Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges

Nicholas A. Heard*†, Christopher C. Holmes‡§, David A. Stephens*, David J. Hand*, and George Dimopoulos†¶

*Department of Mathematics, Imperial College London, Huxley Building, 180 Queens Gate, London SW7 2AZ, United Kingdom; †Oxford Centre for Gene Function, Department of Statistics, University of Oxford, Oxford OX1 3QX, United Kingdom; ‡MRC Mammalian Genetics Unit, Medical Research Council, Harwell, Oxford OX11 0RD, United Kingdom; and §Department of Molecular Microbiology and Immunology, Johns Hopkins School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205

Edited by James O. Berger, Duke University, Durham, NC, and approved August 30, 2005 (received for review November 11, 2004)

We present a method for Bayesian model-based hierarchical coclustering of gene expression data and use it to study the temporal transcription responses of an Anopheles gambiae cell line upon challenge with multiple microbial elicitors. The method fits statistical regression models to the gene expression time series for each experiment and performs coclustering on the genes by optimizing a joint probability model, characterizing gene coregulation between multiple experiments. We compute the model using a two-stage Expectation-Maximization-type algorithm, first fixing the cross-experiment covariance structure and using efficient Bayesian hierarchical clustering to obtain a locally optimal clustering of the gene expression profiles and then, conditional on that clustering, carrying out Bayesian inference on the cross-experiment covariance using Markov chain Monte Carlo simulation to obtain an expectation. For the problem of model choice, we use a cross-validatory approach to decide between individual experiment modeling and varying levels of coclustering. Our method successfully generates tightly coregulated clusters of genes that are implicated in related processes and therefore can be used for analysis of global transcript responses to various stimuli and prediction of gene functions.

microarray | model-based clustering | Markov chain Monte Carlo | Expectation-Maximization

The A. gambiae mosquito is the major vector of human malaria, and its innate immune system, which is capable of killing Plasmodium parasites, is a key area of research (see refs. 1–4). Here, we analyze cDNA microarray data measuring gene transcription of an A. gambiae immune-competent cell line Sua1b in response to challenge with the microbial elicitors Escherichia coli, Salmonella typhimurium (S.t.), Micrococcus luteus (M.l.), Listeria monocytogenes (L.m.), and the yeast wall extract zymosan. The microarray slides used represented ≈2,400 unique genes and expression was assayed at 1, 4, 8, 12, 18, and 24 h after challenge as described in ref. 5. Fig. 1 shows the ranked expression values of the genes plotted as a heat-map.

Cluster analyses of microarray data are used to identify potentially functionally related groups of genes that are transcriptionally coregulated, that is, respond similarly at the transcriptional level to external stimuli and are most likely to be controlled by the same transcription factors and pathways. These analyses thereby allow assignment of putative functions to novel genes based on their shared cluster membership with genes of known functions. Because of the similarity of the different treatments here it is anticipated that we should observe some correlation in the gene transcription patterns across the different experiments. So for a cluster analysis of the genes in this setting, it is natural to consider jointly clustering (“coclustering”) the genes according to their response profiles across all of the experiments. Genes displaying similar expression profiles across the treatments, and hence being part of the same clusters, can be predicted to share similar functions or act in the same processes.

Cluster analysis in gene expression has mainly relied on Pearson’s correlation or Euclidean distance based methods for clustering expression measurements and profiles. Typically, hierarchical clustering is used to obtain a sequence of partitions of N data observations, ranging from a single group containing all observations to N groups each containing just one observation. The partitioning occurs in either a “divisive” fashion (one group to N groups) or by “agglomeration” (N groups to one group), with the hierarchy represented by a tree, or “dendrogram.” For the analysis of time course expression data, such as the data shown in Fig. 1, a multivariate clustering algorithm is required whereby the time dependency is respected.

In ref. 6, we developed a Bayesian model-based agglomerative scheme for clustering time course microarray data, using nonlinear regression splines to capture temporal variation within each cluster. The use of a Bayesian procedure allows us to compute measures of uncertainty for quantities of interest, such as the number of clusters in the data, and to report posterior probabilities that are comparable across all models, experiments, and computational methods. The use of nonlinear regression splines allows us to accommodate nonstationary time-dependence in the data as well as unequal sampling intervals and yet affords analytic calculation of marginal probabilities. See ref. 6 for further details. Model-based clustering of gene expression time series data has also been considered by amongst others, (7–11). Superior performance of nonstationary model-based clustering of time series data against the more standard clustering algorithms was established in ref. 6.

The key extension we present here involves the coclustering of multiple gene expression profiles obtained under related experimental treatments; to our knowledge, no explicit methodology has previously been developed for multiple experiment coclustering of time series; although of course people have considered two way clustering outside of time series application, such as in ref. 12. We extend the models of ref. 6 by allowing the coexpression of genes, not only within experiments at consecutive time points but now also between experiments. The benefits of joint modeling across parallel experiments are twofold. First, we obtain a more robust clustering of the genes, with the borrowing of strength between experiments working to stabilize the low signal to noise ratio inherent in current cDNA microarray study.

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: CV, cross-validation; EM, Expectation-Maximization; MC, Monte Carlo; L.m., Listeria monocytogenes; M.l., Micrococcus luteus; S.t., Salmonella typhimurium.

†To whom correspondence may be addressed. E-mail: n.heard@imperial.ac.uk or gdimopoulos@jhsph.edu.

© 2005 by The National Academy of Sciences of the USA
Second, we can characterize the dependence structure in coexpressed profiles across treatments.

**Methods**

The data are composed of the expression levels of \( n = 2,392 \) putative genes that are represented by DNA spots (probes) on a glass slide microarray (see ref. 5 for details). Of these 2,392 genes, 332 had putatively known function based on sequence similarity. The expression profiles shown in Fig. 1 relate to log2-transformed normalized ratios of intensities of hybridized samples (RNA from challenged vs. nonchallenged cells) to the spotted gene probes as described in ref. 5. Assays were done at six time points (1, 4, 8, 12, 18, and 24 h) after each of the microbial elicitor challenges. Experimental details are in ref. 5.

![Fig. 1. A. gambiae cell line gene-expression profiles at six consecutive time points after microbial challenge. Vertically, each block corresponds to one of the four bacterial or chemical challenges, in order from bottom to top as follows: S.t., M.I., L.m., and the yeast wall extract zymosan. Red corresponds to gene up-regulation; green corresponds to gene down-regulation. Genes are presented in random order.](image)

![Fig. 2. Histograms of empirical cross-correlations, computed by taking the set of gene-by-gene (Pearson) correlation coefficients across all possible pairs of experiments. The bias toward a positive correlation evident in this image indicates that there evidence for dependence between experiments.](image)
Inspection of Fig. 1 indicates that many of the genes display similar expression profiles across the different microbial elicitor challenges, suggesting that it may be possible to cocluster the genes using all of the data collected across the different treatments; this finding is made more explicit in Fig. 2, where for each pair of experiments we show the distribution across the genes of cross-correlations against the global mean of zero of the relative expression time series. That is, letting $y_{ge}$ be the $T$-vector time series of expression levels for gene $g$ in experiment $e$, for each pair of experiments $e_i$, $e_j$, $i \neq j$, we calculate

$$
y_{ge_g}, y_{ge_j} \quad g = 1, \ldots, N. \tag{1}
$$

Note that this quantity is the gene correlation measure used for clustering of a single experiment by ref. 12. Under the scenario of no underlying experimental correlation (easily simulated by permuting the gene indices for one of the pair of experiments) the distribution of Eq. 1 is symmetric about zero. The median values of the distributions for each of our experiment pairs also are shown in the plots in Fig. 2 and are all strongly positive.

A single experiment cluster analysis of the data from the first of the bacterial agents, S.t., appeared in ref. 6. There, Bayesian model-based clustering procedures were developed to accommodate the large sample size and nonstationary time-dependence structure of the data. Here, we extend that methodology to the case of multiple experimental conditions for joint clustering across the four microbial elicitors.

**Bayesian Modeling 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 by means of linear models and nonlinear basis functions. This model takes the form

$$
y_{ge} = X \beta_{ge} + \sigma_e e_{ge}, \quad g = 1, \ldots, N; e = 1, \ldots, E, \tag{2}
$$

where $X$ is a $T \times T$ design matrix made up of nonlinear basis functions evaluated at the experimental time points, $\beta_{ge}$ is a $T$-vector of basis coefficient parameters, $\sigma_e$ is an error variance, and $e_{ge}$ are independent standard Gaussian errors.

The regression model in Eq. 2 readily accommodates nonuniform sampling times and any nonstationarity in the data (which can be seen for our mosquito data most clearly on examination of the clustered data in Fig. 3). In ref. 6, we demonstrated that the use of fixed basis functions ($X$) with random coefficients and error variance induces a nonstationary stochastic process model for the underlying variation in expression for which we can analytically evaluate the marginal likelihood. The marginal likelihood forms the basis of our potential function of our agglomerative clustering scheme.

A fully Bayesian approach to clustering gene expression profiles was described in ref. 6. There, for any clustering of the genes, defined by a partition $C$ of the integers from 1 to $N$, genes in the same cluster of $C$ share common values for the regression coefficients and error variance. In particular, the use of regression models and a computationally efficient agglomerative algorithm for hierarchical clustering was established. As in ref. 6, we are again interested here in clustering together groups of genes that show similar patterns of expression over time or have a similar overall level of up or down regulation, or both. However, it should be noted that were we interested in identifying similar shaped, parallel, but perhaps quite separated, expression curves the model in Eq. 2 is simply extended to include a fixed or random effect term for each gene.

In this work, we demonstrate that these models and the clustering algorithm can be extended to the more general case where multiple, related time course profiles are available. Technical details of the extensions are in Supporting Text, which is published as supporting information on the PNAS web site.

The key extension proposed in this work is the modeling of cross-experiment correlation. This is achieved by means of the covariance matrix $V$ of the regression coefficients $\beta_{ge}$, which models the dependence between these parameters across the experiments for gene profiles in each cluster. Specifically, for $E$ experiments we take

$$
V = \Sigma \otimes I_T, \quad \text{with probability } q
$$

$$
\begin{bmatrix}
D_S \otimes I_T \\
D_E \otimes I_T \\
\end{bmatrix}, \quad \text{with probability } 1 - q. \tag{3}
$$

It transpires that, even under this extended model, the basic approach of ref. 6 can still be implemented. Full details are given in Supporting Text.

**Bayesian Hierarchical Clustering.** As in ref. 6, the method proposed uses the Bayesian posterior distribution on the unknown partition $C$ given the expression data [and now conditional on the experimental covariance parameters $(\Sigma, q)$] as a potential function for agglomerative clustering. The improved clustering performance under this approach and its speed of computation are outlined for the single experiment case in ref. 6.

The full algorithm used to implement the Bayesian hierarchical clustering here incorporates a previously undescribed Markov chain Monte Carlo-based approximation to the Expectation-Maximization (EM) algorithm to enable us to jointly learn about $C$ and $(\Sigma, q)$ and is described in detail in Supporting Text. The algorithm can be viewed as a further approximation to the MCEM method of ref. 13.

It should be noted here that for fully Bayesian inference on the joint parameter space of $(C, \Sigma, q)$, we should, for example, calculate expectations or find maxima over the whole joint posterior distribution of these parameters. However, with such large, high-dimensional data as encountered here, standard Markov chain Monte Carlo methods for exploring the resulting vast joint space prove impractical and in any case severely increase the computation time. As with conventional agglomerative clustering, our algorithm seeks to find a path through the vast model space, providing a (visual) hierarchical decomposition of the association between objects and an ordering of the objects at the base of the tree. However, this structured path comes at a price, and hierarchical clustering is unlikely to pass through the global optimal cluster configuration. For further discussion, again see Supporting Text.

**Model Choice.** We now consider the decision problem of choosing when to perform coclustering. That is, we wish to ascertain whether it is appropriate to cluster the genes using the data from all of the experiments together, treat each experiment separately, or perhaps cocluster across some, but not all, of the experiments. Coclustering cannot always be assumed to be beneficial, because there could be underlying differences in the function groupings of the genes if the treatments are sufficiently different in their action.

Four microbial elicitor challenges were in the study analyzed here, so there are 15 ways of partitioning these experiments into nonempty groups. For any given partition of the experiments (S.t.,
L.m., M.l., zymosan}, each set within the partition can represent a group of experiments we wish to cluster jointly, although independently from the remaining experiments in the other sets of the partition. The 15 partitions thus represent every possible level of joint modeling of the four experiments, including the special cases of modeling all of the experiments individually or all jointly.

A further advantage of using a model-based clustering technique is that we are able to perform a cross-validation (CV) study to identify an appropriate joint-modeling structure by comparing predictive power. Taking any partition of the experiments, we sequentially leave out one of the interior time points in turn from the experimental data; for each group in the partition, we then run our Markov chain Monte Carlo/EM algorithm to find an optimal clustering model and measure how well that model predicts the data points that have been left out. For this measure, we use the log-predictive density (see Supporting Text for further details). This procedure is repeated for each time point and then for each of the possible partitions of the experiments.

Note that we only have four experiments to consider here, so there is no difficulty in looking exhaustively at all of the possible partitions. For larger numbers of experiments, this approach would become prohibitive, and instead one could resort once more to an agglomerative clustering procedure, although now on the experiments, to find a locally optimal joint modeling structure.

### Results

We present the results of our cluster analysis of the *A. gambiae* cell line transcription responses to microbial challenge described above. It is common practice in a cluster analysis of microarray data to filter out all genes whose observed increase or decrease in expression relative to unchallenged cells is never greater than some significance threshold, usually a minimum of 2-fold up-regulation or 2-fold down-regulation (see, for example, ref. 5). For our model-based method, this preprocessing is not necessary, with these nonregulated genes naturally grouping together to form low variance clusters. Thus, we included all of the data in our analysis.

First, to find the appropriate level of joint modeling, we performed a CV study as described above. The best experimental clustering structure found was to model all of the experiments together, and the worst was to model each experiment separately. Therefore, for this study, the borrowing of strength through coclustering achieves more robust clustering. The CV scores for these two models are shown in Table 1, indicating overwhelming support for our coclustering approach.

In fact, the best two- and three-cluster models form a hierarchy, meaning the optimal clusterings at any desired number of clusters could have been found here by agglomerative clustering of the experiments; the merger sequence would have been

1. \(\{S.t\}, \{M.l\}, \{L.m\}, \{zymosan\} \quad (−706.220)\)
2. \(\{S.t\}, \{M.l\}, \{L.m., zymosan\} \quad (−188.080)\)
3. \(\{S.t\}, \{M.l., L.m., zymosan\} \quad (79.2775)\)
4. \(\{S.t., M.l., L.m., zymosan\} \quad (282.058)\)

![Hierarchically clustered expression profiles of bacterial or chemical experiment data. From bottom to top: S.t., L.m., M.l., and zymosan.](image)

**Table 1. Log predictive densities**

<table>
<thead>
<tr>
<th>(t) left out, h</th>
<th>Individual (^1)</th>
<th>Joint (^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>−44.146</td>
<td>799.275</td>
</tr>
<tr>
<td>8</td>
<td>389.805</td>
<td>1056.590</td>
</tr>
<tr>
<td>12</td>
<td>−1366.335</td>
<td>−184.610</td>
</tr>
<tr>
<td>18</td>
<td>−1892.495</td>
<td>−543.024</td>
</tr>
<tr>
<td>Average</td>
<td>−706.220</td>
<td>282.058</td>
</tr>
</tbody>
</table>

\(^{1}\)The particular time point (h) left out in the CV.

\(^{2}\)The log predictive possibilities through modeling the experiments separately and using coclustering.
where the numbers in parentheses are the CV scores for that experimental clustering.

For a simple comparison, we also tried standard hierarchical agglomerative clustering of the four experiments by simply concatenating the time series of all of the genes into one long vector for each experiment. Dendrograms under different distance metrics and linkage choices are given in Supporting Text. In particular, under the Euclidean distance metric we obtained the same optimal experimental merger hierarchy as above using single or average-link clustering.

Having identified the optimal experimental clustering, our Bayesian hierarchical clustering method was used to learn jointly about the experimental covariance matrix and the gene clustering (see Supporting Text for details). The reordered gene expressions after the final hierarchical clustering are shown in Fig. 3, and the estimated correlation matrix giving rise to this clustering is given in Table 2. Note that strong, positive correlations have been found between all of the experiments, a finding supported by our earlier exploratory plots in Fig. 2. This correlation is much more apparent in Fig. 3 than it was for the unordered data plot, with genes generally up-regulated across the four treatments on the left-hand side of the plot and down-regulated across the four treatments on the right-hand side of the plot. The two dendrograms plotted in Fig. 3 indicate the optimal hierarchy of mergers for the genes (horizontal axis) and challenges (vertical axis).

The optimal clustering found had 159 clusters, many more than we get when clustering on the expressions of a single experiment (6). This increase reflects the extra information contained in the combined gene expression time series, so that genes that may have appeared similar by chance under a smaller number of sampling points are revealed as quite different. However, note that if a smaller number of clusters is sought, then because the method is hierarchical such a clustering can be read off from the dendrogram at any desired level. In fact, the partitions for up to, say, \( \leq 50 \) clusters from the optimal clustering had only slightly lower posterior probability and thus offer many plausible alternatives.

**Discussion**

We have tested our clustering method on a data set of mosquito immune responses to microbial challenge. Of the 2,392 assayed genes, 332 had predicted functions and 22 were related to the mosquito immune system, based on DNA-sequence similarity. Three neighboring (in the sense of the dendrogram) clusters in the optimal clustering, clusters 1, 4, and 5, contained 16 genes with predicted function, of which 11 belong to the immunity functional class (see Fig. 4. This result compares favorably with

<table>
<thead>
<tr>
<th></th>
<th>S.t.</th>
<th>L.m.</th>
<th>M.l.</th>
<th>Zymosan</th>
</tr>
</thead>
<tbody>
<tr>
<td>S.t.</td>
<td>1</td>
<td>0.782</td>
<td>0.624</td>
<td>0.614</td>
</tr>
<tr>
<td>L.m.</td>
<td>0.782</td>
<td>1</td>
<td>0.702</td>
<td>0.876</td>
</tr>
<tr>
<td>M.l.</td>
<td>0.624</td>
<td>0.702</td>
<td>1</td>
<td>0.842</td>
</tr>
<tr>
<td>Zymosan</td>
<td>0.614</td>
<td>0.876</td>
<td>0.842</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. Estimated correlation between the experiments from the coclustering algorithm

**Fig. 4.** Three coregulated gene clusters from Bayesian agglomerative clustering (clusters 1, 4, and 5) that are highly enriched with putative immune genes. Each \( 2 \times 2 \) block of plots corresponds to the gene-expression profiles under the four bacterial and chemical challenges. Genes contained in each cluster are presented in Fig. 6, which is published as supporting information on the PNAS web site. Log2-transformed fold regulation is indicated on the vertical axis, and assayed time points are indicated on the horizontal axis.
the single bacterial experiment analysis of ref. 6, where a single immune defense cluster was identified, with 9 of its 27 genes of known function being immunity related.

The three clusters are therefore of particular interest because of the tight coregulation of various immune and other genes that are likely to be involved in common defense mechanisms. Indeed, cluster 1 comprises two pattern-recognition receptors, GNBPB1 and PGRPLB, belonging to the Gram-negative bacteria-binding gene family and the peptidoglycan-recognition protein gene family, respectively. GNBP and PGRPs have been shown to function in the same mechanism implicated in the activation of the Toll immune-signaling pathway in Drosophila melanogaster (14).

Cluster 1 also comprises a leucine-rich repeat (LRR) domain-containing protein, which belongs to a gene family comprising several putative Toll receptor genes. This LRR-domain transcript also may be implicated in the pathway activated by GNBP and PGRP receptors. The Toll pathway controls activation of antimicrobial effectors such as cecropin, which is found in cluster 1 (14). Hence, cluster 1 comprises putative pattern recognition receptors, signaling factors, and an effector gene that may be part of the same immune response process (Fig. 5). Cluster 4 comprises, among other immune genes, a thioester containing protein TEP4 and a LRR-domain protein that may function as pattern-recognition receptors. Other genes in cluster 4 are two serine proteases, CLIPD1 and ENSANGG000013355, and a phenoloxidase gene, PPO5. Phenoloxidases are implicated in melanization defense reactions and are activated by serine protease cascades that, in turn, are triggered by recognition of pathogens by pattern recognition receptors (15, 16). The tight coregulation of these components may be indicative for functional relations in the activation of melanization reactions (Fig. 5). Cluster 5 comprises an antimicrobial peptide gene gambicin and a thioester containing protein gene TEPIV. Gambicin and TEP1 have been shown to possess anti-Plasmodium activity in addition to antimicrobial action in A. gambiae (17, 18).

We have demonstrated a successful implementation of Bayesian hierarchical coclustering on A. gambiae immune-responsive genes, grouping them into putative functionally related clusters. Functional relations between the identified genes can now be tested and validated through other experimental approaches. By examining predictive performance through CV studies, we have seen how clustering can be improved by joint modeling across different, but related, experimental conditions.

Sophisticated computational methods (Markov chain Monte Carlo and approximate EM) enabled us to implement a model that learned about the degree of correlation between the experiments. The algorithm we have proposed also yields a parameter, q in Eq. 3, that measures the probability that a given gene has correlated response across experiments; of the 159 clusters, 109 (68% and including the immune defense clusters) attributed probability of > 0.5 to a model where there was correlation between the experiments.

The method is readily implementable. In each iteration of the EM algorithm (see details in Supporting Text), for the Monte Carlo E-step we ran 20,000 Markov chain iterations, half of which were discarded as a burn-in, and this process along with the clustering M-step took just over 4 and 5 minutes, respectively, for the data analyzed here on a 2-GHz processor PC. Twenty iterations of the overall EM algorithm were performed, and this proved more than sufficient, with the expected experimental covariance matrix stabilizing fairly well after as few as five iterations. C++ code implementing the full EM algorithm, along with an example data set and shell script, are freely available from N.A.H. upon request.

We thank both reviewers and the editor for their comments, which led to considerable improvements in the manuscript. N.A.H. was supported by Wellcome Trust Grant 065822. C.C.H. is partly supported at the Oxford Centre for Gene Function by the U.K. Medical Research Council. G.D. was partly supported by the Johns Hopkins Malaria Research Institute and National Institutes of Health/National Institute of Allergy and Infectious Diseases Grant IR01AI059492-01A1.