively. We also analyzed these three datasets using STEM, with the recommended default settings of $c = 2$, and 50 possible profiles and used Bonferroni correction to control for multiple comparisons. To be consistent, we did not filter out any genes in any of these analysis, but rather used the separately generated noise profiles to specify the hyperparameters. Table 1 reports the clustering results from both our program and STEM, from which we can observe that our program successfully recovered the correct number of patterns, plus the background noise, whereas STEM discovered 17 significant profiles from the noise-only dataset, and 11 significant profiles from the dataset with 6 true patterns.

Figure 2 shows that our program grouped all the gene expression profiles in the noise-only dataset into a single cluster, representing the expected indistinguishability of pure noise. By contrast, STEM found 17 significant profiles in these noise-only data. For the simulated data with 4 different baseline patterns,

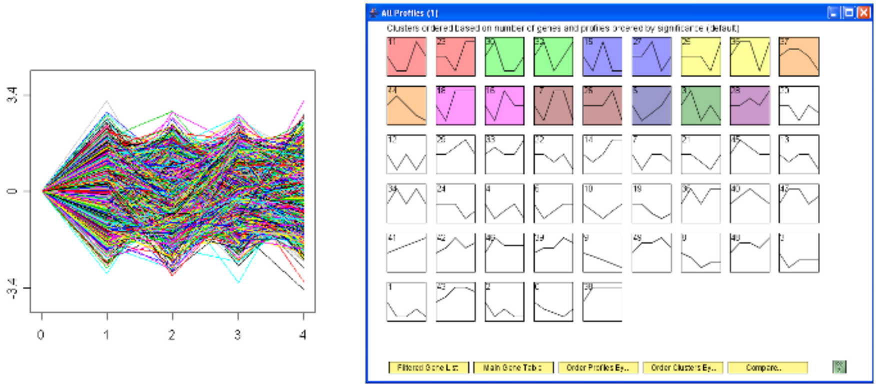
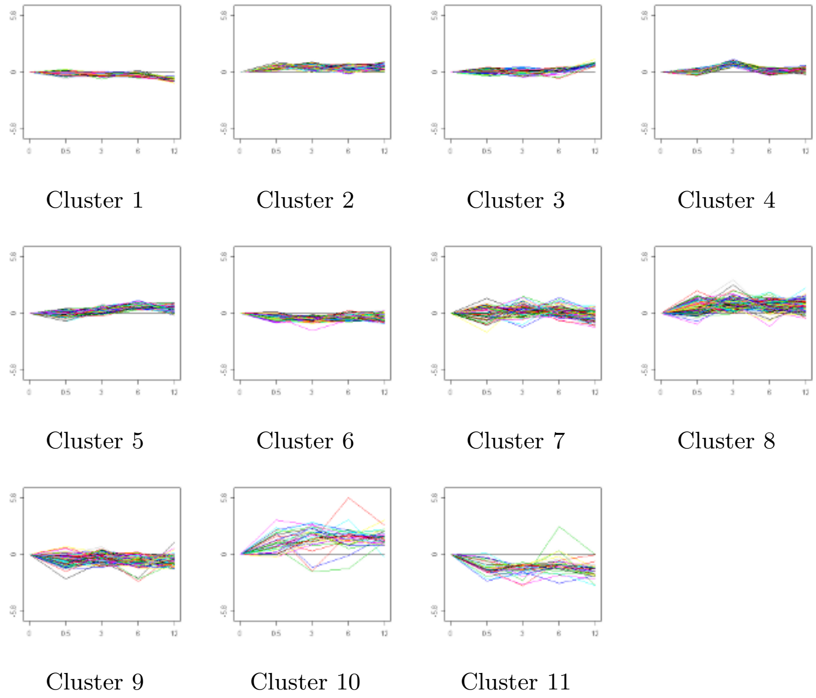


Fig. 2. Left: The only noise cluster found by our program. Right: The results from STEM. The 17 colored profiles are found to be significant by STEM.

our clustering algorithm gave 5 clusters, of which 4 have profiles matching the baseline profiles in Figure 1 left panel, and 1 noise cluster. For these 5,000 genes, 10 were allocated to the wrong cluster, with 3 false negatives (genes with true pattern allocated to noise cluster) and 7 false positives (noise genes allocated to clusters with pattern). The 4 significant profiles that STEM found have 3 profiles that are similar to the baseline patterns, but the up-regulated profile that corresponds to pattern 1 in Figure 1 left panel was not labeled as significant (p value=1). The simulated dataset with 6 different baseline profiles are designed to be harder to discriminate, and our program successfully found 7 clusters, of which 6 had profiles matching the baseline patterns in Figure 1 right panel, and 1 contained only noise. For this set of 5,000 genes, 83 are allocated to the wrong cluster, with 12 false positives and 41 false negatives. STEM analysis found 11 significant profiles for this data.

3.2 Real Data Analysis

We analyzed the data from the microarray experiment on immune response to Helicobacter pylori infection in [4] to further evaluate our clustering algorithm. In this experiment, human cDNA microarrays were used to investigate the temporal behavior of gastric epithelial cells infected with Helicobacter pylori strain G27 and some other mutants. We used the selected 2,243 genes after the data pre-processing in [3] for clustering, and the 17,352 genes that were filtered out were used to specify the hyperparameters. We then normalized and transformed the data into natural log scale, and performed the cluster analysis with polynomial order of 4. The time points we used in the model were the actual time at which the experiments were carried out: 0, 0.5, 3, 6 and 12. Our clustering algorithm returned a total of 11 clusters. Figure 3 shows all the clusters. We then preformed the Gene Ontology enrichment test with EASE [6]. Because there were missing annotations for some genes in each cluster, we carried out the enrichment analysis using only the genes with annotations. Seven out of the 11 clusters



Cluster 1 Cluster 2 Cluster 3 Cluster 4

Cluster 5 Cluster 6 Cluster 7 Cluster 8

Cluster 9 Cluster 10 Cluster 11

Fig. 3. The 11 clusters our program found in the analysis of data from the microarray experiment on immune response to Helicobacter pylori infection

had EASE scores less than 0.05 and hence 63% of the clusters were significantly enriched for GO categories. Cluster 10, which had 38 genes totally and 11 genes with annotations, represented a stable upregulated pattern over time. This cluster is significantly enriched for the immune response GO category (with EASE score 6.85×10^{-3}). Cluster 1 is significantly enriched for mitotic cell cycle genes (EASE score 4.62×10^{-13}) and cell cycle genes (EASE score 2.05×10^{-10}). The STEM analysis described in [3] identified 10 significant profiles, four of which only were found significantly enriched by the GO analysis. Compared to the 63% significantly GO enriched clusters found by our algorithm, the analysis in STEM therefore produces only 40% significantly GO enriched clusters.

4 Conclusions

We have introduced a model reformulation of CAGED using polynomial models of time/dosage with proper prior distributions. We find this formulation to be well suited for clustering analysis of data from short temporal/dosage microarray experiments. The polynomial models that describe the trend are flexible and do not require the gene expression profile to be stationary. We use proper priors in the model so that we can incorporate the background noise information through specifying the hyperparameters with low-intensity genes, and the clustering algorithm becomes invariant to linear transformation on time/dosage. An empirical comparison on simulated data shows that our clustering algorithm can identify the correct number of generating processes, and allocate genes into clusters with low false positives and false negatives. In the analysis of data from the human cDNA microarray experiment on immune response to Helicobacter pylori infection in [4] we found 11 clusters with our algorithm, 7 out of which are significantly enriched by Gene Ontology analysis. In both the empirical study and the analysis of the immune response data to Helicobacter pylori infection, our algorithm performs better than STEM.

References

1. J. M. Bernardo and A. F. M. Smith: *Bayesian Theory.* Wiley, New York, NY, 1994.
2. M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.
3. J. Ernst, G. J. Nau, and Z. Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 21 Suppl. 1:i159-i168, 2005.
4. K. Guillemin, N. Salama, L. Tompkins, and S. Falkow. Cag pathogenicity island-specific response of gastric epithelial cells to helicobacter pylori infection. *Proc. Natl. Acad. Sci. USA*, 99(23):15136–15141, 2002.
5. K. L. Gunderson, S. Kruglyak, M. S. Graigeand F. Garcia, B. G. Kermani, C. Zhao, D. Che, T. Dickinson, E. Wickham, J. Bierle, D. Doucet, M. Milewski, R. Yang, C. Siegmund, J. Haas, L. Zhou, A. Oliphant Ad J. Fan, S. Barnard, and M. S. Chee. Decoding randomly ordered DNA arrays. *Genome Res.*, 14:870–877, 2004.
6. D. A. Hosack, G. Jr. Dennis, B. T. Sherman, H. Clifford Lane, and R. A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(6):4, 2003.
7. R. E. Kass and A. Raftery. Bayes factors. *J. Ameri. Statist. Assoc.*, 90:773–795, 1995.
8. M. Ramoni, P. Sebastiani, and P. R. Cohen. Bayesian clustering by dynamics. *Mach. Learn.*, 47(1):91–121, 2002.
9. M. Ramoni, P. Sebastiani, and I. S. Kohane. Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA*, 99(14):9121–6, 2002.
10. P. Sebastiani, E. Gussoni, I. S. Kohane and M. Ramoni. Statistical challenges in functional genomics (with discussion). *Statist. Sci.*, 18:33–70, 2003.