1

Bayesian Clustering of Gene Expression Dynamics*

Paola Sebastiani†
Marco Ramoni‡
Isaac Kohane§

Abstract

This chapter presents a Bayesian method for model-based clustering of gene expression dynamics and a program implementing it. The method represents gene expression dynamics as autoregressive equations and uses an agglomerative procedure to search for the most probable set of clusters given the available data. The main contributions of this approach are the ability to take into account the dynamic nature of gene expression time series during clustering and an automated, principled way to decide when two series are different enough to belong to different clusters. The reliance of this method on an explicit statistical representation of gene expression dynamics makes it possible to use standard statistical techniques to assess the goodness of fit of the resulting model and validate the underlying assumptions. A set of gene expression time series, collected to study the response of human fibroblasts to serum, is used to illustrate the properties of the method and the functionality of the program.

*To appear in: Parmigiani, G, Garrett, ES, Irizarry, RA and Zeger, SL (eds), *The Analysis of Gene Expression Data*, Springer, New York, NY, in press.

†Department of Mathematics and Statistics, University of Massachusetts, Amherst MA 01002. Email: sebas@math.umass.edu.

‡Children's Hospital Informatics Program, Harvard Medical School, 300 Longwood Avenue, Boston MA 02115. Email: marco_ramoni@harvard.edu.

§Children's Hospital Informatics Program, Harvard Medical School, 300 Longwood Avenue, Boston MA 02115. Email: isaac_kohane@harvard.edu.

1.1 Introduction

Microarray technology [23, 16] enables investigators to simultaneously measure the expression level of the genome of entire organisms under a particular condition and is reshaping molecular biology. The promise of this technology is the ability to observe the entire genome in action and, in so doing, to uncover its underlying expression mechanisms. Cluster analysis is today one of the favorite unsupervised learning approaches to identify these mechanisms [7, 26, 6, 3]. Albeit different, these clustering algorithms share the general strategy of grouping together genes according to the similarity of their behavior across different experimental conditions or different samples. The intuition behind this approach is that genes acting together belong to similar, or at least related, functional categories. Cluster analysis has become widely popular in molecular biology and it has been successfully applied to the genome-wide discovery and characterization of the regulatory mechanisms of several biological processes and organisms [29, 13, 10, 17].

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 at least related, processes. In these experiments, different experimental conditions correspond to the observation of the genome at a particular time point during the temporal evolution of some biological process. In these cases, standard clustering methods cannot be used any longer because they typically rest on the assumption that the set of observations for each gene are *independent and identically distributed* (iid). While this assumption holds when expression measures are taken from independent biological samples, such as different subjects or different experimental conditions, it is no longer valid when the observations are realizations of a time series, where each observation may depend on prior ones (e.g. [5, 30]). 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. Biomedical informatics investigators over the past decade have demonstrated the risks incurred by disregarding the dependency among observations in the analysis of time series [25, 11]. Not surprisingly, the functional genomic literature is becoming increasingly aware of the specificity of temporal profiles of gene expression data, as well as of their fundamental importance in unraveling the functional relationships between genes [22, 12, 1].

A second critical problem of clustering approaches to gene expression data is the arbitrary nature of the actual partitioning process. The method described here 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, it 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 as decades of cognitive science research have shown that the human eye tends to overfit observations, by selectively discount variance and "seeing" patterns in randomness (e.g. [28, 14, 9]). By contrast, a recognized advantage of a Bayesian approach to model selection, like the one adopted in this paper, is the ability to automatically constrain model complexity [18, 27] and to provide appropriate measures of uncertainty.

We describe here a Bayesian model-based clustering method [21] to profile gene expression time series that explicitly takes into account the dynamic nature of temporal gene expression data. This method is a specialized version of a more general class of methods called *Bayesian Clustering by Dynamics* (BCD) [20], which have been applied to a variety of time series data, ranging from cognitive robotics [19] to official statistics [24]. The main novelty of BCD is the concept of similarity: two time series are similar when they are generated by the same stochastic process. With this concept of similarity, the Bayesian approach to the task of clustering a set of time series consists of searching the *most probable* set of processes generating the observed time series. The method presented here models temporal gene expression profiles by autoregressive equations [5], and groups together the profiles with the highest posterior probability of being generated by the same process. Although this chapter will adopt autoregressive equations to model the dynamic of gene expression time series, the method presented here can easily incorporate other representations, such as polynomial trend models [30]. Another important character of the method here presented is its reliance on an explicit statistical model of gene expression dynamics. This reliance makes it possible to use standard statistical techniques to assess the goodness of fit of the resulting model and validate the underlying assumptions. This method is implemented in a computer program, called CAGED (Cluster Analysis of Gene Expression Dynamics). This chapter will first describe the theoretical framework and the clustering method; it will then summarize the functionalities of the computer program implementing this method; and it will finally illustrate the use of the program on a publicly available database of gene expression dynamics.

1.2 Methods

The design of a microarray experiment exploring the temporal behavior of a set of J genes usually consists of a set of n microarrays, each measuring the gene expression level x_{jt} of a set of genes at a time point t. We regard the expression values of a single gene across these measurements as a *time*

series $S_j = \{x_{j1}, \ldots x_{jt}, \ldots x_{jn}\}$, and the entire experiment as a set of J time series $S = \{S_1, S_2, ..., S_J\}$, generated by an unknown number of stochastic processes. The task here consists of merging these expression profiles into groups (*clusters*), so that each cluster groups the time series generated by the same process. Our method searches for the most probable set of processes responsible for the observed gene expression time series. Our clustering method has two components: a stochastic description of a set of clusters, from which we derive a probabilistic scoring metric to rank the different ways of combining gene expression profiles, and a heuristic search procedure to efficiently explore the space of these combinations.

1.2.1 Modeling Time

CAGED takes a Bayesian approach to clustering and searches for the most probable set of processes responsible for the observed data. To do so, CAGED looks for the set of clusters (i.e. ways of combining genes on the basis of their expression values along time) with maximum posterior probability. The critical point, here, is that the expression measurements of each gene along time are not independent and identically distributed. CAGED represents this dependency using autoregressive equations. More formally, a stationary time series $S_j = \{x_{j1}, \ldots x_{jt}, \ldots x_{jn}\}$ of continuous values follows an autoregressive model of order p, say AR(p), if the value of the series at time $t > p$ is a linear function of the values observed in the previous p steps. We can describe this model in matrix form as

$$x_j = F_j \beta_j + \epsilon_j \tag{1.1}$$

where x_j is the vector $(x_{j(p+1)}, \ldots, x_{jn})^T$, F_j is the $(n-p) \times q$ regression matrix whose tth row is $(1, x_{j(t-1)}, \ldots, x_{j(t-p)})$, for $t > p$, and $q = p + 1$. The elements of the vector $\beta_j = \{\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp}\}$ are the autoregressive coefficients and $\epsilon_j = (\epsilon_{j(p+1)}, \ldots, \epsilon_{jn})^T$ is a vector of uncorrelated errors that we assume normally distributed, with expected value $E(\epsilon_{jt}) = 0$ and variance $V(\epsilon_{jt}) = \sigma_j^2$, for any t. The value p is the autoregressive order and specifies that, at each time point t, x_{jt} is independent of the past history before p, given the previous p steps. The time series is stationary if it is invariant by temporal translations. Formally, stationarity requires that the coefficients β_{jk} are such that the roots of the polynomial $f(u) = 1 - \sum_{k=1}^{p} \beta_{jk} u^h$ have moduli greater than unity. The model in Equation (1.1) represents the evolution of the process around its mean μ_j, which is related to the β_j coefficients by the equation $\mu_j = \beta_{j0}/(1 - \sum_{k=1}^{p} \beta_{jk})$. In particular, μ_j is well defined as long as $\sum_{k=1}^{p} \beta_{jk} \neq 1$. When the autoregressive order $p = 0$, the series S_j becomes a sample of independent observations from a normal distribution with mean $\mu_j = \beta_{j0}$ and variance σ_j^2.

Given a time series S_j, we wish to estimate the parameters β_j and σ_j^2 from the data. The Bayesian estimation of β_j and σ_j^2 consists of updating

their prior distribution into a posterior distribution by Bayes' Theorem. So, with $f(\beta_j, \sigma_j^2)$ denoting the prior density, we need to compute the posterior density

$$f(\beta_j, \sigma_j^2 | x_j, p) = \frac{f(x_j|\beta_j, \sigma_j^2, p) f(\beta_j, \sigma_j^2)}{f(x_j|p)}.$$

The likelihood function $f(x_j|\beta, \tau, p)$ is

$$f(x_j \,|\, \beta_j, \sigma_j^2) = \sqrt{\frac{\sigma_j^{2n}}{(2\pi)^n}} \exp\left(-\frac{(x_j - F_j\beta_j)^T (x_j - F_j\beta_j)}{2\sigma_j^2} \right) \tag{1.2}$$

and it is in fact a function of the first p values of the series, but we omit the explicit dependence for simplicity of notation. As prior distributions of β_j and σ_j^2, we assume the family of improper distributions with density $f(\beta_j, \sigma_j^2) \propto {\sigma_j^2}^{\gamma}$, for $\sigma_j^2 > 0$ and $\gamma \geq 0$. When $\gamma = 0$, this formula represents the uniform prior, when $\gamma = 1$ it is the typical reference prior, and when $\gamma = 2$, it becomes the so-called Jeffreys prior [4]. The quantity $f(x_j|p)$ is the averaged likelihood function in which β_j and σ_j^2 are integrated out. Suppose F_j is of full rank, and define

$$\hat{\beta}_j = (F_j^T F_j)^{-1} F_j^T x_j \tag{1.3}$$

$$\text{RSS}_j = x_j^T (I_n - F_j (F_j^T F_j)^{-1} F_j^T) x_j. \tag{1.4}$$

Then, the quantity $f(x_j|p)$ is

$$f(x_j \,|\, p) = \frac{\left(\frac{\text{RSS}}{2}\right)^{(q+\gamma-n)/2} \Gamma\left(\frac{n-q-\gamma}{2}\right)}{(2\pi)^{(n-q)/2} \det(F_j^T F_j)^{1/2}} \tag{1.5}$$

where q is the dimension of β_j. The posterior distribution of σ_j^2 and β_j is normal-inverse-gamma, with

$$\beta_j \,|\, x_j, \sigma_j^2 \sim N(\hat{\beta}_j, [\tau(F_j^T F_j)]^{-1}) \tag{1.6}$$

$$1/\sigma_j^2 \,|\, x_j \sim \text{Gamma}\left(\frac{\text{RSS}_j}{2}, \frac{n-q-\gamma}{2}\right), \tag{1.7}$$

and we define the density function of a Gamma(a, b) by $f(\tau) = a^b \Gamma(b)^{-1} \tau^{b-1} \exp(-\tau a)$. The posterior distributions (1.6) and (1.7) are proper as long as F_j is of full rank, and $n > q + \gamma$.

1.2.2 Probabilistic Scoring Metric

We are actually interested in finding the clustering model with the highest posterior probability given the observed gene expression time series. Since all clustering models are compared with respect to the same data, and we assume uniform prior distributions, the posterior probability of a clustering

model becomes proportional to its *marginal likelihood*. We describe a set of c clusters of gene expression time series as a statistical model M_c, consisting of c autoregressive models with coefficients β_k and variance σ_k^2. Each cluster C_k groups the time series data of J_k genes that are jointly modeled as

$$x_k = F_k \beta_k + \epsilon_k$$

where the vector x_k and the matrix F_k are defined by stacking the J_k vectors x_{kj} and regression matrices F_{kj}, one for each time series, as follows

$$x_k = \begin{pmatrix} x_{k1} \\ \vdots \\ x_{kJ_k} \end{pmatrix} \quad F_k = \begin{pmatrix} F_{k1} \\ \vdots \\ F_{kJ_k} \end{pmatrix}.$$

Note that we now label the vectors x_j assigned to the same cluster C_k with the double subscript kj, and k denotes the cluster membership, so that $\sum_k J_k = J$, and J is the total number of genes. The vector ϵ_k is the vector of uncorrelated errors with zero expected value and constant variance σ_k^2. In principle, given a set of possible clustering models, the task is to rank them according to their posterior probability. The posterior probability of each clustering model M_c is:

$$P(M_c|x) \propto P(M_c) f(x|M_c)$$

where $P(M_c)$ is the prior probability of M_c, x consists of all the time series data $\{x_k\}$, and the quantity $f(x|M_c)$ is the marginal likelihood. The marginal likelihood $f(x|M_c)$ is the solution of the integral

$$\int f(x|\theta) f(\theta|M_c) d\theta$$

where θ is the vector of parameters specifying the clustering model M_c, $f(\theta|M_c)$ is its prior density, and $f(x|\theta)$ is the overall likelihood function. By independence of the series assigned to different clusters, the overall likelihood function is

$$f(x \mid \theta) = \prod_{k=1}^{c} p_k^{m_k} f(x_k \mid F_k, \beta_k, \sigma_k^2)$$

where p_k is the marginal probability that a time series is assigned to the cluster C_k. We assume independent uniform prior distributions on the model parameters β_k, σ_k^2 and a symmetric Dirichlet distribution on the parameters p_k, with hyper-parameters $\alpha_k \propto p_k$ and overall precision $\alpha = \sum_k \alpha_k$. By independence of the time series conditional on the cluster membership, and parameter independence, the marginal likelihood $f(x|M_c)$ can be computed as

$$f(x \mid M_c) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+m)} \prod_{k=1}^{c} \frac{\Gamma(\alpha_k+m_k)}{\Gamma(\alpha_k)} \frac{\left(\frac{\mathrm{RSS}_k}{2}\right)^{(q+\gamma-n_k)/2} \Gamma\left(\frac{n_k-q-\gamma}{2}\right)}{(2\pi)^{(n_k-q)/2} \det(F_k^T F_k)^{1/2}} \tag{1.8}$$

where n_k is the dimension of the vector x_k, and $\mathrm{RSS}_k = x_k^T(I_n - F_k(F_k^T F_k)^{-1}F_k^T)x_k$ is the residual sum of squares in cluster C_k. When all clustering models are a priori equally likely, the posterior probability $P(M_c|x)$ is proportional to the marginal likelihood $f(x|M_c)$, which becomes our probabilistic scoring metric.

1.2.3 Heuristic Search

The Bayesian approach to the clustering task is to choose the model Mc with maximum posterior probability. As the number of clustering models grows exponentially with the number of time series, we use an agglomerative, finite-horizon search strategy which iteratively merges time series into clusters. The procedure starts by assuming that each of the J observed time series is generated by a different process. Thus, the initial model M_J consists of J clusters, one for each time series, with score $f(x|M_J)$. The next step is the computation of the marginal likelihood of the $J(J-1)$ models in which two of the J series are merged into one cluster. The model M_{J-1} with maximal marginal likelihood is chosen and, if $f(x|M_J) \geq f(x|M_{J-1})$, no merging is accepted and the procedure stops. If $f(x|M_J) < f(x|M_{J-1})$, 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 $J-1$ clusters, consisting of the remaining $J-2$ time series and the cluster C_k.

Although the agglomerative strategy makes the search process feasible, the computational effort can still be extremely demanding when the number J 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 of the model. The heuristic search starts by computing the $J(J-1)$ pair-wise similarity measures of the J time series and selects the model M_{J-1} in which the two closest time series are merged into one cluster. If $f(x|M_{J-1}) > f(x|M_J)$, the merging is accepted, the two time series are merged into a single cluster, an *average profile* of this cluster is computed by averaging the two observed time series, and the procedure is repeated on the new set of $J-1$ time series, containing the new cluster profile. If this merging is rejected, the procedure is repeated on pair of time series with decreasing degree of similarity, until an acceptable merging is found. If no acceptable merging is found, the procedure stops. Note that the decision of merging two clusters is actually made on the basis of the posterior probability of the model and that the

similarity measure is only used to improve efficiency and to limit the risk of falling into local maxima.

CAGED includes several similarity measures to assess the similarity of two time series, both model-free — such as Euclidean distance, correlation and lag-correlation — and model-based — such as the symmetric Kullback-Leibler distance. This distance is computed for every pair of parameter vectors β_k, β_j, using the normal distribution of each β_k, conditional on the cluster variance σ_k^2. The variance is then replaced by the posterior estimate. For a clustering model specifying c clusters C_k, with matrices F_k and data x_k, the conditional posterior distribution of $\beta_k|x_k, \sigma_k^2$ is $N(\hat{\beta}_k, \sigma_k^2[(F_k^T F_k)]^{-1})$, and the symmetric Kullback-Liebler divergence between conditional distributions of β_k, β_j is

$$\begin{aligned} d_{kj} &= \int f(\beta|x_k, F_k, \sigma_k^2) \log \frac{f(\beta|x_k, F_k, \sigma_k^2)}{f(\beta|x_j, F_j, \sigma_j^2)} d\beta \\ &+ \int f(\beta|x_j, F_j, \sigma_j^2) \log \frac{f(\beta|x_j, F_j, \sigma_j^2)}{f(\beta|x_k, F_k, \sigma_k^2)} d\beta \end{aligned}$$

where β denotes the generic integration variable, and $f(\beta|x_k, F_k, \sigma_k^2)$ is the density function of a distribution $N(\hat{\beta}_k, \sigma_k^2[(F_k^T F_k)]^{-1})$.

Model-free distances are calculated on the raw data. Since the method uses these similarity measures as heuristic tools rather than scoring metrics, we can actually assess the efficiency of each of these measures to drive the search process toward the model with maximum posterior probability. In this respect, the Euclidean distance of two time series $S_i = \{x_{i1}, \ldots, x_{1n}\}$ and $S_j = \{x_{j1}, \ldots, x_{jn}\}$, computed as

$$D_e(S_i, S_j) = \sqrt{\sum_{t=1}^{n} (x_{it} - x_{jt})^2},$$

performs best on the short time series of our data set. This finding is consistent with the results of [10] claiming a better overall performance of Euclidean distance in standard hierarchical clustering of gene expression profiles.

1.2.4 Statistical Diagnostics

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 as $\hat{\beta}_k = (F_k^T F_k)^{-1} F_k^T x_k$, while $\hat{\sigma}_k^2 = \text{RSS}_k/(n_k - q - \delta)$ is the estimate of the within-cluster variance σ_k^2. The parameter estimates can be used to compute the fitted values for the series in each cluster as $\hat{x}_{kj} = F_{kj}\hat{\beta}_k$, from which we compute the standardized residuals

$r_{kj} = (x_{kj} - \hat{x}_{kj})/\hat{\sigma}_k$. 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 residuals histogram per cluster, are 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 fitted versus observed values and of fitted values versus standardized residuals in each cluster provide further diagnostics. To choose the best autoregressive order, we repeat the clustering for $p = 0, 1, \ldots, w$, for some preset w — by using the same p for every clustering model — and compute a goodness-of-fit score defined as

$$s = c(q+\gamma) - (1+\log(2\pi)) \sum_k n_k + \sum n_k \log(n_k - q - \gamma) - \sum n_k \log(\text{RSS}_k)$$

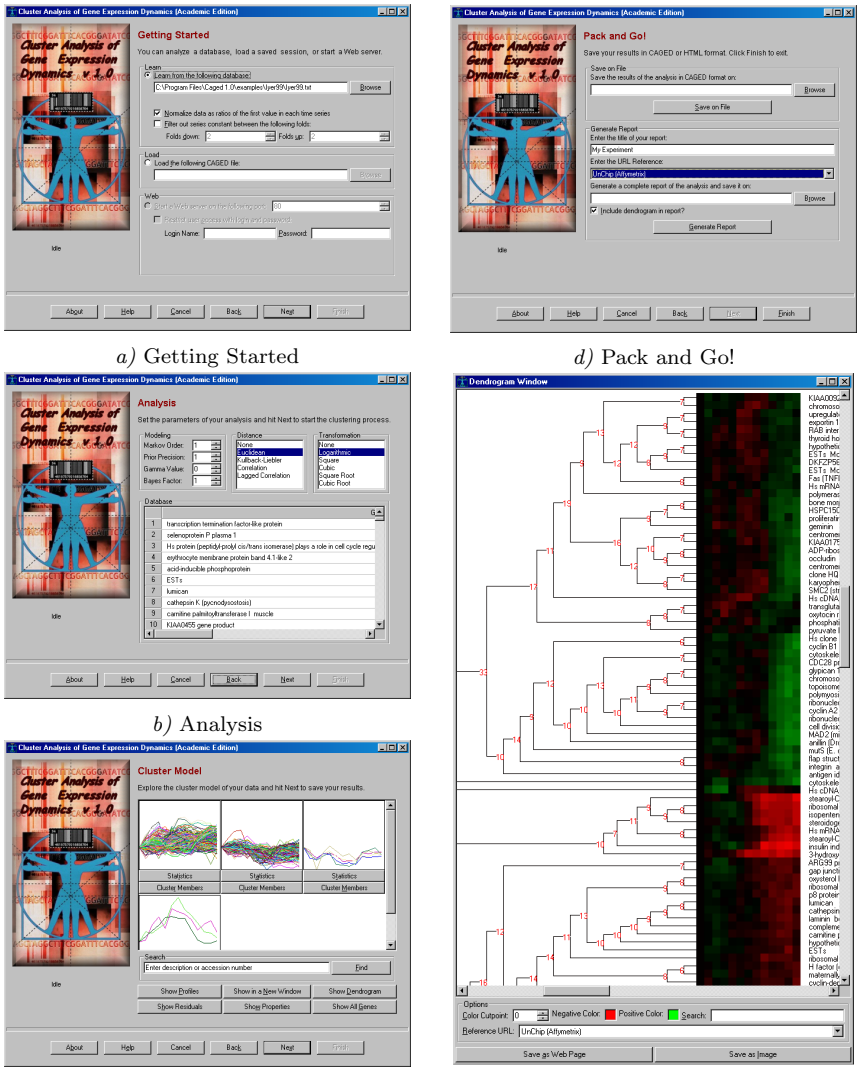
where c is the number of clusters, n_k is the size of the vector x_k in C_k, $q = p + 1$, p is the autoregressive order, and RSS_k is the residual sum of squares of cluster C_k. This score is derived by averaging the log-scores cumulated by the series assigned to each clusters, and details are in Appendix. The resulting score trades off model complexity — measured by the quantity $cq + \sum_k n_k \log(n_k - q)$ — with lack of fit — measured by the quantity $\sum n_k \log(\text{RSS}_k)$, and it generalizes the well known AIC goodness-of-fit criterion of [2] to a set of autoregressive models. We then choose the clustering model with the autoregressive order p that maximizes this goodness-of-fit score.

1.3 Software

The method described in the previous section is implemented in a computer program called CAGED (Cluster Analysis of Gene Expression Dynamics). The program runs under the various version of Microsoft Windows and the graphic user interface is implemented as a *Wizard interface*. The Wizard interface is composed by subsequent screens guide the user through the steps of analyzing a database of gene expression dynamics. This section describes the use of the program screen by screen.

1.3.1 Screen 0: Welcome Screen

When the program is started, a Welcome screen appears, containing a welcome message and a summary of the End-User License. The bottom of the screen contains six buttons, which will remain present in all subsequent screens. The buttons are, from left to right, an About button — containing some basic information about the program — an Help button — evoking an on-line help file system — a Cancel button — to quit the program at any



a) Getting Started

d) Pack and Go!

b) Analysis

c) Cluster Model

e) Dendrogram

Figure 1.1. Snapshots of the main screens of CAGED 1.0.

time — a Back button — to move backward in the succession of screens — a Next button — to move forward — and a Finish button — which will become active at the end of the analysis process. By hitting the Next button, the user is taken to the next screen.

1.3.2 Screen 1: Getting Started

The second screen, shown in Figure 1.1*a* is divided in three parts. The first part allows the user to load a database for the analysis. CAGED expects files in ASCII Tab delimited format. The file format follows the rules of most statistical genomics analysis program: gene expression time series are reported along the rows of the database. The first column contains a description of the gene, the second column an accession number to a internet-available database (such as GenBank, Unigene, or an Affymetrix accession number), and the following columns report the expression values of such genes across time, that is, each column reports the expression value of a gene in a particular microarray. The first row reports, except for the first two columns, a label denoting the experimental conditions and, in this case, it is expected to report a time stamp. Below the loading dialog, the user is given the option to convert the absolute expression values of each gene into ratios between each time point value and the value recorded at first time point. This option can be particularly useful when the investigator is interested in the relative dynamics of each gene expression time series rather than the absolute values. For microarray platforms measuring relative gene expression values, such as cDNA or oligonucleotide microarrays, this is usually the preferred option. For platforms measuring absolute expression values, such as SAGE, the absolute expression values can be by themselves meaningful and this options is usually left unchosen. The user can also decide to filter out gene expression time series where a gene does not show at least one value change higher than a user-defined threshold. The second section of the screen allows the user to load a previously saved analysis session. In this case, the button Next will take the user directly to Screen 3 (Cluster Model). The last section of the screen allows the user to open a web-site on the current machine to serve the program over the internet.

1.3.3 Screen 2: Analysis

If the user has chosen to load a database for analysis, the Next button will display the Analysis screen, shown in Figure 1.1*b*. This screen allows the user to choose the statistical hypotheses to run the analysis, the distance to guide the heuristic search, and some optional data transformations. Statistical hypotheses are encoded by some model parameters in the Modeling panel in the top left corner of the screen. Here, the user can set the *order* p in equation 1.1, representing the length of the memory of the model. The user can also set the prior precision α and the parameter γ used to compute marginal likelihood (Equation 1.5) and the cluster score (Equation 1.8). These three parameters are usually dealt with using sensitivity analysis: the user will run various analysis using different settings of the parameters (typically, 0, 1 and 2 for all three) to check the robustness of the cluster-

ing model with respect to these assumptions. The last parameter in the panel sets a threshold on the Bayes factor: the ratio between the marginal likelihood of two alternative models. Setting this parameter to δ will tell the program to merge two gene expression time series or two clusters if the marginal likelihood of the model resulting from such a merging is at least δ times larger than the marginal likelihood of the model in which the two gene expression time series or clusters are kept separated This screen also offers the user the opportunity to choose the distance used by the heuristic search process and described in the previous section. Since this measure is meant to simply guide the search process, the best distance will be the one leading to the clustering model with the largest marginal likelihood. The program also offers the possibility to impose some optional transformations over the data, ranging from the common logarithmic transformation to some standard power transformations.

1.3.4 Screen 3: Cluster Model

When the user hits the Next button from the Analysis screen, the program will run the clustering process described in the previous section. The results of this analysis are displayed in the Cluster Model screen, shown in Figure 1.1*c*. The model is primarily described by a set of plots collecting the gene time series members of each cluster. The members and the basic statistical properties of each cluster can be viewed by clicking the button below each plot. The screen also offers the display of a dendrogram with a binary tree representing the clustering order of the gene expression time series. The nodes of the tree in this representation report the Bayes factor of the merging, i.e. how many times the marginal likelihood of the model is increased by the merging of the two sub-trees with respect to the model in which the two sub-trees are kept separated. General properties of the clustering model are displayed by the Property window, which also contains a validation program to list which repeated genes fall in the same cluster. An example of such a tree is given the dendrogram shown in Figure 1.1*e*.

1.3.5 Screen 4: Pack and Go!

The last screen of the program, shown in Figure 1.1*d*, allows the user to save the results of the analysis in two formats. The first option saves the analysis results in CAGED format, so that the results will be loadable and viewable through the program. The second option generates a complete report, including images and statistical diagnostics, in HTML format, which can be posted on the World Wide Web or loaded in some word processing program. An optional dialog allows the user to insert a URL template to link the Accession numbers in the second column of the input data with the appropriate internet resource database, such as Genbank, Unigene, or an Affymetrix accession numbers repository.

1.4 Application

This section illustrates the properties of this method and the use of CAGED using a data set of gene expression dynamics. Iyer *et al.* [13] report the results of a study of the temporal deployment of the transcriptional program underlying the response of human fibroblasts to serum. The study uses two-dye cDNA microarrays to measure the changes of expression levels of 8613 human genes over 24 hours, at not equally spaced time points. The actual data described in the study comprise a selection of 517 genes whose expression level changed in response to serum stimulation. At the time of their original publication, 238 genes were unknown expressed sequence tags (ESTs). We relabeled the data set using the most recent UniGene database (http://www.ncbi.nlm.nih.gov/UniGene) and 45 genes were left unknown. The UniGene classification was used to identify repeated genes in the data set. We found that 20 genes appear at least twice in the data set and were not known to be part of the same UniGene cluster at the time of the original report.

1.4.1 Analysis

The analysis of a database of gene expression dynamics typically involves more than one run of the program. Statistical diagnostics play the fundamental role of assessing the best fitting model and, in general, validating the soundness of the conclusions. After the database is loaded in Screen 1, with no filter on the minimum required change, Screen 2 contains all the parameters to set in order to explore different statistical hypotheses. As the data in questions are actually ratios of two conditions, we choose to log transform the data in order to treat symmetrically positive and negative fold changes.

For this analysis, we choose a uniform prior ($\gamma = 0$) and a minimal prior precision ($\alpha = 1$), the default values in the Modeling panel in the top left corner of Screen 2. Also, because the time point were not equally spaced, we assumed that the spacing of the time points was irrelevant. In other words, intervals of different lags were taken as equally informative about the underlying process. We ran the clustering algorithm with four autoregressive orders $p = 0, 1, 2, 3$ and the similarity measures available in the Distance panel and described in the Method section. Since the role of the distance is simply to guide the search process, we can assess the best working distance by checking the *Marginal Likelihood* of the resulting model in the Properties window in Screen 3. For all values of p, Euclidean distance gave the best results, i.e. the model with highest marginal likelihood. At this point, we must choose the best *Model 0rder p*, the first parameter in Modeling panel. The number of clusters found for $p = 0, 1, 2, 3$ varied between 4 ($p = 0, 1$) and 3 ($p = 2, 3$). To choose a clustering model among these four, we used the goodness-of-fit score called *Autoregressive*

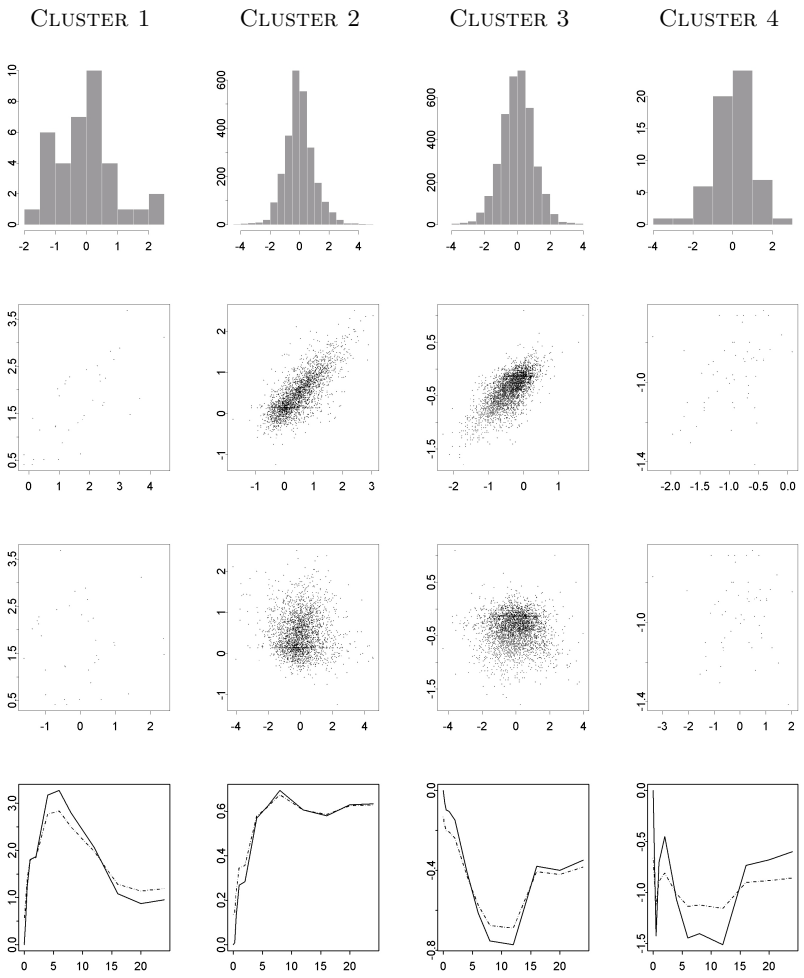


Figure 1.2. Diagnostic plots for the clustering model identified by the method when the autoregressive order is $p = 1$. The first row reports histogram of standardized residuals. The second row reports the scatter plot of fitted values versus observed values. The third row shows the scatter plot of fitted values versus standardized residuals. The fourth row displays the four cluster average profiles (continuous lines) computed as averages of the time observed time series in each cluster and the averages of the fitted time series in each cluster (dashed lines). In these plots, the x-axis reports time in hours.

Score in the Properties window of Screen 3. The scores for the four models were, for increasing p, 10130.78, 13187.15, 11980.38, and 11031.12, and the model with order $p = 1$ was therefore selected. This model merges the 517 gene time series into four clusters of 3, 216, 293, and 5 time series,

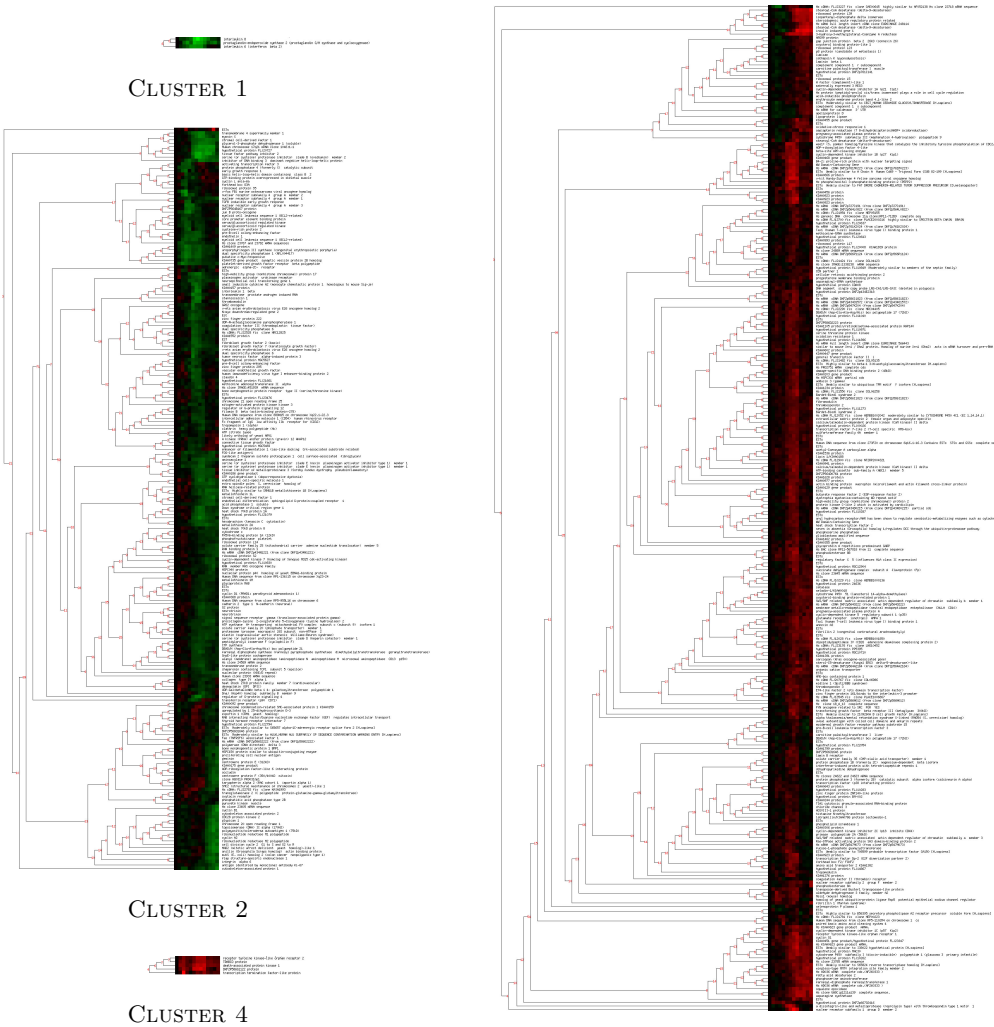


Figure 1.3. Binary tree (dendrogram) and labeled gene expression display showing the clustering model obtained by our method on the data reported in Iyer *et al.* [13]. The numbers on the branch points of the tree represent how many times the merging of two series renders the model more probable. The model identifies four distinct clusters containing 3 (Cluster 1), 216 (Cluster 2), 293 (Cluster 3) and 5 (Cluster 4) time series.

with estimates of the autoregressive coefficients and within-cluster variance $\hat{\beta}_{10} = 0.518; \hat{\beta}_{11} = 0.708; \hat{\sigma}_1^2 = 0.606$ in cluster 1, $\hat{\beta}_{20} = 0.136; \hat{\beta}_{21} = 0.776; \hat{\sigma}_2^2 = 0.166$ in cluster 2, $\hat{\beta}_{30} = -0.132; \hat{\beta}_{31} = 0.722; \hat{\sigma}_3^2 = 0.091$ in cluster 3, and $\hat{\beta}_{40} = -0.661; \hat{\beta}_{41} = 0.328; \hat{\sigma}_4^2 = 0.207$ in cluster 4.

1.4.2 Statistical Diagnostics

In the selected model, merging any of these clusters decreases the posterior probability of the clustering model of at least 10.05 times, a *strong* evidence in favor of their separation [15]. The symmetry of the standardized residuals in Figure 1.2, together with the lack of any significant patterns in the scatter plot of the fitted values versus the standardized residuals and the closeness of fitted and observed values, suggests that AR(1) models provide a good approximation of the processes generating these time series. This impression is further reinforced by the averages of the fitted time series in each cluster, shown in Figure 1.2, which follow closely their respective cluster average profiles. In CAGED, the button *Show Residuals* on Screen 4 shows the residuals plot and their basic statistics for each cluster.

1.4.3 Understanding the Model

The most evident difference between the model in Figure 1.3 and the model obtained in the original article by visual inspection [13] is the number of clusters: our method detects four distinct clusters, characterized by the autoregressive models described above, while hierarchical clustering merges all 517 genes into a single cluster and leaves it to the investigator to identify subgroups by visual inspection. For example, Iyer *et al.* identify, by visual inspection, eight subgroups of genes — labeled A, B, C,..., I, J — from eight large contiguous patches of color. With the exception of a few genes, our cluster 2 merges the subgroups of time series labeled as D, E, F, G, H, I and J, and cluster 3 merges the majority of time series assigned to subgroups A, B and C. Interestingly enough, the members of subgroups A, B and C differ, on average, by one single temporal value and, similarly, members of groups D and G differ by a single temporal value, as well as F, H, J and I. The assignment of time series to different groups on the basis of one temporal points could be a consequence of the fact that the human eye tends to overfit.

Across the four clusters, both average profiles and averages of the fitted time series appear to capture different dynamics. Our cluster 1 collects the temporal patterns of three genes — interleukin 8, prostaglandin-endoperoxide synthase 2, interleukin 6 (interferon beta 2). These time series were assigned by [13] to the subgroups F, I, and J, respectively. Cluster 4 collects the time series of five genes — receptor tyrosine kinase-like orphan receptor, TRABID protein, death-associated protein kinase, DKFZP586G1122 protein, transcription termination factor-like protein. Three of these time series were assigned by [13] to the A and B subgroups. These two smaller clusters — cluster 1 and 4 — are particularly noteworthy because they illustrate how our method automatically identifies islands of particular expression profiles. The first of these two clusters merges cytokines involved in the processes of the inflammatory response

Gene name	Group	Cluster
serum/glucocorticoid regulated kinase	J, J	2, 2
pre-B-cell colony-enhancing factor	J, NA	2, 2
myeloid cell leukemia sequence 1	J, J	2, 2
serine proteinase inhibitor	I, I	2, 2
stromal cell-derived factor 1	NA, H	2, 2
neurotrimin	H, H	2, 2
dual specificity phosphatase 6	F, F	2, 2
v-ets avian erythroblastosis virus E26	F, F	2, 2
ESTs	H, H	2, 2
DKFZP566O1646 protein	B, A	2, 3
stearoyl-CoA desaturase	C, C, C	3, 3, 3
pregnancy-associated plasma protein A	C, C	3, 3
DEAD/H box polypeptide 17	B, B	3, 3
KIAA0923 protein	B, B, B, B	3, 3, 3, 3
WW Domain-Containing Gene	B, B	3, 3
Bardet-Biedl syndrome 2	B, B	3, 3
calcium/calmodulin-dependent protein kinase	B, B	3, 3
Tax1 (human T-cell leukemia virus type I)	A, B	3, 3
AD036 protein	A, A	3, 3
DKFZp586I1823	A, A	3, 3

Table 1.1. Assignment of gene repeats to subgroups by Iyer *et al.* [13] (column 2) and by our method (column 3). The first column reports the UniGene name of the repeated genes. Subgroups in column 2 are identified by A to J letters,with NA denoting a placement outside the eight clusters identified by the authors.

and chemotaxis, and the signal transduction and cell-cell signaling underlying these processes. The cluster includes Interleukin 8, Interleukin 6, and Prostaglandin-endoperoxide synthase 2, which catalyzes the rate-limiting step in the formation of inflammatory prostaglandins. The second small cluster includes genes that are known to be involved in the cell-death/apoptosis processes. It includes kinases and several transcription factors reported to be involved in these processes. The cluster includes receptor tyrosine kinase-like orphan receptor 2, TRAF-binding protein domain, and Death-associated protein kinase. The cluster also includes the transcription termination factor-like protein, which plays a central role in the control of rRNA and mRNA synthesis in mammalian mitochondria [8], and DKFZP586G1122 protein, which has unknown function but has strong homology with murine zinc finger protein Hzf expressed in hematopoiesis.

The number of clusters found by our algorithm is directly inferred from the data, which also provide evidence in favor of a temporal dependency of the observations: the goodness-of-fit score of the AR(0) clustering model, where the observations are assumed to be marginally independent, is lower than the goodness-of-fit score of the AR(1) clustering model, which assumes that each observation depends on its immediate predecessor. The allocation of the 20 repeated genes in the data set seems to support our claim that identifying subgroups of genes by visual inspection may overfit the data:

with the exception of the two repeats of the DKFZP566O1646 protein, our model assigns each group of repeated genes to the same cluster, whereas four of the repeated genes are assigned to different subgroups in [13]. Details are shown in Table 1.1. The risks of overfitting by visual inspection can be easily appreciated by looking at the color patterns in Figure 1.3. As the dendrogram is built, genes with highly similar temporal profiles are merged first, thus producing subtrees with similar patterns of colors. However, according to our analysis, the data do not provide enough evidence to conclude that such subtrees contain time series generated by different processes and they are therefore merged into a single cluster.

An example of this phenomenon is shown in detail in the dendrogram displayed in Figure 1.1*e*, which enlarges part of the dendrogram in Figure 1.3 around the breaking point between cluster 1 and cluster 2. The first 28 time series from the top of the image, which appear to be more homogeneous by visual inspection, are merged in a subtree showing that at each step of the iterative procedure, merging the time series induces a model more likely than the model determined by not merging them. Similarly, the next 18 time series are merged in a subtree labeled, at the top, by a Bayes factor of 10, in logarithmic scale. These two subtrees are then merged in a larger tree, with a Bayes factor of $exp(33)$, meaning that the model in which the two subtrees are merged together is $exp(33)$ times more likely than the model in which these subtrees are taken as two separate clusters. The dramatic change of the color patterns two time series below the end of this tree marks the beginning of cluster 2.

1.5 Conclusions

The analysis of gene expression data collected along time is at the basis of critical applications of microarray technology. This contribution addresses a fundamental property of temporal data — their directed dependency along time — in the context of cluster analysis. We have represented the dependency of temporal observations as autoregressive equations and we have taken a Bayesian approach to the problem of selecting the number and members of clusters. To explore the exponential number of possible clustering models, we have devised a heuristic search procedure based on pairwise distances to guide the search process. In this way, our method retains the important visualization capability of traditional distance-based clustering and acquires a principled measure to decide when two time series are different enough to belong to different clusters. It is worth noting that the measure here adopted, the posterior probability of the clustering model, takes into account all the available data and such a global measure also offers a principled way to decide whether the available evidence is sufficient to support an empirical claim. Our analysis shows that sometimes

the available evidence is not sufficient to support the claim that two time series are generated by two distinct processes. Figure 1.3 shows contiguous patches of colors, but the posterior probability of the model does not support the claim that these subgroups are sufficiently distinct to be viewed as distinct processes. This finding has interesting implications for experiment design and sample size determination, because it allows the analyst to assess whether the available information is sufficient to support significant differentiations among gene profiles and, if necessary, collect more data. A third feature of the method presented here is the reliance of the clustering process on an explicit statistical model. Contrary to other approaches [12], our method builds the clustering model using the parametric content of the statistical model rather than providing statistical content to an established clustering model. This stochastic content allows us to use standard statistical techniques to validate the goodness of fit of the clustering model, as illustrated at the end of the Application section. While the biological validation of microarray experiments plays a critical role in the development of modern functional genomics, practical considerations often limit this validation to few genes, while the claims and the scope of a microarray experiment involve thousands. A proper use of available statistical diagnostics provides analytical tools to independently assess the global validity of a clustering model.

Autoregressive equations are very simple representations of process dynamics and they rely on the assumption that the modeled time series are stationary. Our reason to choose this representation is its simplicity: since the time series of gene expression experiments are typically very short, more sophisticated representations could be prone to overfitting. Stationarity conditions can be checked using the method described at the end of the Methods section but, both in the data analyzed here and in our general experience, the clustering process seems to be largely unaffected by the presence of non stationary time series. In principle, however, other representations can be integrated within the Bayesian framework described in this paper. The forthcoming version of CAGED will include, besides autoregressive models, also polynomial trend models, to tackle the problem of shorter time series and the explicit dependency upon time, and state-space models to handle comparative experiments along time and multiple arrays.

Acknowledgments

Authors thank Stefano Monti (Whitehead Institute) and Alberto Riva (Harvard Medical School) for their insightful comments on an early draft of this article. This research was supported in part by the National Science Foundation (Bioengineering and Environmental Systems Division — Biotechnology) under contract ECS-0120309.

References

[1] J. Aach and G.M. Church.Aligning gene expression time series with time warping algorithms.*Bioinformatics*, 17:495–508, 2001.

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

[3] O. Alter, P. O. Brown, and D. Botstein.Singular value decomposition for genome-wide expression data processing and modeling.*Proc Natl Acad Sci USA*, 97:10101–10106, 2000.

[4] J. M. Bernardo and A. F. M. Smith.*Bayesian Theory.*Wiley, New York, NY, 1994.

[5] G. E. P. Box and G. M. Jenkins.*Time Series Analysis: Forecasting and Control.*Holden-Day, San Francisco, CA, 1976.

[6] A.J. Butte, P. Tamayo, D. Slonim, T.R. Golub, and I.S. Kohane.Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks.*Proc Natl Acad Sci USA*, 97:12182–12186, 2000.

[7] 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.

[8] P. Fernandez-Silva, F. Martinez-Azorin, V. Micol, and Attardi G.The human mitochondrial transcription termination factor (mterf) is a multizipper protein but binds to DNA as a monomer, with evidence pointing to intramolecular leucine zipper interactions.*Embo J*, 16(5):1066–79, 1997.

[9] T. Gilovich, R. Vallone, and A. Tversky.The hot hand in basketball: On the misperception of random sequences.*Cognitive Psychol*, 17:295–314, 1985.

[10] R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, 1 M. L. Loh H. Coller, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander.Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.*Science*, 286:531–537, 1999.

[11] I.J. Haimowitz, Le P.P., and Kohane I.S.Clinical monitoring using regression-based trend templates.*Artif Intell Med*, 7(4):471472, 1995.

[12] N.S. Holter, A. Maritan, M. Cieplak, N.V. Fedoroff, and J.R. Banavar.Dynamic modeling of gene expression data.*Proc Natl Acad Sci USA*, 98(4):1693–1698, 2001.

[13] V.R. Iyer, M.B. Eisen, D.T. Ross, T. Schuler, G. Moore, J.M. Lee, J.C. Trent, L.M. Staudt, J. Hudson, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and Brown P.O.The transcriptional program in the response of human fibroblasts to serum.*Science*, 283:83–7, 1999.

[14] D. Kahneman, P. Slovic, and A. Tversky.*Judgment under Uncertainty: Hueristic and Biases.*Cambridge University Press, New York, NY, 1982.

[15] R. E. Kass and A. Raftery.Bayes factors.*J Amer Statist Assoc*, 90:773–795, 1995.

[16] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown.Expression monitoring by hybridization to high-density oligonucleotide arrays.*Nat Biotech*, 14:1675–1680, 1996.

[17] I.S. Lossos, A.A. Alizadeh, M.B. Eisen, W.C. Chan, P.O. Brown, D. Botstein, L.M. Staudt, and R. Levy.Ongoing immunoglobulin somatic mutation in germinal center b cell-like but not in activated B cell-like diffuse large cell lymphomas.*Proc Natl Acad Sci USA*, 97(18):10209–10213, 2000.

[18] D.J.C. MacKay.Bayesian interpolation.*Neural Comput*, 4:415–447, 1992.

[19] M. Ramoni, P. Sebastiani, and P. R. Cohen.Multivariate clustering by dynamics.In *Proceedings of the 2000 National Conference on Artificial Intelligence (AAAI-2000)*, pages 633–638, San Francisco, CA, 2000. Morgan Kaufmann.

[20] M. Ramoni, P. Sebastiani, and P. R. Cohen.Bayesian clustering by dynamics.*Mach Learn*, 47(1):91–121, 2002.

[21] M. Ramoni, P. Sebastiani, and I.S. Kohane.Cluster analysis of gene expression dynamics.*Proc Natl Acad Sci USA*, 2002.Published online before print on June 24.

[22] B.Y. Reis, A.S. Butte, and I.S. Kohane.Extracting knowledge from dynamics in gene expression.*J Biomed Inform*, 34(1):15–27, 2001.

[23] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown.Quantitative monitoring of gene expression patterns with a complementary DNA microarray.*Science*, 270:467–70, 1995.

[24] P. Sebastiani and M. Ramoni.Common trends in european school populations.*Research in Official Statistics*, 4(1):169–183, 2001.

[25] Y. Shahar, S. Tu, and M. Musen.Knowledge acquisition for temporal abstraction mechanisms.*Knowl Acquis*, 1(4):217236, 1992.

[26] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub.Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation.*Proc Natl Acad Sci USA*, 96:2907–2912, 1999.

[27] J.B. Tenenbaum and T. L. Griffiths.Generalization, similarity, and bayesian inference.*Behav Brain Sci*, 24(3), 2001.

[28] A. Tversky and D. Kahneman.Judgment under uncertainty: Heuristics and biases.*Science*, 185:1124–1131, 1974.

[29] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi.Large-scale temporal gene expression mapping of central nervous system development.*Proc Natl Acad Sci USA*, 95:334–339, 1998.

[30] M. West and J. Harrison.*Bayesian Forecasting and Dynamic Models*.Springer, New York, NY, 1997.