A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves

Nicholas A. Heard, Christopher C. Holmes and David A. Stephens *

ABSTRACT

Malaria represents one of the major worldwide challenges to public health. A recent breakthrough in the study of the disease follows the annotation of the genome of the malaria parasite *Plasmodium falciparum* and the mosquito vector [1] *Anopheles*. Of particular interest is the molecular biology underlying the immune response system of Anopheles which actively fights against Plasmodium infection. This paper reports a statistical analysis of gene expression time profiles from mosquitoes which have been infected with a bacterial agent. Specifically, we introduce a Bayesian model-based hierarchical clustering algorithm for curve data to investigate mechanisms of regulation in the genes concerned; that is, we aim to cluster genes having similar expression profiles. Genes displaying similar, interesting profiles can then be highlighted for further investigation by the experimenter. We show how our approach reveals structure within the data not captured by other approaches. One of the most pertinent features of the data is the sample size, which records the expression levels of 2771 genes at six time points. Additionally, the time points are unequally spaced and there is expected non-stationary behaviour in the gene profiles. We demonstrate our approach to

*Nicholas Heard (Email: n.heard@imperial.ac.uk) and David Stephens (Email: d.stephens@imperial.ac.uk) are Lecturers in Statistics, Department of Mathematics, South Kensington Campus, Imperial College London, SW7 2AZ, U.K.; Christopher Holmes (Email: cholmes@stats.ox.ac.uk) is Lecturer in Statistics, Oxford Centre for Gene Function, Department of Statistics, University of Oxford, 1 South Parks Rd, Oxford, OX1 3TG, U.K. and MRC Mammalian Genetics Unit, Harwell, Oxford, OX11 0RD, U.K. The research in this paper was supported by grant 065822 from the Wellcome Trust.

[1] An organism that spreads an infectious disease

be readily implementable under these conditions, and highlight some crucial computational savings that can be made in the context of a fully Bayesian analysis.

KEYWORDS: Microarrays, gene expression profiles, Bayesian hierarchical clustering.

1 INTRODUCTION

The objective of this paper is to describe an exploratory statistical analysis of data relating to gene transcription in the immune response system of Anopheline mosquitoes. The data, illustrated by a colour 'heat map' in Figure 1, were collected using cDNA microarray technology, and represent the relative gene expression of a large number of genes measured at a small number of time points within mosquitoes following their infection with the bacterial agent *Salmonella typhi*.

These data are the expression levels of 2771 genes/sequence tags spotted on a cDNA array, with probes selected from a constructed cDNA library (see Dimopoulos et al., 2000, for details). Of these 2771 probes, 356 had known function. The expression profiles in Figure 1 relate to expression level measurements at $T = 6$ time points, taken at 1, 4, 8, 12, 18 and 24 hours after infection. The measurements were taken relative to unchallenged cells, and their rank values have been plotted in Figure 1 to prevent the scale being dominated by outliers. Further details of the experimental set-up, and pre-processing of the data, are given in Dimopoulos et al. (2002).

Our goal was to assist the experimental biologists by providing statistical procedures to detect and highlight structure within the data, by grouping together genes that exhibit similar dynamics. The aim was to identify groups of genes that appear to be *co-regulated*, that is, are controlled by the same biological mechanism, or form part of the same genetic *regulatory network*. Our task then was an exercise in statistical cluster analysis of longitudinal data. Pertinent aspects of the data include the large number of observations and the temporal dependence between observations in each series. More subtle aspects include the non-stationarity of many of the series and the non-uniform sampling intervals. These features led us to develop tailored methodology for clustering time course data using hierarchical Bayesian model-based procedures which we describe in detail in sections 2-5.

The analysis forms part of a much wider study by researchers in the Centre for Molecular Microbiology & Infection at Imperial College London who have been at the centre of recent and important genomic investigations in malaria (Christophides et al., 2002; Zdobnov et al., 2002;

Alphey et al., 2002; Carlton et al., 2002; Florens et al., 2002). The mosquito immune response system, the biological focus of the data in this paper, is of particular interest since it has been discovered (Dimopoulos et al., 1998, for example) that the mosquito activates immune-responsive genes during critical transition stages of the parasite life cycle, and so it is thought that such genes play a crucial role in the development and spread of the disease in the vector and ultimately in humans. Several indicators of immune response have been used for monitoring temporally and spatially these defense reactions at the molecular level and have shown clear correlation of immune responses with the passage of Plasmodium through the vector (Kumar et al., 2003).

Hence, the biologists' goal is to understand the underlying system biology and through this the gene pathways and regulatory networks involved. Exploratory statistical tools can help in this regard. In the next section we provide some background on the task of clustering time series. Further details on the biological context and on the microarray technology used to generate the data can be found in Dimopoulos et al. (2002).

1.1 CLUSTER ANALYSIS OF TIME COURSE/LONGITUDINAL DATA

Traditionally, cluster analysis has focused on univariate observations or on multivariate independent observations. Under these circumstances, Euclidean distance or correlation based *hierarchical* clustering is typically used either on the raw or transformed data. The term hierarchical refers to the sequential, conditional partitioning of the data from a single group containing all the observations to a partition where each group contains just one observation. The partitioning can be constructed in a top down fashion, starting from the global cluster, in which case the procedure is termed *divisive*; or more commonly from the bottom up by merging groups together, known as *agglomerative clustering*. The hierarchy, usually represented by a tree known as a *dendrogram*, is informative in that it provides a visual display of group homogeneity within the data at various clustering levels.

Standard clustering techniques were not appropriate for our data due to the time dependency in the observations. Essentially, this converts the clustering problem to one involving the clustering and analysis of *curves* rather than *random variables*, establishing links with functional data analysis as described in, for example, Ramsay and Silverman (1997).

This led us to develop a Bayesian model-based agglomerative scheme for clustering the data.

Our approach uses non-linear regression splines to capture the temporal variation within each cluster. The use of a Bayesian procedure allows us to compute measures of uncertainty for marginal quantities, such as the number of clusters in our data, and to report posterior probabilities that are comparable across all models, experiments, and computational methods. The use of non-linear regression splines allows us to accommodate the non-stationarity in the data as well as the unequal sampling intervals and yet affords analytic calculation of marginal probabilities. For background on Bayesian model-based clustering see the excellent reviews of Banfield and Raftery (1993), Fraley and Raftery (2002) and references therein. Denison et al. (2002) provided an overview of Bayesian regression splines.

Model-based clustering of gene expression time series data has been recently considered by, amongst others, Wakefield et al. (2003), Ramoni et al. (2002), Yeung et al. (2001) and Luan and Li (2003). In addition, extensions of standard Euclidean distance or correlation clustering and modelling of gene expression have been proposed based on projections of the data after performing singular value decomposition on the expression matrix to identify the eigenvectors (or *eigengenes*) as representative expression profiles (see Hastie et al., 2000; Holter et al., 2001, and references therein). These methods are not specificlly tailored to time series and as such are invariant to permutations of the time points; see the discussion below.

Wakefield et al. (2003) perform clustering using a full MCMC Bayesian approach with a basis function representation for the time series with random effects. The marginal likelihood is not analytically available under their model, and inference is made on the basis function coefficients. The size of our data set makes this approach infeasible as the run-time to obtain reasonably accurate approximations to the marginal likelihood for a full hierarchy would be excessive.

Ramoni et al. (2002) propose a pseudo-Bayesian autoregressive model for the time series with improper priors on the coefficients. This method is not appropriate in our case for two reasons; firstly, we do not believe our time series to be stationary processes, and secondly we are interested in the posterior probability distribution on the number of clusters underlying the data, and in particular the most probable number of clusters. The use of an improper prior as in Ramoni et al. (2002) in a fully Bayesian setting would dictate that the most probable number of clusters be one, regardless of the data. This is a consequence of the Lindley-Bartlett paradox, see the discussion of Holmes (2002) for example. However, Ramoni et al. (2002) uses a heuristic search strategy with

a pseudo-Bayesian marginal likelihood criterion and this appears to give useful clusterings for the data analyzed there; we apply Ramoni's approach to our data and report the results in section 6.

Yeung et al. (2001) use the MCLUST software of Fraley and Raftery (1999) (available at `http://www.stat.washington.edu/fraley/mclust`), a generic Bayesian clustering tool, to analyze gene expression profiles. MCLUST fits Gaussian process clusters which are optimized against a BIC criterion. The covariance matrix of the Gaussian process for each cluster is determined via an eigen-decomposition of the empirical covariance matrix obtained from the data in the cluster. In contrast we impose a parametric form, a model, for the covariance function which captures our beliefs about temporal dependencies in the data. In this respect MCLUST is much more general than our method which is specifically tailored for time series. An implication of this is that MCLUST is invariant to permutations of the time points in a series. That is, if we randomly permute the time points, fit a clustering using MCLUST, and then map the clusters obtained back onto the original time domain, we obtain the same clustering regardless of the permutation. Our method is time dependent. The extra structure we impose appears to lead to more cohesive clusters in our example, as highlighted in section 6. In addition, we use a full Bayesian specification, clustering on the exact marginal probabilities rather than an approximating measure such as BIC.

Finally and most recently, Luan and Li (2003) report the use of mixed-effect B-splines to cluster gene expression profiles. This non-hierarchical scheme based on an expectation-minimization algorithm should be contrasted with our fully Bayesian hierarchical approach which explicitly integrates out the spline base parameter values within each cluster to obtain marginal probabilities for the cluster memberships in the tree hierarchy.

To summarise, we propose a method for agglomerative clustering of multiple non-stationary time series using fully probabilistic Bayesian measures of cluster heterogeneity. Moreover, the method readily accommodates non-uniform sampling intervals and missing values in the series. None of the current methods described above capture all of these features which are pertinent to our study.

In order to perform model-based clustering we must first define a probability model for observations within a group. In section 2 we introduce the modelling strategy based on non-linear modelling of gene expression profiles using time indexed basis function representations, and discuss Bayesian inference methods. In section 3 we outline the choice of basis functions used here for the reconstruction of the profiles, whilst noting the modelling and computational constraints imposed

by the structure of the data. In section 4 we describe prior specification. In section 5 we describe our Bayesian hierarchical clustering approach based on maximizing the marginal probability, and demonstrate its implementability. In section 6 we describe the results of our analysis of the malaria-related expression profile data described above, and compare the output with those from competing methods. Finally, in section 7, we offer a discussion and point to possible extensions of the method including full MCMC sampling.

2 MODELLING OF GENE EXPRESSION PROFILES

Model-based clustering requires the specification of a probability distribution for the data residing within a group. We choose to model the gene expression profiles in a regression context via linear models and non-linear basis functions. This approach readily accommodates the non-stationarity and non-uniform sampling aspects of the data (which can be seen most clearly on examination of the clustered data in Figure 2). We highlight that the use of fixed basis functions with random coefficients induces a non-stationary stochastic process model for the underlying variation in expression for which we can analytically evaluate the marginal likelihood.

Generically, we wish to capture the behaviour of the relative gene expression y as a function of time t and measurement error. The basis of our modelling strategy is to use models that are able to capture the characteristic behaviour of expression profiles which we can expect to observe due to different forms of regulation in the immune response.

2.1 THE REGRESSION APPROACH

It is convenient to use a basis function representation for the time series data in Figure 1. In particular, adopting a regression framework, we model the expression level for an individual gene i at time t as

$$y_{it} = X_i(t)\beta + \varepsilon_{it}$$

where $X_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))$ is in general a p-vector of specified basis functions of t, β is a p-vector of basis coefficient parameters and $\{\varepsilon_{it}\}$ is some error process that we shall model as independent and Gaussian.

In vector representation, for gene i we have for expression levels $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ at times

$\underline{t} = (t_1, t_2, \ldots, t_T)$

$$y_i = X_i(\underline{t})\beta + \varepsilon_i \tag{2.1}$$

where for our data $i = 1, \ldots, N = 2771$, $T = 6$. The model is simply a linear regression in a time dependent base $X_i(\underline{t})$. The precise form of design matrix $X_i(\underline{t})$ is at the moment left unspecified. From here on we shall suppress the dependence on $\underline{t}$, writing $X_i(\underline{t})$ as X_i.

Now consider a partition $\mathcal{C}$ of the genes dividing them into C groups of sizes $\{N_1, \ldots, N_C\}$ ($\sum_{i=1}^{C} N_i = N$). Then for the kth set of genes, let the vector $y^{(k)} = \left(y_1^{(k)\prime} \ \ldots \ y_{N_k}^{(k)\prime}\right)'$ be their concatenated expression profiles. The key assumption underlying our clustering method will be that within each set of the partition, the genes follow the regression model (2.1) with a coefficient vector β_k and error variance σ_k^2 specific to that group. Under the assumption that the random error terms $\{\varepsilon_{i_t}\}$ form an i.i.d Gaussian sequence with variance σ_k^2, the conditional distribution of the random variable $Y^{(k)}$ is multivariate normal

$$Y^{(k)}|X^{(k)}, \beta_k, \sigma_k^2 \sim \mathrm{N}\left(X^{(k)}\beta_k, \sigma_k^2 I_{N_k T}\right) \tag{2.2}$$

where now $X^{(k)}$, the design matrix of the group, is of size $N_k T \times p$ and $I_{N_k T}$ is the $N_k T$-dimensional identity matrix.

The form of X_i (or $X^{(k)}$) in (2.1) and (2.2) relates to the specific *basis function representation* used. The class of suitable basis function models for time series is wide and includes Fourier representations, splines, wavelets, and radial basis functions. Basis function representations form some of the most flexible and convenient approaches to nonlinear modelling as, conditional on the basis functions, the model is simply a linear regression in a non-linear design space. Within the Bayesian framework we are in a position to be able to make inference about the most suitable basis representation by comparison of marginal likelihood values for different choice of bases.

The conditional linear structure allows for many of the standard computational and methodological techniques surrounding Bayesian linear models to be employed when making inference. This is an essential feature for our application where the dimensionality of the data is large, and computationally efficient procedures are required in order to make inference in reasonable time.

For general discussions of basis function representations and the so-called *Extended Linear Model*, see, for example, Vidakovic (1999), Schimek (2000), Denison et al. (2002) and Hansen and Kooperberg (2002).

2.2 BAYESIAN REGRESSION

In a Bayesian analysis of the model in (2.2) a joint prior distribution is specified for $\left(\beta_k, \sigma_k^2\right)$. It is convenient to adopt a conjugate prior specification where

$$p\left(\beta_k|\sigma_k^2\right) \equiv \mathrm{N}\left(m, \sigma_k^2 V\right) \qquad p\left(\sigma_k^2\right) \equiv \mathrm{IGamma}\left(\frac{\alpha}{2}, \frac{\gamma}{2}\right) \tag{2.3}$$

m is $p \times 1$, V is $p \times p$ positive definite and symmetric and all other parameters are scalars. Using this prior independently for each gene group index k in the partition, standard Bayesian calculations (see, for example, Denison et al. (2002)) show that conditional on the observed data

$$p\left(\beta_k|y^{(k)}, \sigma_k^2\right) \equiv \mathrm{N}\left(m_k^*, \sigma_k^2 V_k^*\right) \qquad p\left(\sigma_k^2|y^{(k)}\right) \equiv \mathrm{IGamma}\left(\frac{N_k T + \alpha}{2}, \frac{d_k + \gamma}{2}\right) \tag{2.4}$$

where

$$V_k^* = \left(X^{(k)\prime} X^{(k)} + V^{-1}\right)^{-1} \qquad m_k^* = \left(X^{(k)\prime} X^{(k)} + V^{-1}\right)^{-1} \left(X^{(k)\prime} y^{(k)} + V^{-1} m\right)$$
$$d_k = y^{(k)\prime} y^{(k)} + m' V^{-1} m - \left(X^{(k)\prime} y^{(k)} + V^{-1} m\right)' \left(X^{(k)\prime} X^{(k)} + V^{-1}\right)^{-1} \left(X^{(k)\prime} y^{(k)} + V^{-1} m\right)$$

In regression modelling, it is usual to consider a centred parameterization for β_k so that $m = 0$, giving

$$m_k^* = \left(X^{(k)\prime} X^{(k)} + V^{-1}\right)^{-1} X^{(k)\prime} y^{(k)}$$
$$d_k = y^{(k)\prime} \left(I_{N_k T} - X^{(k)} \left(X^{(k)\prime} X^{(k)} + V^{-1}\right)^{-1} X^{(k)\prime}\right) y^{(k)} \tag{2.5}$$

2.3 MARGINAL LIKELIHOOD

The critical quantity in our clustering procedure will be the marginal likelihood or prior predictive distribution for each cluster k,

$$p\left(y^{(k)}\right) = \int\int p\left(y^{(k)}|\beta_k, \sigma_k^2\right) p\left(\beta_k|\sigma_k^2\right) p\left(\sigma_k^2\right) d\beta_k d\sigma_k^2. \tag{2.6}$$

The marginal likelihood is an attractive measure of cluster integrity as it explicitly quantifies the probability that a group of measurements arose from the same underlying stochastic process.

Combining (2.2) and (2.3) gives

$$p\left(y^{(k)}|\sigma_k\right) \equiv \mathrm{N}\left(0, \sigma_k^2 \left[X^{(k)} V X^{(k)\prime} + I_{N_k T}\right]\right) \tag{2.7}$$

which, after marginalizing over σ_k, leads to

$$\begin{aligned} p\left(y^{(k)}\right) &= \left(\frac{1}{\pi}\right)^{N_k T/2} \frac{\gamma^{\alpha/2}\Gamma\left(\frac{N_k T+\alpha}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \frac{|V_k^*|^{1/2}}{|V|^{1/2}} \frac{1}{\{d_k+\gamma\}^{(N_k T+\alpha)/2}} \\ &= g\left(N_k T, \alpha, \gamma\right)|V|^{-1/2} \frac{1}{\left|X^{(k)\prime}X^{(k)}+V^{-1}\right|^{1/2}\{d_k+\gamma\}^{(N_k T+\alpha)/2}} \end{aligned} \tag{2.8}$$

where $g\left(N_k T, \alpha, \gamma\right)|V|^{-1/2}$ is a normalizing constant independent of the data. We note here that the vague prior specification $V^{-1} \to 0$ in (2.8) leads to $p(y^{(k)}) \to 0$ and impropriety or indeterminacy. This fact prevents a fully non-informative prior specification being used and, as mentioned in the introduction, can lead to the Lindley-Bartlett paradox when considering models of non-fixed dimension; see the discussion of Holmes (2002). For our data analysis, V was chosen to approximately maximize the marginal likelihood as discussed in section 4.1.

We shall use this marginal likelihood as the *potential function* of a Bayesian hierarchical clustering procedure that is readily computable for large data samples; within each cluster, the gene expression profiles will be assumed to have originated from a common Gaussian process, giving rise to a marginal likelihood for each cluster of the form (2.8).

3 CHOICE OF DESIGN MATRIX AND BASIS FUNCTION

We now consider the design matrix, X_i, for a single gene expression curve y_i that appears in (2.1). This $T \times p$ matrix consists of rows of possibly non-linear functions of the time ordinates at which the expression measurements are taken. That is, for $s = 1, \ldots, T$, the row s denotes the response of p basis functions $g_1, \ldots, g_p$ at the time point t_s,

$$\left[g_1\left(t_s\right), \ldots, g_p\left(t_s\right)\right].$$

For our data analysis, we use a flexible family of basis functions called the *truncated power spline basis*, taking the form

$$g_1(t_s) = 1 \qquad g_j(t_s) = (t_s - t_{j-1})_+^q, j = 2, \ldots, T,$$

where $(\cdot)_+ = \max\{0, \cdot\}$ and q is a positive integer.

Of particular note are the special cases of $q = 1$, which gives a continuous piecewise linear model and $q = 3$, which is the piecewise *cubic regression spline* model. Each alternate basis function choice

induces a different, non-stationary marginal covariance structure on the Gaussian process via (2.7). The truncated power base is a very flexible form for modelling curves, as seen in Denison et al. (2002), Chapter 3.

Recall that for our data the expression measurements are taken for all genes at the same time points. Thus the design matrix for a single gene X_i will not differ between genes and so from now on will be referred to generically as X.

4 PRIOR MODELLING

4.1 CHOICE OF PRIOR COVARIANCE

Ideally, for a fully Bayesian approach we would like to treat the prior covariance matrix V as fully unknown and following some multivariate prior distribution. However, this would carry with it great extra computational burden, so instead we simply assume independence of the $\{\beta_j\}$ by letting V be a scalar multiple v of the p-dimensional identity matrix, so $V = \text{diag}(v)$. The value of the multiplier v is then chosen in an empirical Bayes fashion to optimize the marginal likelihood of the resulting clustering.

To test sensitivity in the prior, we also considered variations of this scheme which introduced some prior correlation between the $\{\beta_j\}$, such as calculating an estimated sample covariance matrix of the $\{\beta_j\}$. When such alternatives were used in the model to analyze our data they gave very similar results and so these are not included in this paper.

4.2 NUMBER OF CLUSTERS AND CLUSTER SIZES

A Bayesian specification of a clustering regression model also requires a prior model for the clustering $\mathcal{C}$. Assuming exchangeability in the assignment of genes to clusters, it is sufficient to specify prior distributions for the number of clusters C and the cluster sizes $N_1, \ldots, N_C$ to satisfy this requirement. We use a default specification placing a uniform prior on C over the range $\{1, 2, \ldots, N\}$ and for the sizes of those C clusters, a Multinomial-Dirichlet distribution. This gives

$$p(\mathcal{C}|\xi_1, \ldots, \xi_C) = \frac{1}{N} \frac{\Gamma\left(\sum_{i=1}^{C} \xi_i\right)}{\prod_{i=1}^{C} \Gamma(\xi_i)} \frac{\prod_{i=1}^{C} \Gamma(N_i + \xi_i)}{\Gamma\left(N + \sum_{i=1}^{C} \xi_i\right)}$$

with uniform settings on the Multinomial-Dirichlet parameters $\xi_1 = \ldots = \xi_C = 1$ leading to

$$p(\mathcal{C}) = \frac{(C-1)!N_1!N_2!\ldots N_C!}{N(N+C-1)!}. \tag{4.1}$$

Whilst a Multinomial-Dirichlet prior theoretically allows the possibility of empty clusters, this can be disregarded when we come to present our agglomerative clustering algorithm for maximizing posterior probability, as it is easily seen that the prior probability (4.1) of a clustering containing empty components is strictly less than the probability of an identical partition with no empty sets.

5 BAYESIAN HIERARCHICAL CLUSTERING

Hierarchical clustering is a method of organizing a collection of objects into disjoint sets using a similarity/discrepancy measure or by some overall potential function, such that objects in sets are more similar to each other than objects across sets. Agglomerative clustering initially places each of the N items in its own cluster. At the first level, two objects are to be clustered together, and the pair is selected such that the potential function increases by the largest (best, or least worst) amount, leaving $N-1$ clusters, one with two members, the remaining $N-2$ each with one. At the next level, the optimal configuration of $N-2$ clusters is found, by joining two of the existing clusters. This process continues until a single cluster remains containing all N items. Most commonly the similarity measure is based on Euclidean distance between data sequences i and j, $\|y_i - y_j\|^2 = (y_i - y_j)'(y_i - y_j)$. The method we propose forms clusters on the basis of the covariance structure induced by the underlying stochastic process (or, at least, our Gaussian process representation of it). So in effect, our hierarchical clustering approach assigns profiles to the same cluster if they are similar in covariance terms. The potential advantages of 'covariance clustering' are evident: it respects the time ordering of the data, and by judicious selection of the design and prior matrices can lead to biologically appropriate covariance structures being discovered, as it can be used to incorporate knowledge of the dynamics of the underlying processes involved in the regulation of expression.

5.1 COMPUTATIONALLY EFFICIENT BAYESIAN CLUSTERING

A principal feature of our data is the dimensionality. This is typical of gene microarray studies, where the technology enables thousands of gene expression measurements to be taken simultane-

ously, and in our case, repeatedly at a series of time points. Therefore when constructing statistical methods to analyze these data, it is crucial to examine the implications of this dimensionality on the feasibility of implementation.

On the face of it, implementing a fully Bayesian cluster analysis of non-stationary time series data, without resorting to time-consuming MCMC, appears challenging. However, as shown in section 2, by modelling the curves using non-linear regression splines we are able to adopt conjugate priors on the coefficients and thus obtain an analytic expression for the marginal likelihood. We will see how the choice of covariance matrix induced by our basis function representation leads to some further important simplifications in the calculation of the marginal likelihoods of each cluster.

In our modelling framework, expression profiles in the same cluster k have the same (unknown) regression parameters β_k and variance parameter σ_k^2, and thus, conditional on $\left(\beta_k, \sigma_k^2\right)$ the data sequences for clustered data are mutually independent, and the likelihood, posterior and marginal likelihood can be evaluated as shown in section 2. For example, conditional on $\left(\beta_k, \sigma_k^2\right)$ the likelihood of the N_k profiles in the kth cluster $y^{(k)}$ is given by (2.2) where now, since in our data the time points at which the expression profiles are observed are identical, ${X^{(k)}}' = \begin{bmatrix} X' & X' & \cdots & X' \end{bmatrix}$. This implies

$$ {X^{(k)}}' X^{(k)} = \left(X'X + X'X + \ldots + X'X\right) = N_k X'X $$

and

$$ {X^{(k)}}' y^{(k)} = \sum_{i=1}^{N_k} X' y_i^{(k)}. $$

Hence the quantities in (2.5) and thus (2.8) can be presented in simple form. In particular, (2.8) becomes

$$ p\left(y^{(k)}\right) = g\left(N_k T, \alpha, \gamma\right) |V|^{-1/2} \frac{1}{|N_k X'X + V^{-1}|^{1/2} \left\{d_k + \gamma\right\}^{(N_k T + \alpha)/2}}. \tag{5.1} $$

where now

$$ d_k = \left(\sum_{i=1}^{N_k} {y_i^{(k)}}' y_i^{(k)}\right) - \left(\sum_{i=1}^{N_k} X' y_i^{(k)}\right)' \left(N_k X'X + V^{-1}\right)^{-1} \left(\sum_{i=1}^{N_k} X' y_i^{(k)}\right) $$

Then the marginal posterior probability of a particular clustering of the genes given the data is given up to proportionality by

$$ \pi(C, N_1, N_2, \ldots N_C | y) = p(C) p(N_1, N_2, \ldots N_C) \prod_{k=1}^{C} p\left(y^{(k)}\right) \tag{5.2} $$

where the third term on the right hand side is the product of the likelihoods of the expression profiles for each cluster k given by (5.1).

When two clusters k and l are merged to form $y^{(kl)} = (y^{(k)} \;\; y^{(l)})$, (5.1) becomes

$$p\left(y^{(kl)}\right) = \frac{g\left(\left(N_k + N_l\right)T, \alpha, \gamma\right)|V|^{-1/2}}{\left|\left(N_k + N_l\right)X'X + V^{-1}\right|^{1/2}\left\{d_{kl} + \gamma\right\}^{((N_k+N_l)T+\alpha)/2}} \tag{5.3}$$

where

$$d_{kl} = \left(\sum_{i=1}^{N_k+N_l} y_i^{(kl)\prime} y_i^{(kl)}\right) - \left(\sum_{i=1}^{N_k+N_l} X'y_i^{(kl)}\right)' \left((N_k + N_l)X'X + V^{-1}\right)^{-1} \left(\sum_{i=1}^{N_k+N_l} X'y_i^{(kl)}\right) \tag{5.4}$$

In terms of computation, clearly the key quantities in (5.3) and (5.4) are

$$W_{N_k+N_l} = \left(N_k + N_l\right)X'X + V^{-1} \qquad \text{and} \qquad |W_{N_k+N_l}|.$$

This leads to considerable simplification in the hierarchical clustering as we can compute the quantities

$$W_n^{-1} = \left(nX'X + V^{-1}\right)^{-1} \qquad \text{and} \qquad |W_n|$$

off-line for all $n \in \{1, 2, \ldots, N\}$. In addition, for each gene i we can at the start compute

$$\left\{y_i'y_i,\, X'y_i\right\}$$

and then simply take sums over each gene in a cluster to get the required quantities in (5.4).

The procedures above lead to considerable savings in computation time. First, the use of non-linear regression splines with conjugate priors allows for explicit calculation of the marginal likelihood for any clustering. Second, the basis function covariance matrix for a cluster of size n is simply $nX'X$. Third, the necessary matrix multiplications and inversions can be calculated off-line and stored in a look-up table. This makes the analysis of our data feasible, with a final run time generating the full cluster hierarchy of just over half a minute reported in section 6.

It should be noted that although assuming all profiles are observed at the same set of time points gives some of the computational savings we have just indicated, this assumption can be relaxed without losing any of the distributional results and the clustering algorithm proceeds in the same way. This is not the case for conventional hierarchical clustering which breaks down if there are different time points or different numbers of time points. This is then a further advantage of the method we propose as a general method for functional clustering.

5.2 BAYESIAN HIERARCHICAL CLUSTERING ALGORITHM

The algorithm proceeds as follows:

Step 1: Start with $C = N$ clusters, each cluster containing the expression levels for one gene. Calculate the marginal posterior unnormalized probability kernel π_N in (5.2).

Step 2: For each pair of clusters k, l, letting $N_{kl\cdot}$ denote the vector of cluster sizes other than $\{N_k, N_l\}$, we calculate the multiplicative increase in marginal posterior that would be gained by merging the two clusters to obtain an inter-cluster closeness

$$\begin{aligned} c_{kl} = c_{lk} &= \frac{p(N_k + N_l | N_{kl\cdot}) p(y^{(kl)})}{p(N_k, N_l | N_{kl\cdot}) p(y^{(k)}) p(y^{(l)})} \\ &= \frac{(N + C - 1)(N_k + N_l)! p(y^{(kl)})}{(C-1) N_k! N_l! p(y^{(k)}) p(y^{(l)})} \end{aligned} \tag{5.5}$$

which follows from the prior (4.1) and where $p(y^{(\cdot)})$ is given by (5.1) or (5.3) ($N(N-1)/2$ calculations).

Step 3: For each cluster k, identify the closest other cluster according to the metric (5.5) and the corresponding maximum closeness value

$$k' = \arg\max_l c_{kl}, \quad c_k = c_{kk'}.$$

Step 4: Find the cluster $\hat{k}$ with largest c_k value, and merge with cluster $\hat{k}'$ to form a new cluster $\hat{k}$. Set $C = C - 1$ and relabel the other remaining clusters accordingly. Calculate the revised marginal unnormalized posterior kernel value

$$\pi_C = c_{\hat{k}\hat{k}'} \pi_{C+1}.$$

Step 5: For each cluster $l \neq \hat{k}$, calculate the closeness to cluster $\hat{k}$, $c_{\hat{k}l}$ (C calculations), and identify the new nearest cluster $\hat{k}'$.

Step 6: For each cluster $l \neq \hat{k}$, update the stored nearest cluster l'; unless the stored cluster l' was just merged, we only need to check the value of c_l against $c_{\hat{k}l}$.

Step 7: Repeat *Steps 4-6* until $C = 1$.

Step 8: Looking back over the clusterings visited, find the number of clusters C in the hierarchy maximizing the posterior distribution, $\arg\max_C \pi_C$. This is our optimal clustering.

This algorithm is repeated for a collection of candidate settings of the prior covariance matrix $V = vI_p$, $v \in \{v_1, \ldots, v_J\}$ and the maximum reported. For the results presented in section 6 below, we used 10 candidate points equally spaced on the log scale with $v_1 = 10^{-3}$, $v_{10} = 10^5$.

6 RESULTS

We applied the clustering methodology introduced over the preceding sections to the Anopheles gene expression data in Figure 1. The piecewise linear model was used ($q = 1$ in the notation of section 3) and the spline coefficients β were assumed independent in our choice of prior covariance matrix V, as this combination gave the maximum marginal probability for this data.

As we had little prior information about the variance parameter σ^2, we chose the hyperparameters (α, γ) to both be small ($\alpha = \gamma = 10^{-2}$). This ensures an uninformative prior specification, as can be seen by examination of the marginal likelihood (2.8).

For a comparison we also obtained the output from three other methods applied to the data: Euclidean distance average-link hierarchical clustering, the Bayesian model-based clustering software MCLUST of Fraley and Raftery (1999), and the autoregressive model-based approach of Ramoni et al. (2002) using the accompanying software CAGED (`http://www.genomethods.org/caged`).

Of the 2900 genes in the original data set of Dimopoulos et al. (2002), 129 of the expression profiles had missing data. These have been discarded to enable a fair comparison with the other methods. However, as noted earlier, the presence of missing data presents no theoretical problems to our proposed methodology except for some extra computational burden. The complete run on the reduced data set took 38 seconds on a 2GHz processor PC, using C++ code. The code is freely available for download from (`http://stats.ma.imperial.ac.uk/~naheard/software/splinecluster`).

Figure 2 is an image plot of the clustered data under the most probable model visited (which contains 19 clusters). The left hand side of the figure showing the reordered data should be compared to Figure 1. To the right are the (ranked) mean profiles for each cluster and finally a dendrogram indicating the order in which the 19 clusters would merge to a single group under further agglomerative clustering. Comparing the cluster means on the right hand side gives a visual

measure of between cluster heterogeneity; while comparing the group means to the raw observations gives a sense of within cluster homogeneity.

The most probable clustering is shown in further detail in Figure 3. The optimal number of clusters maximizing the log unnormalized marginal probability (-3,250.9) using our Bayesian method was 19. Recall that *a priori* the number of clusters was assumed uniformly distributed across the range. The top left hand plot of Figure 3 shows the number of clusters versus the log unnormalized marginal probability. This is generally well behaved, with a clear global maximum. Particularly to the right of this maximum the decrease of the curve is fairly shallow, suggesting many other plausible alternative clusterings of different sizes besides the optimum. As ratios of marginal likelihoods are Bayes factors, qualitative interpretation of the relative plausibility of the different clusterings is readily available (see Kass and Raftery, 1995).

The remaining plots in Figure 3 are scatter plots of the raw data for each of the individual clusters under the optimal clustering, along with the cluster mean profiles and bars indicating plus or minus two posterior standard deviations of the profile from the mean (solid lines).

From examination of Figure 3, several clusters are noteworthy:

Cluster 12 - Slight down regulation for the first hour before steadily increasing throughout the remaining period. This group highlights genes which are progressively more transcribed further into the course of infection.

Clusters 13 - Significant initial up-regulation up to the second time point (four hours after infection) followed by gradual decrease through the time course. This cluster contains a high proportion of the genes known to be related to the immune-defense system; the dataset has 356 'labelled' genes of known function, of which 23 are related to the immune defense system; 9 of the immune defense genes appear in the 27 labelled genes in this cluster.

Cluster 16 - No significant up or down regulation until halfway through the time course, where expression is suddenly heavily down regulated before beginning to pick up again at 18 hours.

Cluster 17 - Contains just three genes, all highly down regulated for the whole of the time course. These patterns could be of significance, or may simply be outliers. Our analysis cannot distinguish between these two cases without further prior insight, but these genes can be highlighted for further investigation.

Additionally, it is perhaps worth noting that our method successfully clustered together some large groups of genes with almost no change in relative expression over time (particularly cluster 4), suggesting application of the bacterial challenge does not change the regulation of these genes.

This hierarchical clustering provides a qualitative tool to highlight potentially interesting structure within the data. Prior to our involvement the biologists were using Euclidean based measures which they found provided little insight. Using our approach, the biologists are able to target those genes of unknown functionality for further investigation which are grouped alongside labelled genes known to be involved in immune response.

Figure 4 shows the scatter plots for the clusters obtained from Euclidean hierarchical clustering, where the number of clusters to be formed was taken to be the optimal number from our Bayesian method. It is clear that this method fails to capture the same dynamical structure revealed by our approach. For instance, there do not appear to be any clusters representing no overall change in expression. Moreover, most of the genes are contained in the three large clusters 10, 11 and 12, and the latter of these, which is by far the largest containing 85% of the genes, does not exhibit any real temporal cohesion.

To further compare our approach we analyzed the data using the software packages MCLUST and CAGED described in Yeung et al. (2001) and Ramoni et al. (2002) respectively.

Figure 5 gives the corresponding plot obtained when using the MCLUST software, which selected the most general model (VVV in their notation) with 11 clusters using a BIC criterion. The expression profiles are visually less cohesive than those shown in Figure 3. This is particularly apparent in clusters 5, 9 and 10. As with our method, MCLUST was successful in picking out a group of genes showing almost no change in regulation (cluster 2). It is noteworthy that compared to our optimal clustering, the gene partitions suggested by Euclidean clustering and MCLUST have significantly lower log unnormalized marginal probability under our probability model, -6,423.5 and -9,309.2 respectively.

When the data was analyzed using the autoregressive method of Ramoni et al. (2002) using the lagged correlation option of the provided software CAGED, either one or two clusters were fitted for the different choices of Markovian order. This unintuitive result may be due to the non-stationarity in some of the expression profiles in our data, or the non-uniform spacing of the sampling times.

To put these results in context, Table 1 shows the distributions of all the genes of known function

through the 19 clusters from the Bayesian method. These functions have been grouped into classes such as involvement in immunity (I), oxidation/reduction reactions or expression in mitochondria (R), or encoding ribosomal proteins or other components of protein metabolism (P); for more details, see Dimopoulos et al. (2002). Whilst the immune defence genes appear to have clustered together fairly well, the other functions seem fairly uniformly distributed. This is unsurprising, as it is the immune defence genes which we would expect to be most stimulated by the bacterial challenge. Similar tables were prepared for Euclidean clustering and MCLUST, and neither had clusters as pure in immune defence genes as Cluster 13 from our method; overall, the Bayesian clustering method was the most successful at producing groups with a high density of immunity genes, though both this method and MCLUST were much more successful than Euclidean hierarchical clustering.

7 CONCLUSIONS AND EXTENSIONS

We have demonstrated the utility of Bayesian hierarchical clustering procedures in the construction of biologically interpretable gene clusters, and have shown the agglomerative hierarchical clustering method based on covariance clustering can outperform the standard clustering methods currently used by biologists. We developed the use of non-linear splines to accommodate the key features in our data; the high dimensionality, the non-stationarity of the profiles and the unequally spaced sampling time points. This lead us to discover better, more visually cohesive expression profile clusters than the other standard methods. Computationally, as shown in section 5.1 the method is readily implementable with the full agglomerative clustering in our real data example, starting with 2771 genes each in their own cluster and successively combining up to a single global cluster containing all the genes, taking just over half a minute.

This speed of computation also suggests the possible use of the optimal clustering model as a good starting point for a full Bayesian Markov chain Monte Carlo (MCMC) model-based clustering analysis. Our experiences with full MCMC cluster analysis, however, are somewhat mixed; for small sample sizes, standard MCMC methods work well, similar to the variable dimensional MCMC analysis of mixture models. But for mixture models in dimensions greater than one, for large sample sizes, it is rather difficult to construct efficient dimension changing moves in the vast space of possible clusterings. This remains a significant problem in all such mixture modelling problems.

REFERENCES

Alphey, L., Beard, C. B., Billingsley, P., Coetzee, M., Crisanti, A., Curtis, C., Eggleston, P., Godfray, C., Hemingway, J., Jacobs-Lorena, M., James, A. A., Kafatos, F. C., Mukwaya, L. G., Paton, M., Powell, J. R., Schneider, W., Scott, T. W., Sina, B., Sinden, R., Sinkins, S., Spielman, A., Toure, Y. and Collins, F. H. (2002) Malaria Control with Genetically Manipulated Insect Vectors. *Science*, **298**, 119–121. URL `http://www.sciencemag.org/cgi/content/abstract/298/5591/119`.

Banfield, J. D. and Raftery, A. E. (1993) Model-based gaussian and non-gaussian clustering. *Biometrics*, **49**, 803–821.

Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T. W., Pertea, M., Silva, J. C., Ermolaeva, M. D., Allen, J. E., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., f. Shumway, M., l. Bidwell, S., Shallom, D. J., e. Van Aken, S., b. Riedmuller, S., Feldblyum, T. V., Che, J. K., Quackenbush, J., Sedegah, M., Shoaibi, A., leda m. Cummings, Florens, L., Yates, J. R., Raine, J. D., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., h. Van Lin, L., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., j. Craig venter, Fraser, C. M., l. Hoffman, S., Gardner, M. L. and Carucci, D. J. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature*, **419**, 512–519.

Christophides, G. K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P. T., Collins, F. H., Danielli, A., Dimopoulos, G., Hetru, C., Hoa, N. T., Hoffmann, J. A., Kanzok, S. M., Letunic, I., Levashina, E. A., Loukeris, T. G., Lycett, G., Meister, S., Michel, K., Moita, L. F., Muller, H.-M., Osta, M. A., Paskewitz, S. M., Reichhart, J.-M., Rzhetsky, A., Troxler, L., Vernick, K. D., Vlachou, D., Volz, J., von Mering, C., Xu, J., Zheng, L., Bork, P. and Kafatos, F. C. (2002) Immunity-Related Genes and Gene Families in Anopheles gambiae. *Science*, **298**, 159–165. URL `http://www.sciencemag.org/cgi/content/abstract/298/5591/159`.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002) *Bayesian Methods for Nonlinear Classification and Regression*. Chichester: Wiley.

Dimopoulos, G., Casavant, T. L., Chang, S., Scheetz, T., Roberts, C., Donohue, M., Schultz, J.,

Benes, V., Bork, P., Ansorge, W., Bento Soares, M. and Kafatos, F. C. (2000) Anopheles gambiae pilot gene discovery project: Identification of mosquito innate immunity genes from expressed sequence tags generated from immune-competent cell lines. *PNAS*, **97**, 6619–6624.

Dimopoulos, G., Christophides, G. K., Meister, S., Schultz, J., White, K. P. and Barillas-Mury, C andKafatos, F. C. (2002) Genome expression analysis of Anopheles gambiae: Responses to injury, bacterial challenge and malaria infection. *PNAS*, **99**, 8814–8819.

Dimopoulos, G., Seeley, D., Wolf, A. and Kafatos, F. C. (1998) Malaria infection of the mosquito Anopheles gambiae activates immune-responsive genes during critical transition stages of the parasite life cycle. *The EMBO Journal*, **17**, 6115–6123.

Florens, L., Washburn, M. P., Raine, J., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinded, R. E., Yates, J. R. and Carucci, D. J. (2002) A proteomic view of the Plasmodium falciparum life cycle. *Nature*, **419**, 520–526.

Fraley, C. and Raftery, A. E. (1998) How many clusters? which clustering method? - answers via model-based cluster analysis. *Computer Journal*, **41**, 578–588.

— (1999) Mclust: software for model-based cluster analysis. *J. Classification*, **16**, 297–306.

— (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, **97**, 611–631.

Hansen, M. and Kooperberg, C. (2002) Spline adaptation in extended linear models. *Statistical Science*, **17**, 2–51.

Hastie, T., Tibshirani, R., B, E. M., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D. and Brown, P. (2000) 'gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, **1**, 1–21.

Holmes, C. C. (2002) Discussion of *Spline Adaptation in Extended Linear Models*, by M Hansen and C Kooperberg. *Statistical Science*, **17**, 2–51.

Holter, N. S., Maritan, A., Marek, C., Fedoroff, N. V. and Banavar, J. R. (2001) Dynamic modeling of gene expression data. *Proc. Nat. Acad. Sci.*, **98**, 1693–1698.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *J. Amer. Statist. Assoc.*, **90**, 773–795.

Kumar, S., Christophides, G, K., Cantera, R., Charles, Band Han, Y. S., Meister, S., Dimopoulos, G., Kafatos, F. C. and Barillas-Mury, C. (2003) The role of reactive oxygen species on Plasmodium melanotic encapsulation in Anopheles gambiae. *PNAS*, **100**, 14139–14144.

Luan, Y. and Li, H. (2003) Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, **19**, 474–482.

Ramoni, M., Sebastiani, P. and Kohane, P. R. (2002) Cluster analysis of gene expression dynamics. *Proc. Nat. Acad. Sci.*, **99**, 9121–9126.

Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis*. Springer-Verlag, New York.

Schimek, M. G. (ed.) (2000) *Smoothing and Regression: Approaches, Computation and Application*. John Wiley.

Vidakovic, B. (1999) *Statistical Modelling by Wavelets*. John Wiley.

Wakefield, J., Zhou, C. and Self, S. (2003) Modelling gene expression over time: curve clustering with informative prior distributions. In *Bayesian Statistics 7* (eds. J. M. Bernardo, M. J. Bayarri, B. J. O, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West). Oxford: Clarendon Press.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987. URL `http://bioinformatics.oupjournals.org/cgi/content/abstract/17/10/977`.

Zdobnov, E. M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R. R., Christophides, G. K., Thomasova, D., Holt, R. A., Subramanian, G. M., Mueller, H.-M., Dimopoulos, G., Law, J. H., Wells, M. A., Birney, E., Charlab, R., Halpern, A. L., Kokoza, E., Kraft, C. L., Lai, Z., Lewis, S., Louis, C., Barillas-Mury, C., Nusskern, D., Rubin, G. M., Salzberg, S. L., Sutton, G. G., Topalis, P., Wides, R., Wincker, P., Yandell, M., Collins, F. H., Ribeiro, J., Gelbart, W. M., Kafatos, F. C. and Bork, P. (2002) Comparative Genome and Proteome Analysis of Anopheles gambiae and Drosophila melanogaster. *Science*, **298**, 149–159. URL `http://www.sciencemag.org/cgi/content/abstract/298/5591/149`.

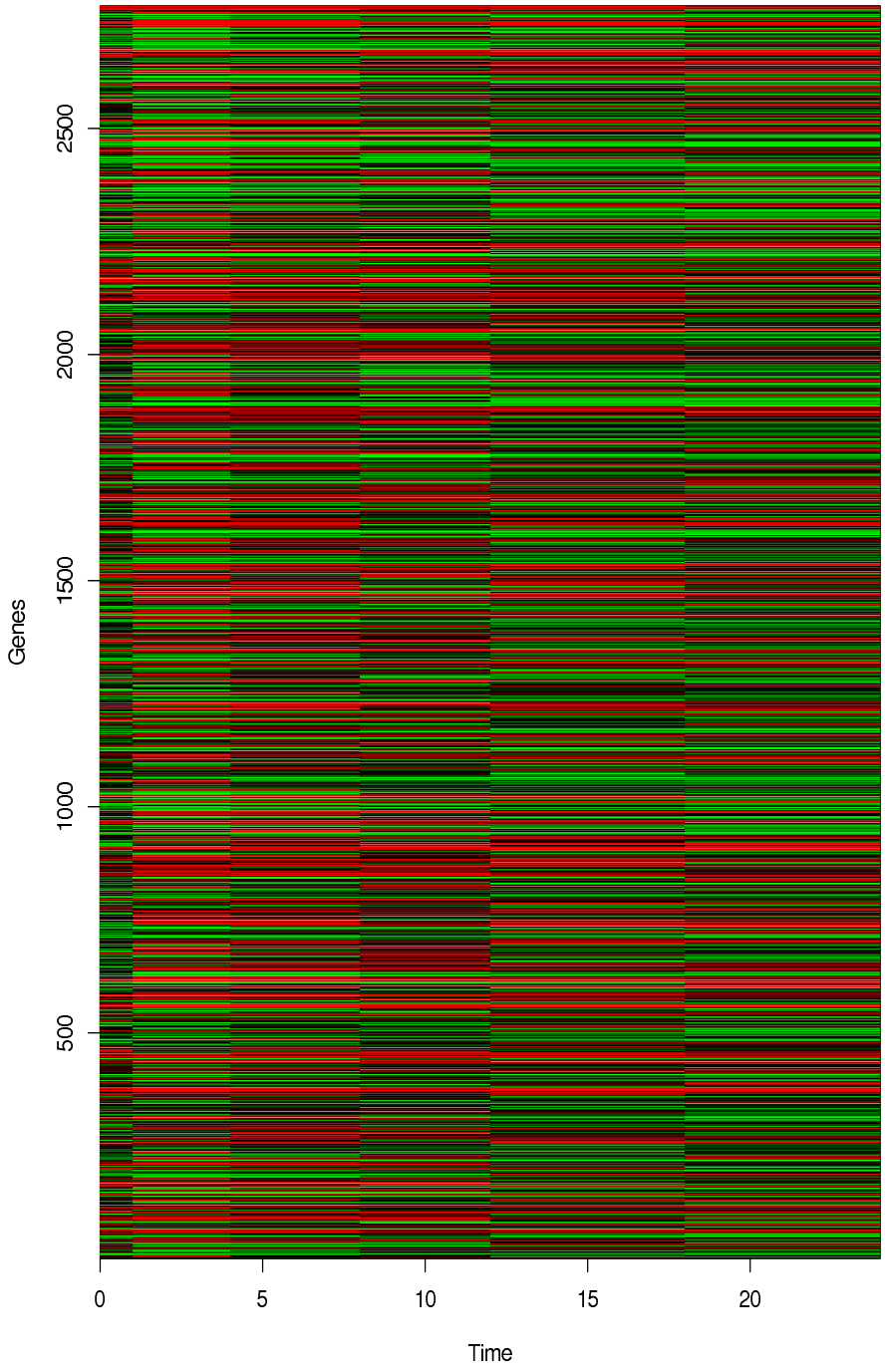

Figure 1: *Piecewise constant plots of unclustered gene expression profiles for the* Salmonella typhi *data set. Brighter red (green) colours correspond to higher (lower) expressions. (Note we use rank relative expression to avoid saturation due to a few outlying gene expression levels).*

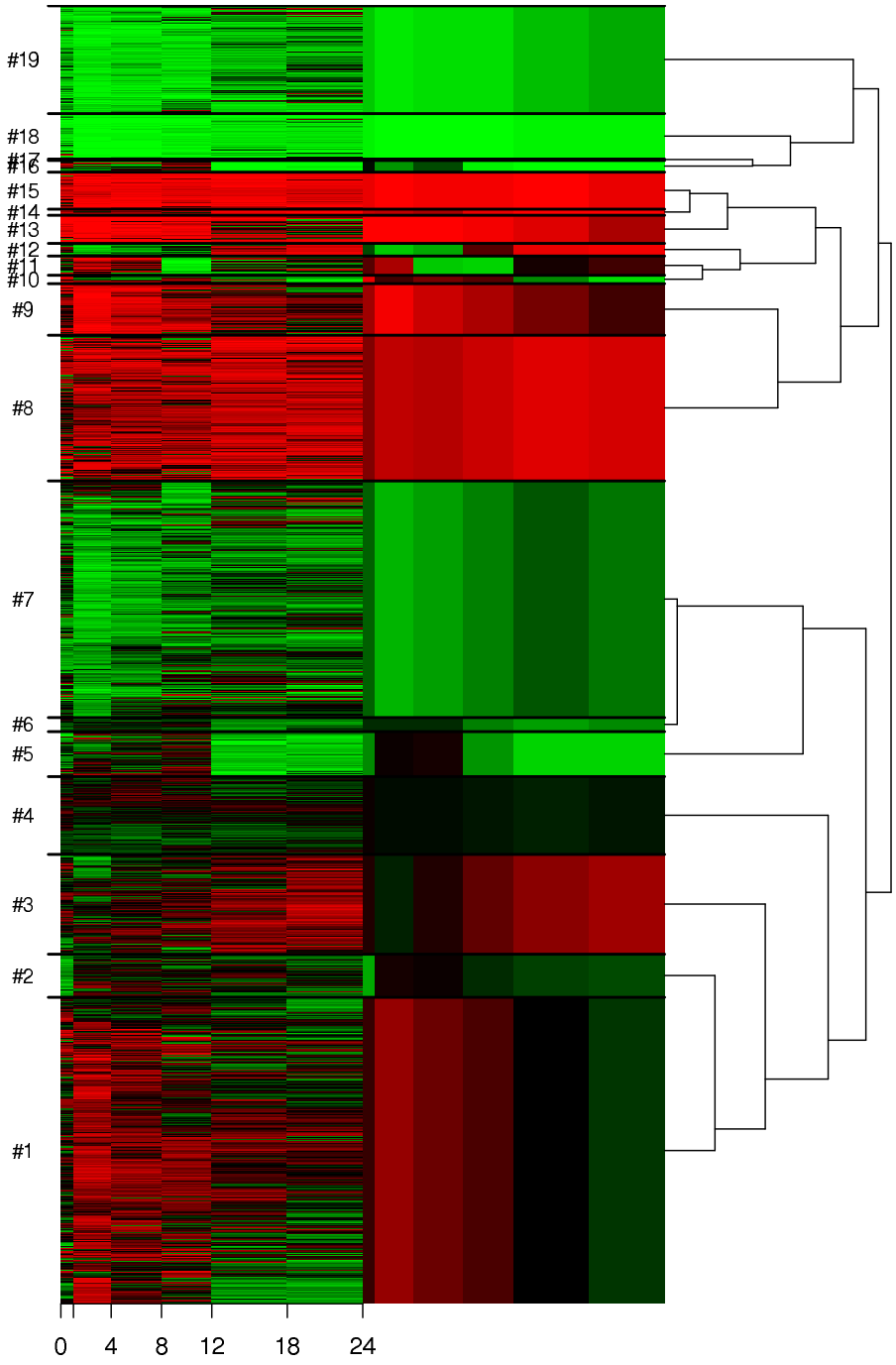

Figure 2: *Clustered gene expression profiles from the* Salmonella typhi *data.*

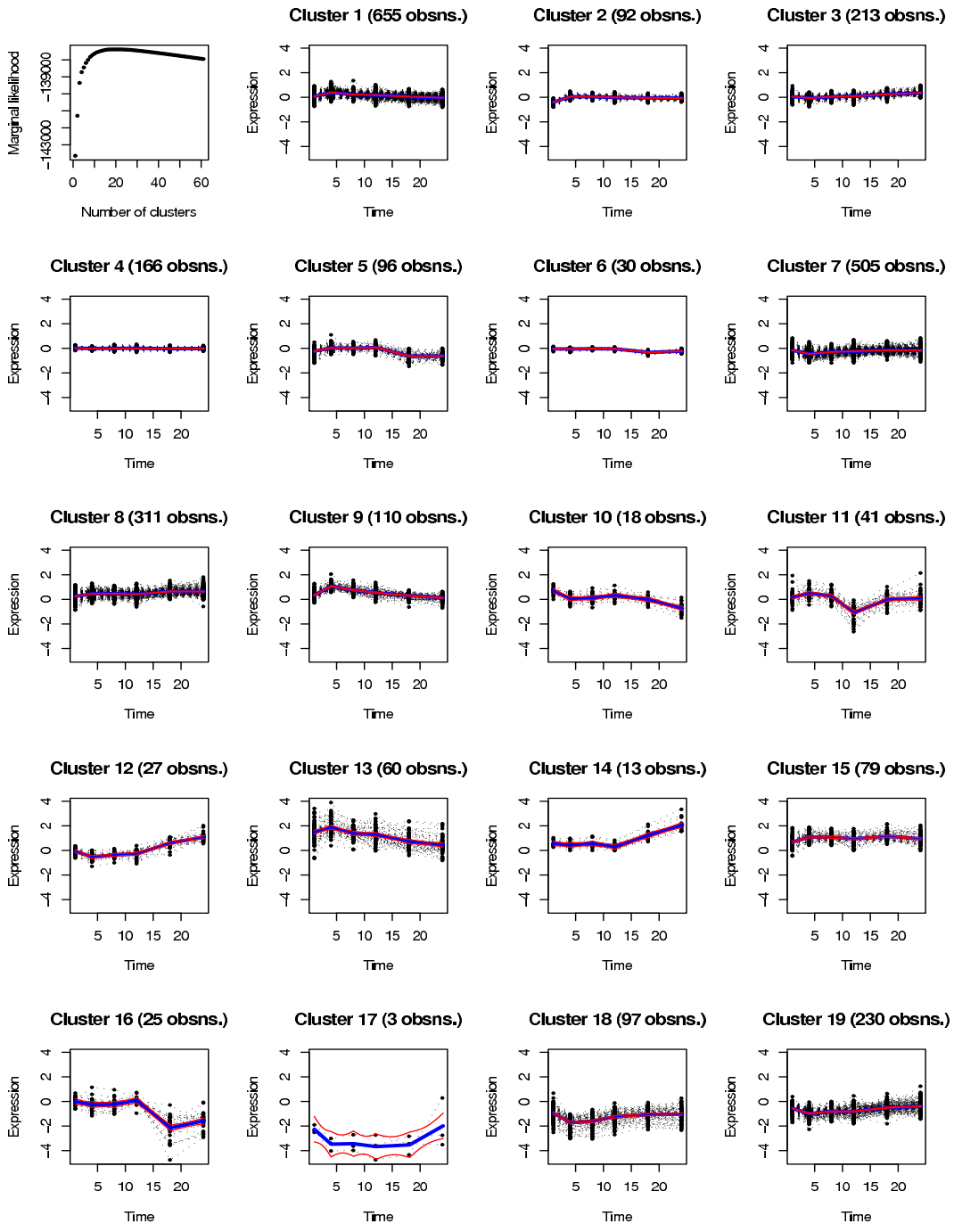

Figure 3: *Plot of marginal likelihood against number of clusters, and the optimal clusters after Bayesian agglomerative clustering of the* Salmonella typhi *data.*

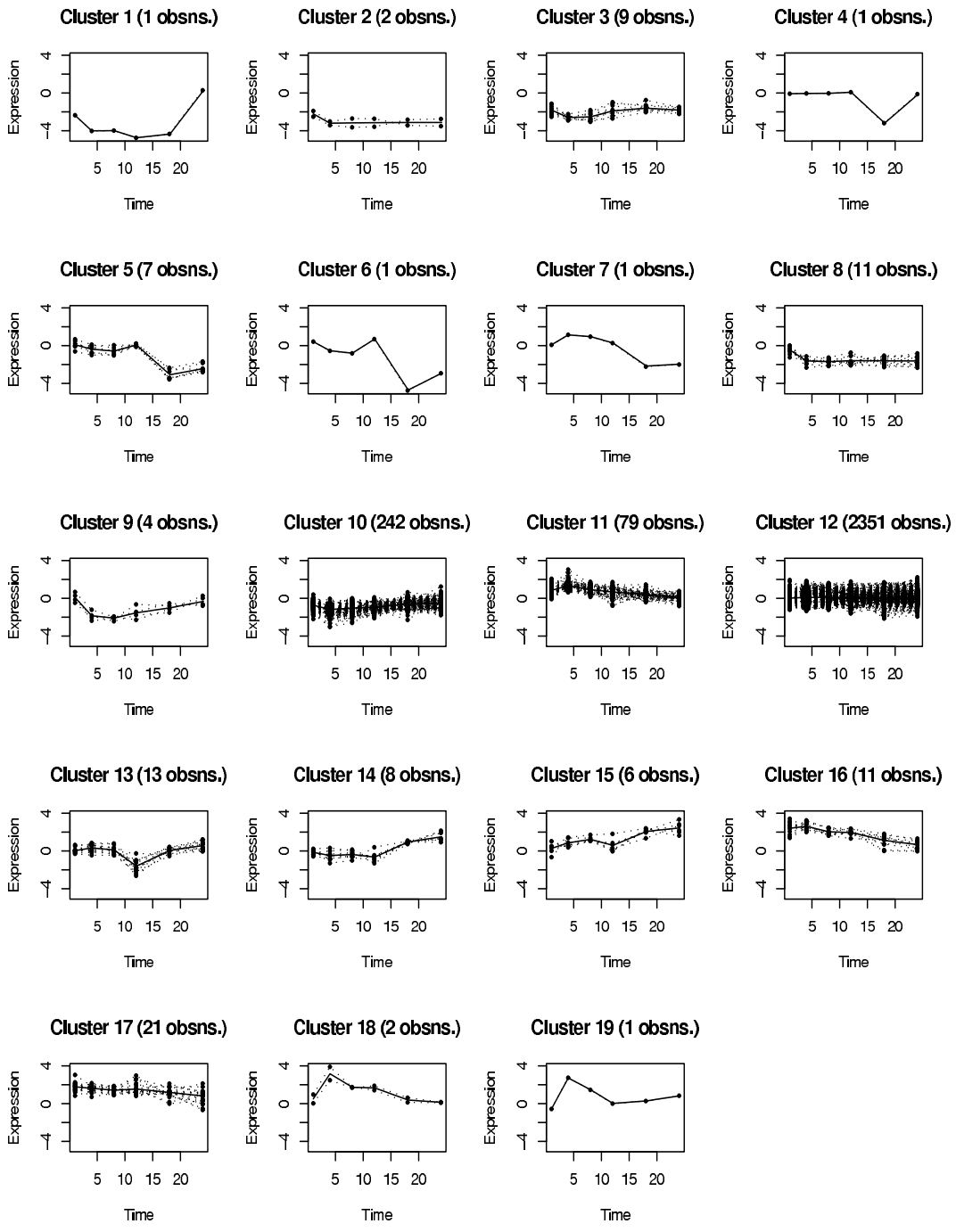

Figure 4: *Plots of optimal clusters after Euclidean agglomerative clustering of the* Salmonella typhi *data.*

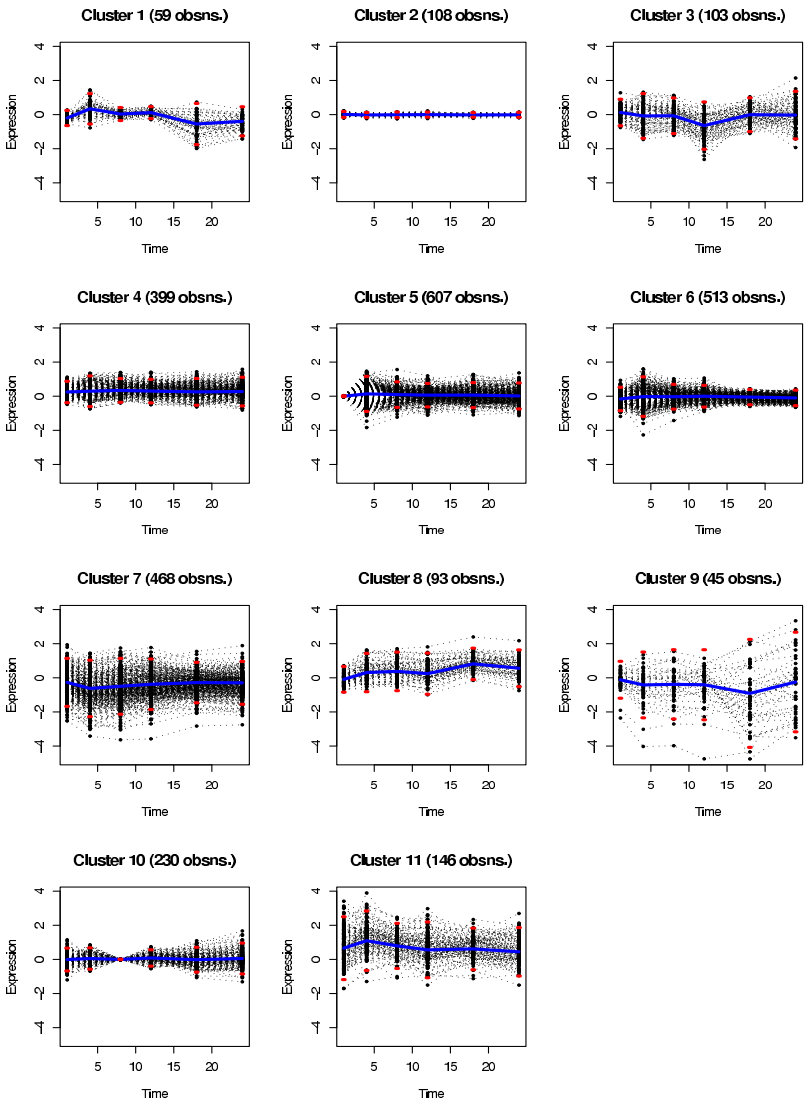

Figure 5: *Plots of optimal clusters after unconstrained maximum likelihood Gaussian clustering of the* Salmonella typhi *data using the* MCLUST *software (Fraley and Raftery, 1998).*

Cluster	I	R	PS	P	C	L	S	K	TR	TP	O	Total
1	3	1	1	0	0	0	3	3	3	2	9	25
2	0	1	0	1	0	0	0	1	0	1	2	6
3	0	1	1	1	0	0	1	0	0	1	1	6
4	1	1	0	1	1	0	0	0	0	0	3	7
5	0	0	1	1	0	1	1	0	2	4	6	16
6	0	0	0	0	0	0	0	0	1	0	1	2
7	2	9	2	9	0	2	5	1	3	2	21	56
8	2	6	1	4	1	0	0	0	1	2	12	29
9	1	2	0	1	0	0	1	1	2	5	9	22
10	1	0	1	0	0	1	0	1	0	2	0	6
11	0	1	0	0	1	0	0	0	0	0	2	4
12	0	2	0	0	0	0	0	0	0	0	1	3
13	9	2	0	0	1	0	2	0	3	2	8	27
14	0	1	0	0	0	0	0	0	1	0	1	3
15	3	0	1	2	1	1	0	1	1	4	9	23
16	0	1	0	0	0	0	1	1	0	2	1	6
17	0	0	0	0	0	1	0	0	0	0	0	1
18	1	9	1	5	5	2	8	1	1	3	12	48
19	0	10	1	17	2	2	7	2	6	2	17	66
Total	23	47	10	42	12	10	29	12	24	32	115	356

Table 1: Distribution of genes of known function for Bayesian clustering. Key: I = Immunity; R = Redox/Mitoch.; PS = Proteasome syst.; P = Protein metab.; C = Carbohydr. metab.; L = Lipid metab.; S = Str./Cytosk./Adh.; K = Kinases; TR = Transcription; TP = Transport; O = Other.