Bayesian Clustering of Gene Expression Dynamics: An Application

Sara Lodi[1], Paola Sebastiani[2], Daniela Cocchi[3], and Marco Ramoni[4]

[1] Dipartimento di Scienze Statistiche,
Università di Bologna, Italy
saralodi@yahoo.com
[2] Department of Biostatistics,
Boston University, MA
sebas@bu.edu
[3] Dipartimento di Scienze Statistiche,
Università di Bologna, Italy
cocchi@stat.unibo.it
[4] Children Hospital Informatics Program,
Harvard Medical School, MA
marco_ramoni@harvard.edu

Abstract. This paper describes an application of Caged (Cluster Analysis of Gene Expression Dynamics) to a data set of gene expression temporal profiles from Saccaromicys Cerevisiae. The goal of the analysis is to identify groups of genes with similar temporal patterns of expression during the cell cycle. We show that Caged groups gene expression temporal profiles into meaningful clusters and identifies genes with a putative role in the cell cycle.

1 Introduction

Several applications of genome-wide clustering methods focus on the temporal profiling of gene expression patterns. Temporal profiling offers the possibility of observing the cellular mechanisms in action and tries to break down the genome into sets of genes involved in the same or related processes. Standard clustering methods, such as the hierarchical clustering method of Eisen et al. (1998) or the self organizing maps (Tamayo et al., 1999), should not be used to analyse the data generated from these experiments because they typically rest on the assumption that the set of observations for each gene are independent and identically distributed (iid). On the other hand, gene expression data collected from temporal experiments are realizations of time series, where each observation may depend on prior ones (Box and Jenkins, 1976; West and Harrison, 1997), and standard similarity measures currently used for clustering gene expression data, such as correlation or Euclidean distance, are invariant with respect to the order of observations: if the temporal order of a pair of series is permuted, their correlation or Euclidean distance will not change.

A second critical problem of clustering approaches to gene expression data is the arbitrary nature of the partitioning process. This operation is often done by visual inspection, by searching for groups of genes with similar expression patterns.

Permutation tests are sometimes used to validate the partitions found by this procedure (Eisen et al., 1998), and a bootstrap-based validation technique is presented in Kerr and Churchill (2001). The gap statistic of Tibshirani et al. (2001) is also used to find the optimal number of groups in the data.

CAGED (Cluster Analysis of Gene Expression Dynamics) is a model based, Bayesian clustering procedure developed by Ramoni et al. (2002) to cluster gene expression profiles measured with microarrays in temporal experiments. Contrary to popular clustering methods, CAGED takes into account explicitly the fact that expression profiles in temporal experiments may be serially correlated and uses a model-based, Bayesian procedure to identify the best grouping of the gene expression data in an automated way. An important property of CAGED is that it automatically identifies the number of clusters and partitions the gene expression time series in different groups on the basis of the principled measure of the posterior probability of the clustering model. In this way, CAGED allows the investigator to assess whether the experimental data convey enough evidence to support the conclusion that the behavior of a set of genes is significantly different from the behavior of another set of genes. This feature is particularly important because decades of cognitive science research have shown that the human eye tends to overfit observations by selectively discounting variance and "seeing" patterns in randomness, see Kahneman et al. (1982). By contrast, a recognized advantage of a Bayesian approach to model selection, such the one adopted in this chapter, is the ability to automatically constrain model complexity and to provide appropriate measures of uncertainty.

We apply CAGED to cluster a data set of gene expression temporal profiles from Saccaromicys Cerevisiae. The goal of the analysis is to detect those genes whose transcript levels vary periodically within the cell cycle. Cell cycle is a very complex ordered set of events that consists of several phases culminating in cell growth and division into daughter cells (*mitosis*). During this period, the cell is constantly synthesizing RNA, producing protein and growing in size. In the *G1* phase, the cell increases in size, produces RNA and synthesizes proteins. The next step is the synthesis phase *S* in which DNA replication occurs. This phase is followed by the *G2* phase in which the cell continues to grow and to produce new proteins, and by the mitosis (*M* phase). Many genes are involved in DNA synthesis, budding and cytokinesis that occur only once per cell cycle. In addition, many of these genes are also involved in controlling the cell cycle itself. For this reason, the expression levels of the genes that have a regulatory role in cell cycles are expected to show periodical behaviors across time and to present at least one peak during the phase in which they are activated. The data were originally analyzed by Spellman et al. (1998), using Fourier models, and the authors identified several clusters by visual inspections. We show that Caged finds automatically clusters of gene expression temporal profiles that exhibit periodic behavior.

The next section gives a brief description of the model based clustering procedure that is implemented in CAGED. Section 3 provides details of the analysis and conclusions and suggestions for further work are in Section 4.

2 Caged

The clustering method implemented in CAGED is based on a novel concept of similarity for time series: two time series are similar when they are generated by the same stochastic process. Therefore, the components of CAGED are a model describing the dynamics of gene expression temporal profiles, a metric to decide when two gene expression temporal profiles are generated by the same stochastic process, and a search procedure to efficiently explore the space of possible clustering models.

CAGED models gene expression temporal profiles by autoregressive equations (West and Harrison, 1997). Let $S_j = \{x_{j1}, \ldots, x_{jt}, \ldots, x_{jn}\}$ denote a stationary time series. An autoregressive model of order p, say AR(p), for the time series can be described in matrix form as

$$x_j = X_j \beta_j + \epsilon_j$$

where x_j is the vector $(x_{j(p+1)}, \ldots, x_{jn})^T$, X_j is the $(n-p) \times (p+1)$ regression matrix whose tth row is $(1, x_{j(t-1)}, \ldots, x_{j(t-p)})$ for $t > p$, β_j is the vector of autoregressive coefficients and ϵ_j the vector of uncorrelated errors that are assumed normally distributed with expected value $E(\epsilon_{jt}) = 0$ and precision τ_j, for any time point t. Given the data, the model parameters can be estimated using standard Bayesian procedures, and details are in Ramoni et al. (2002).

To select the set of clusters, CAGED uses a novel model-based Bayesian clustering procedure. A set of clusters $C_1, \ldots, C_c$, each consisting of m_k time series, is represented as a model M_c. The time series assigned to each cluster are treated as independent realizations of the dynamic process represented by the cluster, which is described by an autoregressive equation. The posterior probability of the model M_c is computed by Bayes theorem as $P(M_c|y) \propto P(M_c) f(x|M_c)$ where $P(M_c)$ is the prior probability of M_c and $f(x|M_c)$ is the marginal likelihood. Assuming independent uniform prior distributions on the model parameters and a symmetric Dirichlet distribution on the cluster probability p_k, the marginal likelihood of each cluster model M_c can be easily computed in closed form by solving the integral:

$$f(x|M_c) = \int f(x|\theta_c) f(\theta_c) d\theta_c,$$

where θ_c is the vector of parameters that describe the likelihood function, conditional on a clustering model M_c, and $f(\theta_c)$ is the prior density. In this way, each clustering model has an explicit probabilistic score and the model with maximum score can be found. In practice, we assume that each clustering model has the same prior probability so that the marginal likelihood $f(x|M_c)$ is the scoring metric of the clustering model M_c.

As the number of clustering models grows exponentially with the number of time series, CAGED uses an agglomerative search strategy, which iteratively merges time series into clusters. The procedure starts by assuming that each of the m gene expression time series is generated by a different process. Thus, the initial model M_m consists of m clusters, one for each time series, with score $f(x|M_m)$. The next step is the computation of the marginal likelihood of the $m(m-1)$ models in which two of the m profiles are merged into one cluster. The model M_{m-1} with maximal marginal likelihood is chosen and the merging is rejected if $f(x|M_m) \geq f(x|M_{m-1})$ and the procedure stops. If $f(x|M_m) < f(x|M_{m-1})$ the merging is accepted, a cluster C_k merging the two time series is created, and the procedure is repeated on the new set of $m-1$ time series that consist of the remaining $m-2$ time series and the cluster profile. Although the agglomerative strategy makes the search process feasible, the computational effort can be extremely demanding when the number of time series is large. To further reduce this effort, we use a heuristic strategy based on a measure of similarity between time series.

The intuition behind this strategy is that the merging of two similar time series has better chances of increasing the marginal likelihood. The heuristic search starts by computing the $m(m-1)$ pair-wise similarity measures of the time series and selects the model M_{m-1} in which the two closest time series are merged into one cluster. If the merging increases the marginal likelihood, the two time series are merged into a single cluster, a profile of this cluster is computed by averaging the two observed time series, and the procedure is repeated on the new set of $m-1$ time series. If this merging is rejected, the procedure is repeated on the two time series with second highest similarity until an acceptable merging is found. If no acceptable merging is found, the procedure stops. Note that the clustering procedure is actually performed on the posterior probability of the model and the similarity measure is only used to increase the speed of the search process and to limit the risk of falling into local maxima. Similarity measures implemented in CAGED are Euclidean distance, correlation and cross correlation. Empirical evaluations, see Sebastiani et al. (2003b), have shown that this heuristics makes the search process faster, without loosing accuracy. Compared to other clustering methods such as hierarchical clustering or self organizing maps, CAGED identifies the set of clusters with maximum posterior probability without requiring any prior input about the number of clusters and avoids the risk of overfitting.

Standard statistical diagnostics are used as independent assessment measures of the cluster model found by the heuristic search. Once the procedure terminates, the coefficients β_k of the AR(p) model associated with each cluster C_k are estimated by Bayesian Least Squares, while $\hat{\sigma}_k^2 = RSS_k/(n_k - p)$ is the estimate of the within-cluster variance and RSS_k is the within cluster residual sum of squares. The parameter estimates can be used to compute the fitted values $\hat{x}_{ik}$, for the series in each cluster, from which we compute the residuals $x_{ik} - \hat{x}_{ik}$. If AR(p) models provide an accurate approximation of the processes generating the time series, the standardized residuals should behave like a random sample from a standard normal distribution. A normal probability plot, or the residual histogram per cluster, can be used to assess normality. Departures from normality cast doubt on the autoregressive assumption, so that some data transformation, such as a logarithmic transformation, may be needed. Plots of the fitted values versus the observed values and of the fitted values versus the standardized residuals in each cluster provide further diagnostics. To choose the best autoregressive order, we repeat the clustering for different autoregressive orders, $p = 0, 1, \ldots, w$ for some preset w, and compute a goodness-of-fit score defined as $s = cq + \sum n_k \log(n_k - q) - \sum n_k \log(\mathrm{RSS}_k)$, where c is the number of clusters, n_k the total number of time points in cluster C_k. This score trades off model complexity with lack of fit and it generalizes the well known AIC goodness-of-fit criterion of Akaike (1973) to a set of autoregressive models. We use the goodness of fit score to choose the best autoregressive order.

3 Application

3.1 Materials

The data we analyze are expression levels of genes from the budding yeast Saccaromyces Cerevisiae that were collected on spotted cDNA microarrays. Data were drawn from time courses during the cell cycle after synchronization by alpha factor arrest in 18 time points. The original data set is available at http://genome-www.stanford.edu/Saccharomyces and consists of expression profiles of 6178 genes. About 1500 expression profiles had missing data and because the shortness of time series would make traditional imputation methods not very reliable those gene expression profiles were disregarded. To reduce the noise, we excluded those time series in which the ratio between the minimum and maximum expression level did not exceed 2. This filter is justified by the fact that significant biological events are characterized by at least a 2-fold change, see Sebastiani et al. (2003a) for a discussion and further references. With this filter we selected 1574 temporal profiles.

3.2 Methods

We analyzed the data set with CAGED. The software is freely available and can be downloaded from http://www.genomethods.org/caged. Details about the software are described in Sebastiani et al. (2003b). For selecting the most probable cluster model given the data, the user needs to specify the autoregressive order, the distance to be used during the heuristic search, a threshold on the Bayes factor that determines the odds for merging similar time series, and some prior hyper-parameters. We performed the analysis several times varying both the hyper-parameters and the autoregressive order. Among the available similarity measures, Euclidean distance always lead to cluster models with greater marginal likelihood. The Bayes Factor was set to 10 so that an aggregation occurs only if the model in which the clusters are merged is 10 times more probable than the model in which they are separated. This choice is recommended in Kass and Raftery (1995) to force the model selection toward significant dependencies. We also run some sensitivity analysis to prior settings that lead to set the prior precision to 10. Once the parameters were chosen, the last step was the choice of the autoregressive order that was chosen by comparing the goodness of fit of clustering models induced by different autoregressive orders. The autoregressive model that best fitted the data was an autoregressive model of order 2. Therefore, the method found a temporal dependency in the data:

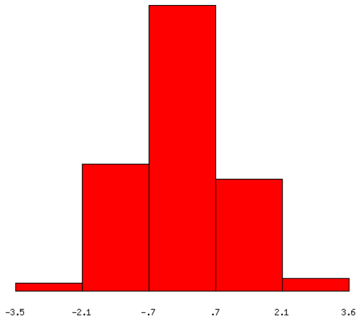


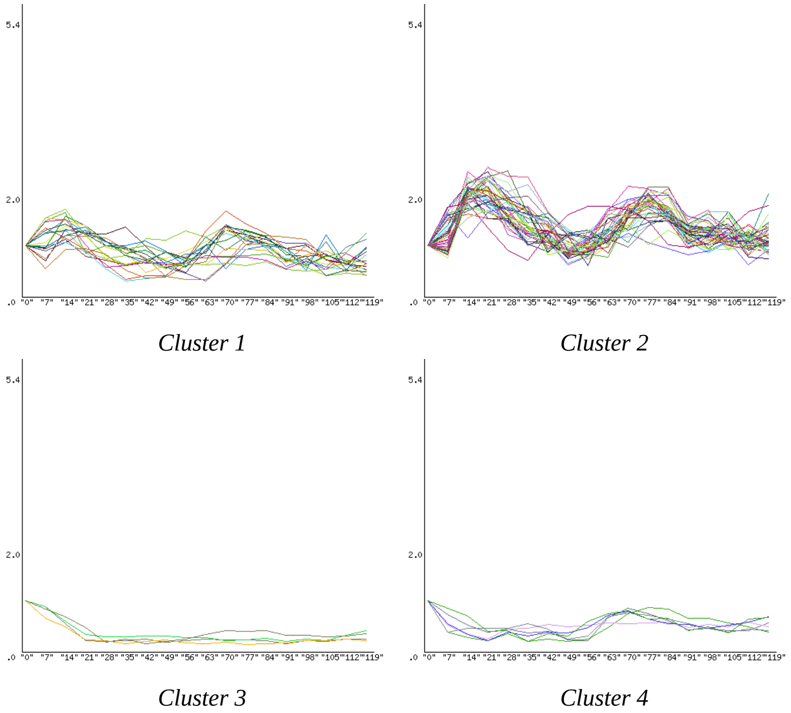
Fig. 1. *Histogram of the standardized residuals for the time series assigned to Cluster 1.*

each observation of a gene expression time series depends on its two immediate predecessors. This model fits the data better than the autoregressive model of order 0, in which observations are assumed to be marginally independent. Goodness of fit statistics confirm that autoregressive models of order 2 provide a good fit. As an example, Figure 1 shows the standardized residuals for the cluster of time series in Figure 2 (Cluster 1 in the top-left panel). The symmetry confirms that the model assumptions are reasonable.

3.3 Results

CAGED merged the 1574 genes time series into 13 distinct clusters and by querying the GeneOntology database (http://www.geneontology.org/), genes assigned to each cluster were annotated by their biological processes. Five of the thirteen clusters have periodical profiles, while four of the clusters have at least one spike of expression during the cell cycle, and two clusters group genes that are systematically upregulated or downregulated during the cell cycle. Two clusters group genes that do not appear to have cell-cycle related change of expression. Among the clusters found by the algorithm, particularly noteworthily are four clusters in which either a periodical trend is detected (Cluster 1 and 2 in Figure 2) or only one spike is detected (Cluster 3 and 4 in Figure 2). All genes belonging to these clusters are strongly cell cycle regulated.

Cluster 1 contains 18 genes that spike at 7 and 70 minutes, so that one can conjecture that they are coregulated during the M/G1 transition of the cell cycle. Peak expressions occur in early G1 phase that consists of growth and preparation of chromosomes for replication. Most of the genes are involved in cell wall, which is laid out during the division of the cell. Five of the 18 genes have unknown functions, and the fact that they are merged into a cluster of genes involved with cell wall suggests that they may have the same function. Cluster 2 contains 66 genes that are characterized by two spikes at time 21 minutes and 77 minutes and, because of the time shift, the conjecture is that these genes are involved in the S/G2 phase. A large proportion of these genes have DNA replication and repair functions, thus



Cluster 1

Cluster 2

Cluster 3

Cluster 4

Fig. 2. *Plot of the gene expression profiles assigned to significant clusters by* CAGED.

confirming the conjecture that the cluster groups genes involved in the S/G2 phase. The third cluster contains four genes that are systematically down regulated during the cell cycle. All genes have a role in cell fusion. The six genes assigned to Cluster 4 are down regulated during the first hour and then spike at about 70 minutes. The functions of the genes assigned to this clusters have already been associated with the cell cycle, and include cell fusion, cell cycle arrest, and completion of separation.

4 Conclusions

Several applications of genome-wide clustering methods focus on the temporal profiling of gene expression. The intuition behind this analytical approach is that genes showing a similar expression profile over time are acting together, because they belong to the same, or similar, functional categories. The novelty and strength of the clustering algorithm implemented in CAGED is that it takes into account the dynamic nature of gene expression data in temporal microarray experiments and the analysis presented in this paper confirms the capability of CAGED to detect groups of gene expression temporal profiles with periodical patterns and genes having related functions in a complex event such as the cell cycle.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pp. 267–281 Budapest, Hu. Kiado.

Box, G. E. P., and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control* (2nd edition). Holden-Day, San Francisco, CA.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA*, *95*, 14863–14868.

Kahneman, D., Slovic, P., and Tversky, A. (Eds.). (1982). *Judgment under Uncertainty: Hueristic and Biases*. Cambridge University Press, New York, NY.

Kass, R. E., and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kerr, M. K., and Churchill, G. A. (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments.. *98*, 8961–8965.

Ramoni, M., Sebastiani, P., and Kohane, I. S. (2002). Cluster analysis of gene expression dynamics. *Proceedings on the National Academy of Sciences of the USA*, *99*, 9121–9126.

Sebastiani, P., Gussoni, E., Kohane, I. S., and Ramoni, M. (2003a). Statistical challenges in functional genomics (with discussion). *Statistical Science*, *18*, 33–70.

Sebastiani, P., Ramoni, M., and Kohane, I. (2003b). Bayesian model-based clustering of gene expression dynamics. In G., P., Irizarry, R., and Zeger, S. L. (Eds.), *The Analysis of Microarray Data: Methods and Software*, pp. 400–427. Springer, New York.

Spellman, P. T., Sherlock, G., and et al (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, *9*, 3273–297.

Tamayo, P., Slonim, D., Mesirov, J., and et. al (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings on the National Academy of Sciences of the USA*, *96*, 2907–2912.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a dataset via the Gap statistic.. *63*, 411–423.

West, M., and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models* (2nd edition). Springer, New York, NY.

Acknowledgements P. Sebastiani and M. Ramoni were supported by the National Science Foundation: Grant ECS-0120309.