Cluster-based network model for time-course gene expression data

LURDES Y. T. INOUE*

Department of Biostatistics, University of Washington, F-600 Health Sciences Building, Campus Mail Stop 357232, Seattle, WA 98195, USA and
Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, MP 665, PO Box 19024, Seattle, WA 98109, USA
linoue@u.washington.edu

MAURICIO NEIRA

The Prostate Centre, Vancouver General Hospital, 2660 Oak Street, Vancouver, British Columbia, Canada

COLLEEN NELSON, MARTIN GLEAVE

The Prostate Centre, Vancouver General Hospital, 2660 Oak Street, Vancouver, British Columbia, Canada and
Department of Surgery, University of British Columbia, Vancouver, British Columbia, Canada

RUTH ETZIONI

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, MP 665, PO Box 19024, Seattle, WA 98109, USA

SUMMARY

We propose a model-based approach to unify clustering and network modeling using time-course gene expression data. Specifically, our approach uses a mixture model to cluster genes. Genes within the same cluster share a similar expression profile. The network is built over cluster-specific expression profiles using state-space models. We discuss the application of our model to simulated data as well as to time-course gene expression data arising from animal models on prostate cancer progression. The latter application shows that with a combined statistical/bioinformatics analyses, we are able to extract gene-to-gene relationships supported by the literature as well as new plausible relationships.

Keywords: Bayesian network; Bioinformatics; Dynamic linear model; Mixture model; Model-based clustering; Prostate cancer; Time-course gene expression.

*To whom correspondence should be addressed.
Time-course gene expression data arise when microarray experiments are performed over time. Such experiments allow for observing cellular mechanisms in action while trying to break down the genome into sets of genes involved with related processes.

Several methods have been devised to analyze time-course microarray data. Some methods focus on the question of identifying genes with differential expression over time (see, for example, Storey and others, 2005). Other methods focus on aggregating genes with similar temporal profiles (see, for example, Ramoni and others, 2002) or on identifying networks (see, for example, Murphy and Mian, 1999).

Some authors have explored connections between these seemingly different tasks, for example, using methods that allowed them to aggregate genes and define networks over the aggregated genes. Holter and others (2001), for example, used singular value decomposition to extract few characteristic modes that determined essential features of the gene expression patterns and used them to infer networks. Ong and others (2002) used instead biological knowledge to aggregate genes, thought to be co-regulated, over which they defined a network. In both cases, the underlying clustering allows for dimension reduction and, therefore, inferences on regulatory networks—a problem that is often computationally intractable with a large (or even moderate) number of genes.

In this paper, we provide a model-based approach to unify the processes of inferring networks and clustering genes. Note that previous work (Holter and others, 2001; Ong and others, 2002) used deterministic clusters. We provide a probabilistic framework for inferring clusters from gene expression profiles. Genes within the same cluster are expected to share a similar expression profile. We build a network over clusters using state-space models.

We apply our model to microarray data arising from time-course experiments in the Shionogi tumor model (see Section 4.2). Our goal in the analysis is to identify genes with similar expression profiles and relationships between genes at the cluster level.

This manuscript is organized as follows: we start with a review of methods to clustering and networks in Section 2. In Section 3, we describe our model. We discuss applications of our model in Section 4; first, using simulated data in Section 4.1 and then, in Section 4.2 we apply our model to the time-course microarray data arising from animal models on prostate cancer progression. Finally, in Section 5 we provide a discussion.

2. REVIEW OF METHODS

Cluster analysis allows for grouping genes on the basis of similarities between them. This technique has been widely used for analyzing gene expression data as it provides clues to unknown gene function as inferred from clusters of genes similarly expressed. A wide range of clustering methods have been used including hierarchical clustering (Eisen and others, 1998), the k-means algorithm (Tavazoie and others, 1999), self-organizing maps (Tamayo and others, 1999), singular value decomposition (Wall and others, 2001), and support vector machines (Brown and others, 2000). However, none of these methods provides formal inferences. Alternatively, model-based approaches provide a framework for addressing this issue with the use of probabilistic models to describe clusters (Fraley and Raftery, 2002). Most applications using time-course microarray data are based on models that account for the serial dependence in the gene expression over time (Ramoni and others, 2002; Schliep and others, 2003; Wakefield and others, 2003; Zhou and Wakefield, 2006).

Network models are now the focus of a growing number of researchers concerned with discovering novel gene interactions and regulatory relationships from expression data. Friedman and others (2000) were among the first to use a Bayesian network to describe interactions between genes. Recent papers applied dynamic Bayesian networks (DBNs) to time-course microarray data. Such models extend Bayesian networks in that they allow for cyclic temporal relationships between genes. To our knowledge, while
Cluster-based network model

Murphy and Mian (1999) first proposed the use of DBN to model time-course microarray data, Ong and others (2002) were among the first to implement DBNs for this purpose. Other applications include, for example, Perrin and others (2003) and Beal and others (2005).

Our modeling efforts bring together clustering and network modeling to analyze time-course microarray data. This approach allows us to capture within-cluster relationships as well as relationships between clusters of genes. From a computational viewpoint, by looking at relationships defined at the cluster level, we reduce the number of relationships to be estimated in the network, making the problem of inferring networks more tractable when considering large sets of genes.

Our model is motivated by earlier experimental work suggesting that biological networks are modular (Barabasi and Oltvai, 2004; Petti and Church, 2005) with modules defined over groups of genes, proteins, and other molecules that are involved in a common subcellular process. The underlying idea for clustering genes is that if co-regulation indicates shared functionality, then clusters defined to this level of abstraction represent biological modules. This idea underlies our model where modules are clusters of genes sharing similarity in the patterns of expression. Specifically, we use mixture models for inferences on the gene-cluster membership and state-space models to describe the time-course gene expression. Our approach utilizes the general framework of state-space models to probabilistically infer networks—the key aspect of it is that we target at relationships between sets of genes (clusters) rather than gene-to-gene relationships. We note, however, that a model for gene-to-gene relationships is a special case in which each gene defines its own cluster.

Our networks explicitly incorporate time-delayed relationships, allowing, for example, expression within a cluster of genes at a given time to affect expression within the same or a different cluster at a subsequent observation time. The biological motivation for this comes from earlier work on the existence of time-delayed relationships as seen, for example, between transcription factors and their targets (see, for example, Qian and others, 2001; Yu and others, 2003). This idea also underlies our network. Specifically, in our model we seek to identify first-order relationships where the expression of genes in a given cluster at time \( t \) affects the expression of genes in the same or in a different cluster at time \( t + 1 \).

3. Model specification

3.1 Model

We consider aggregating genes into one of \( K \) clusters. Here \( K \) is a fixed number such that \( K \ll N \). Let \( Y_{it} \) denote the expression level of gene \( i \) at time \( t \), where \( i = 1, \ldots, N \) and \( t = 1, 2, \ldots, n \). Define \( Y_i = (Y_{i1}, \ldots, Y_{in})', Y_t = (Y_{1t}, \ldots, Y_{Nt})', \) and \( Y = (Y_1', \ldots, Y_N')' \). Similarly, let \( \theta_{kt} \) denote the parameter associated with the expression level in cluster \( k \) (\( k = 1, \ldots, K \)) at time \( t \). Let \( \theta_k = (\theta_{k1}, \ldots, \theta_{kn})' \) denote the parameter vector associated with the expression profile in cluster \( k \), while \( \theta_t = (\theta_{1t}, \ldots, \theta_{Kt})' \) denote the parameter vector associated with time \( t \). Moreover, \( \theta = (\theta_1', \ldots, \theta_K')' \). In what follows we use \( f(\cdot) \) to represent both probability density functions and probability functions.

We assume that the observation vector for any gene \( i \) comes from a mixture of multivariate distributions, that is,

\[
f(y_i|\theta) = \sum_{k=1}^K \pi_k f(y_i|\theta_k).
\]

Here clusters are defined over expression profiles. The mixing probability \( \pi_k \) is the probability that the profile of gene \( i \) belongs to cluster \( k \).

We denote cluster membership for the \( i \)th gene using a latent variable \( Z_i \). The latent variable \( Z_i \) is assumed to be conditionally independent and identically distributed with a multinomial distribution, that is, \( Z_i \sim \text{Multinomial}(K, \pi) \), where \( \pi = (\pi_1, \ldots, \pi_K)' \). Conditional on the latent variable, we break the
mixture with
\[ f(y_i | Z_i = k, \theta) = f(y_i | \theta_k). \]
For any gene \( i \) that falls within cluster \( k \), we model its expression profile at time \( t \) with the “observation equation”
\[ y_{it} = \theta_{kt} + v_t, \quad v_t \sim \text{Normal}(0, V_{kt}). \] (3.1)

Next we can assume that parameters \( \theta_{kt} \) evolve over time according to the “evolution equation”
\[ \theta_{kt} = \theta_{k(t-1)} + \omega_t, \quad \omega_t \sim \text{Normal}(0, W_k). \] (3.2)

We assume that the sequence of errors \( v_t \) and \( \omega_t \) are independent. Moreover, both are normally distributed with mean zero and with cluster-specific variances, \( V_{kt} \) and \( W_k \), respectively. The model defined with (3.1) and (3.2) is known as the univariate state-space model (see, for example, West and Harrison, 1997).

The above model allows us to describe, dynamically, the evolution of the expression profiles within a given cluster. We build on it to define a network over the clustering parameters. We replace the evolution equation (3.2) with the following equation to describe the network:
\[ \theta_{kt} = \beta_{k(t-1)}^t + \omega_t, \quad \omega_t \sim \text{Normal}(0, W_k), \] (3.3)
where \( \beta_{k} = (\beta_{k1}, \ldots, \beta_{kK})' \) and let \( \beta = (\beta_{11}, \ldots, \beta_{1K})' \). We call \( \beta \) the network parameters. In other words, the above model implies that the mean expression level at time \( t \) in cluster \( k \) is a “weighted” average of the mean expression level of each cluster at the previous observation time except that we are not restricting the weights to be positive and to sum up to one. Moreover, the “weights” are cluster specific, but not time dependent. (Figure 1 of the supplementary material available at Biostatistics online [http://www.biostatistics.oxfordjournals.org] schematically shows the relationships between clusters described by (3.1) and (3.3).)

A multivariate representation, combining data from all genes at time \( t \) and conditional on the gene-cluster membership, is provided by the following equations:
\[ Y_t = F'\theta_t + v_t, \quad v_t \sim \text{Normal}(0, V_t), \] (3.4)
\[ \theta_t = G\theta_{t-1} + \omega_t, \quad \omega_t \sim \text{Normal}(0, W), \] (3.5)
where \( F \) and \( G \) are defined as follows. Suppose that gene \( i \) (\( i = 1, \ldots, N \)) belongs to cluster \( k \) (\( k = 1, \ldots, K \)). Then the \( i \)th row of matrix \( F' \) is such that
\[ (0, \ldots, 0, \underbrace{1}_{k\text{th column}}, 0, \ldots, 0), \]
that is, it is filled with zeroes except in column \( k \) where it takes the unit value. The matrix \( G \) contains the network parameters in the following form:
\[ G = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1K} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{K1} & \beta_{K2} & \cdots & \beta_{KK} \end{bmatrix}. \]

Moreover, \( V_t \) is a diagonal matrix with elements \( V_{kt} \) in the diagonal, that is, if gene \( i \) is in cluster \( k \), then \( V_t \) in its \((i, i)\) cell has element \( V_{kt} \). Moreover, \( W = \text{diag}(W_1, \ldots, W_K) \). This defines a multivariate state-space model or multivariate dynamic linear model.
3.2 Estimation

We take a Bayesian approach to estimate the above model. The full parameter vector is \( \xi = (\theta, \beta, \pi, z, V, W)' \), where \( V \) denotes the vector with all observation variances \( V_{kt} \). We fully specify the Bayesian model with priors on \( \theta_0 \), that is, the prior on the state parameter at time 0, and on parameters \( \beta, \pi, V, W \). Thus, the posterior distribution, given \( y \), follows from

\[
\begin{align*}
  f(\xi | y) &= f(\theta, \beta, \pi, z, V, W | y) \\
  &\propto f(y | \theta, z) f(z | \pi) f(\theta | \beta) f(\beta) f(\pi) f(V) f(W),
\end{align*}
\]

where, in the above, we are assuming that \( \theta_0, \beta, \pi, V, W \) are a priori independent. In the applications discussed in Section 4, we specifically assume that

\[
\begin{align*}
  \theta_0 &\sim \text{Normal}(m_0, C_0), \\
  \beta &\sim \text{Normal}(m_\beta, S_\beta), \\
  \pi &\sim \text{Dirichlet}(p), \\
  1/V_{kt} &\sim \text{iid Gamma}(a_v, b_v), \\
  1/W_k &\sim \text{iid Gamma}(a_w, b_w), \quad k = 1, \ldots, K, \quad t = 1, \ldots, n.
\end{align*}
\]

(In the above, \( \text{Gamma}(a, b) \) denotes a Gamma distribution with parameters \( a \) and \( b \)—parametrized such that the mean is \( a/b \).)

Although the model does not provide a closed-form posterior distribution, we can estimate it using Markov chain Monte Carlo methods. Using the above priors, the full conditionals are available in closed form, and so we implemented the Gibbs sampler (Geman and Geman, 1984). In particular, for the estimation of the cluster profile parameters \( \theta \), we used the forward filtering backwards sampling algorithm (see, for example, Frühwirth-Schnatter, 1994; Carter and Kohn, 1994). The full conditional distributions for the model parameters are provided in the Appendix A of the supplementary material available at Biostatistics online (http://www.biostatistics.oxfordjournals.org).

4. Applications

In this section, we apply our method to simulated data and to a data set arising from animal studies on progression of prostate cancer.

4.1 Simulation study

Simulation 1.

Simulation of the data. We simulated data considering \( K = 4 \) clusters. In each cluster, we simulated 250 trajectories over 20 equally spaced time points. In our simulation, cluster-specific parameters were fixed with \( \theta_t = G_0 \theta_{t-1} \), where \( \theta_0 = (1.0, 0.0, 0.0, 0.0)' \) and

\[
G = \begin{bmatrix}
  0.8 & 0.0 & -0.1 & 0.0 \\
  0.8 & 0.6 & 0.0 & 0.0 \\
 -0.8 & 0.4 & 0.8 & 0.0 \\
 -0.6 & 0.0 & 0.4 & 0.2 \\
\end{bmatrix}.
\]

We simulated data \( y_{it} \) at time \( t \) using \( y_{it} = \theta_{kt} + \nu_t, \quad \nu_t \sim \text{Normal}(0, V) \), with \( V = 0.01 \) (we will also present results using, alternatively, \( V = 0.05 \)).

Figure 1 shows the cluster-specific trajectories in panel (a) with simulated trajectories in panel (b) using \( V = 0.01 \). We note that clusters 3 and 4 have very similar profiles. The basic premise of our model is that genes within the same cluster share similar expression profile. Thus, in using our model, genes from the simulated clusters 3 and 4 could, depending on the variability in the observations, be in a single
cluster. We examine this in the next section. We also note that trajectories simulated with $V = 0.05$ have more overlap across clusters after time 10 and higher within-cluster variability (figure not shown).

For our estimation, we assume that $\theta_{x0}$ has, a priori, a multivariate normal distribution with null mean vector and that the covariance matrix is diagonal with all variances equal to 0.1. Likewise, we assume that the network parameters are, a priori, all independent normals centered at zero and with variance 0.1. Parameter $\pi$ has a Dirichlet(1, 1, 1) (when $K = 3$) prior distribution. Moreover, we assume that both $1/V_k$ and $1/W_k$ are independent and identically distributed Gamma(1, 1). Priors on $\pi$, $1/V_k$, and $1/W_k$ are somewhat noninformative relative to the data. We call this specification as prior 1. Sensitivity to our prior choice is discussed in the next section. Specifically, we discuss sensitivity of our analysis to prior assumptions on the network parameters using prior 2, as well as on the state parameter at time 0 using prior 3, both defined with a five-fold increase in the respective prior variances.
Cluster-based network model

Results. First, let us consider simulated trajectories from clusters 1, 2, and 3. In this case, all profiles have different shapes. Our model correctly separates all genes. This was observed under both values of the observation variance (that is, $V = 0.01$ and $V = 0.05$).

Figure 2 in the supplementary material available at Biostatistics online (http://www.biostatistics.oxfordjournals.org) shows the 95% credible intervals (CIs) for the network parameters under different priors and using data simulated with $V = 0.01$ (top panels) or $V = 0.05$ (bottom panels). We note that, under all three priors, the estimation was fairly insensitive to a higher variance $V$ in the observations. This is because, even though the observation variance was increased by five-fold, information in the prior distributions and in the trajectories allows for cluster discrimination. We also note that the true values of the network parameters (represented with asterisks in the figure) are all included in the 95% CIs.

Under prior 1, there is some uncertainty on the network parameters, with most of the CIs covering the null value except for relationships $1 \rightarrow 2$, $2 \rightarrow 2$, and $3 \rightarrow 3$ signifying that the state evolution in cluster 2 is influenced by the states in the previous period in both clusters 1 and 2. State evolution in cluster 3 only depends on the state of cluster 3 in the previous time period. We also note that there is also a borderline relationship $1 \rightarrow 3$.

Increasing the variability on the network parameters (under prior 2) or the state parameters at time 0 (under prior 3) did not change our ability to distinguish clusters. In both cases, we see that the estimates of the network parameters change. Under prior 2, we still detect the relationships $1 \rightarrow 2$, $2 \rightarrow 2$, and $3 \rightarrow 3$, but we also detect relationships $1 \rightarrow 1$, $1 \rightarrow 3$, and $2 \rightarrow 3$. Under prior 3, we detect relationships $1 \rightarrow 2$, $2 \rightarrow 2$, $3 \rightarrow 3$, and a borderline relationship $1 \rightarrow 3$.

Next, we considered simulated data from all clusters. Using $K = 4$ and data simulated with $V = 0.01$ all genes were correctly separated. Panel (c) in Figure 1 shows the estimated posterior cluster-specific profiles while panel (d) shows the 95% CIs for the network parameters. We note that most of the CIs cover the null value with the exceptions being on the relationships $1 \rightarrow 2$, $2 \rightarrow 2$, and $3 \rightarrow 3$. When analyzing the data simulated with $V = 0.05$, however, clusters 3 and 4 were aggregated into a unique cluster while the “pseudo”genes from the remaining clusters were correctly separated. As we indicated in our earlier discussion, this can be explained by the fact that the profiles in clusters 3 and 4 are similar and the relatively large variability in the simulated data does not allow for cluster discrimination.

Simulation 2.

Simulation of the data. In our simulation study in the previous subsection, our model identified (first-order) relationships between clusters. The issue arises as to whether our model can still capture higher-order relationships or even nonlinear relationships. We investigate this issue using data simulated under two settings. For each setting, we simulated 250 trajectories in each of the three clusters for $t = 1, \ldots, 20$ time points.

In the first simulation setting, trajectories in the first cluster are described by $h_1^{(1)}(t) = \cos(0.2 \times t + 0.5)$, trajectories in the second cluster by $h_2^{(1)}(t) = \cos(t/3+3)$ and finally, in the third cluster, trajectories are described by $h_3^{(1)}(t) = h_2(t - 2) - 1.2h_1(t - 1)$. Thus, cluster 3 is related to both clusters 1 and 2 through first- and second-order relationships. In the second setting, trajectories in cluster 1 are described by $h_1^{(2)}(t) = \cos(0.25 \times t)$, in cluster 2 by $h_2^{(2)}(t) = h_1^{(2)}(t - 8)$ and in cluster 3 by a nonlinear function of $h_1^{(2)}$ and $h_2^{(2)}$, namely, with trajectories described by $h_3^{(2)}(t) = \log(h_1^{(2)}(t) + h_2^{(2)}(t)/2 + 1.0) + 0.8$. Thus, there is a linear relationship between clusters 1 and 2, and cluster 3 is related to both clusters 1 and 2, but the relationship is nonlinear. Random noise was added to all trajectories.

Results. Figure 2 shows the simulated data and results from our model using data from setting 1, with results shown in panel (a), and data from setting 2, with results shown in panel (b). Specifically, the figure shows the simulated trajectories in the first row of panels (a) and (b) where the added noise had variances...
0.1, 0.2, and 0.4. The mid row of panels (a) and (b) shows the estimated profiles along with the 95% CIs. Finally, the last row of panels (a) and (b) shows the 95% CIs for the network parameters.

In both simulation settings, the estimation of the profiles and network parameters was fairly robust to the variances in the trajectories. Under setting 1, our model identified relationships between clusters 1 $\rightarrow$ 3 and 2 $\rightarrow$ 3, that is, the model was able to identify not only the first-order relationship but also the second-order relationship in the simulated data. Next, under setting 2, our model identified relationships between clusters 2 $\rightarrow$ 1, 3 $\rightarrow$ 2, and a borderline relationship 1 $\rightarrow$ 3. This simulation shows that our model can identify relationships between clusters beyond the first-order and linear relationship in (3.3). We note that the directions of the relationships are not all in agreement with the true functional relationships. However, as we look at the cluster profiles the inferred significant relationships seem plausible. To see this, note that, for example, features in cluster profile 1 are delayed and negatively correlated with those seen in cluster profile 2; in particular, for a maximum in the profile in cluster 2, there is at a later time a minimum in the profile in cluster 1 which makes the relationship 2 $\rightarrow$ 1 plausible.

4.2 Case study

Background. The development of the normal prostate as well as tumors of the prostate depends on the ratio of cells that proliferate to those that die and this dynamic depends on male hormones known as androgens (e.g. testosterone). This has motivated prostate cancer treatments that, by lowering the level of male hormones, cause massive tumor cell death greatly reducing tumor size. However, despite this initial response, cells become independent and the tumor grows again, a phenomenon known as “progression to androgen independence.”

Understanding how androgen-sensitive tumors become androgen independent is crucial for the development of novel prostate cancer treatments. The progression of tumors to androgen independence is, however, a complex process thought to involve several factors such as clonal selection, upregulation of anti-apoptotic survival genes, ligand-independent activation of the androgen receptor, and alternative growth factor pathways. Feldman and Feldman (2001) describe some pathways for this process.

The relationship between prostate tumor cells and their dependence on androgens has been investigated using animal tumor models. One example is the Shionogi tumor model (Bruchovsky and others, 1990). This androgen-dependent mammary tumor reliably regresses following castration and predictably recurs in a progressed state within approximately a month. Because patterns of tumor progression after castration mimic those seen in humans, this model is considered to be representative of human disease.

The data we utilize for our analysis are from 6- to 8-week-old mice implanted with Shionogi xenografts and castrated at day 14 postimplantation. Shionogi tumors were isolated at different time points: pre-castration (day 0) and from day 1 to 25 post-castration with mRNA obtained for microarray analysis. Details of microarray hybridizations, differential expression analysis, clustering and bioinformatic analysis (extraction of gene sets with overrepresented functional categories and visualization and analysis of interaction networks) are given in Appendix B of the supplementary material available at Biostatistics online (http://www.biostatistics.oxfordjournals.org).

Finally, we note that in this data set observations are not equally spaced in time. A small modification of our algorithms allows us to model such data in a straightforward manner (see Appendix A, in the supplementary material [http://www.biostatistics.oxfordjournals.org], for some details).

Fig. 2. Simulation study 2. In both panels, the different columns show results using data simulated with variances equal to 0.1, 0.2, and 0.4, respectively. The first row of panels (a) and (b) shows 10 randomly selected trajectories per cluster. The mid row of panels (a) and (b) shows the estimated profiles along with the 95% CIs while the last row of panels (a) and (b) shows the 95% CIs for the network parameters.
Results.

Choice of priors. In this application, we assume priors similar to those described in Section 4.1. As seen in our simulation study, posterior estimation is sensitive to priors set on the network parameters. To determine whether our results under these priors were calibrated to detect some known biological relationships, we set aside a small set of seven genes classified into two clusters.

Time-course expression data for these seven genes are shown in panel (a) of Figure 3. Panel (b) of Figure 3 shows the known biological network of the genes in Ingenuity (http://www.ingenuity.com) with the edges showing the different interactions among them. These genes are part of the gene set “bcl2family_and_reg_network” in the Molecular Signature Database (http://www.broad.mit.edu/gsea/msigdb/index.html) and were selected because they showed a significant enrichment score in our data and also because they were represented in two different clusters, labeled cluster A (B-cell CLL/lymphoma 2 [BCL2], BCL2-associated X protein [BAX], myeloid cell leukemia sequence 1 [MCL1], BCL2-interacting killer [BIK], nuclear cap binding protein subunit 2, 20kDa [NCBP2]) and cluster B (actin beta [ACTB], v-akt murine thymoma viral oncogene homolog 1 [AKT1]). In the previous description, the gene symbols in parentheses are Hugo approved (http://www.gene.ucl.ac.uk/nomenclature/index.html). Ingenuity pathways show that these genes influence each other’s expression levels in different systems and are important in processes related to cell survival and apoptosis.

We analyzed this set of genes under two situations. First, we used the known clusters. This requires a slight modification of our model with inferences made only on the network parameters and not on cluster membership. Second, we assumed that the cluster membership was not known. In this case, we used our full model for joint inferences on the network and clusters. Using priors in which the network parameters were centered at zero with a variance of 0.1, our estimated model recovered the known biological relationships between the two known sets of clusters. Specifically, we detected relationships $A \rightarrow A$ (posterior median $= 0.38$, 90% posterior CI $= [0.11, 0.66]$) and $B \rightarrow A$ (posterior median $= -0.30$, 90% CI $= [-0.59, -0.03]$). When using the model also for clustering, we also found the same relationships ($A \rightarrow A$ with posterior median $= 0.38$, 90% CI $= [0.10, 0.67]$ and $B \rightarrow A$ with posterior...
Cluster-based network model

median= $-0.27$, 90% CI = $[-0.58, -0.01]$). Moreover, the cluster classifications of the genes were the same as in the case where the clusters were known. Given this initial validation of our prior distributions, we moved into analyzing a larger set of genes.

Analysis and results. For this application, we selected a subset of 340 genes which included transcription factors with their potential targets and/or regulators, and were differentially expressed over time, based on the results of our preliminary analysis. In this section, we present the results of our analysis using four clusters automatically obtained from Hopach, and also when using our model to classify genes where, for comparison of results, we use $K = 4$.

Panels (a) and (b) of Figure 4 shows the time-course expression data for some randomly selected genes per cluster classification (specifically, data for the same randomly selected set of genes are shown using the cluster classification from Hopach and also using our joint model). Table 1 of the supplementary material available at Biostatistics online (http://www.biostatistics.oxfordjournals.org) shows the number of genes per cluster in the cross-classification obtained from Hopach and our joint model. We see, for example, that all genes from our cluster 2 are in Hopach’s cluster 3 and that most genes from our cluster 1 are in Hopach’s cluster 1. As seen in the figure and also in the table, there are differences in the classification. In particular, clusters obtained with our model seem more homogeneous than those obtained with Hopach. Although both approaches utilize similarity of the expression profiles to cluster genes, our classification criterion also utilizes the relationships across clusters to classify genes.

Our next comparison is on the network inferences. Hopach does not provide inferences on networks. So, we discuss the results of our analysis under two situations. First, we assume that cluster memberships are known, as provided by Hopach, and use a restricted version of our model that only looks into network inferences. Second, we assume unknown gene-cluster membership and use our full model for joint inferences in the cluster membership and networks.

Figure 5 shows image plots of the posterior probabilities of network positivity (that is, $P(\beta_{ij} > 0|\text{Data})$) along with the 90% posterior CIs for the network parameters. Using the image plots, we look for black (or white) spots signifying highest (or smallest) posterior probabilities for the network positivity. This plot is useful when analyzing data sets with larger number of clusters. In the figure, the image plots show that, under prior 1, there are potentially four non-null relationships between clusters (namely, using Hopach’s clusters the relationships are $1 \rightarrow 1$, $3 \rightarrow 1$, $1 \rightarrow 3$, and $3 \rightarrow 3$ [borderline], while when using our model-based clusters the relationships are $1 \rightarrow 1$, $3 \rightarrow 1$, $1 \rightarrow 3$, and $1 \rightarrow 2$). These relationships are confirmed using the 90% posterior CIs for the network parameters (obtained under prior 1).

As seen in our simulation study, inferences are sensitive to prior choices. So, we also analyzed the data when the prior variances for parameters $\theta$ and $\beta$ are five-fold higher (results shown under prior 2). Under this prior, we only identify relationship $1 \rightarrow 1$.

It is important to note that despite the fact that our joint model allows for greater uncertainty as we estimate not only the network parameters but also gene-cluster membership, we were able to identify the same number of relationships. Moreover, the relationships obtained under the two different clustering methods are related, but not necessarily the same as they may involve slightly different sets of genes. For example, under prior 1, we found the relationship $1 \rightarrow 1$. As seen in Table 1 of the supplementary material available at Biostatistics online (http://www.biostatistics.oxfordjournals.org) genes in Hopach’s cluster 1 are split, when using the joint model, into clusters 1, 3, and 4—with the majority of them being classified into cluster 1. In other words, although cluster 1 under both algorithms share some genes, they are not the same.

Panels (c) and (d) of Figure 4 shows the estimated cluster-specific posterior profiles under prior 1 (profiles under prior 2 are similar and thus omitted). The effect of observations being at unequal time points is reflected by regions where the credible bands are more spread out reflecting the greater uncertainty at those particular time points.
Fig. 4. Case study: subset for analysis. Panels (a) and (b): Time-course expression data for a subset of 340 used for analysis. Here we show some randomly selected trajectories per cluster to facilitate visualization. Panels (c) and (d): Cluster-specific posterior profiles. Posterior medians are shown in full lines along with the 90% posterior CIs in dotted lines. The wider intervals at some time points reflect the higher uncertainty due to missing data.
So far we have analyzed the data assuming the knowledge of the number of clusters $K$, in particular, to draw comparisons of our results with those obtained by Hopach. A natural question that arises is how to choose $K$. In answering this question, as suggested by one of the reviewers, we used the BIC (Bayesian Information Criterion) which is calculated as

$$\text{BIC} = -2 \log(\text{maximized likelihood}) + \log(\text{Number of observations}) \times (\text{Number of parameters}).$$

The optimal number of clusters is the value $K^*$ with minimal BIC. Alternative criteria could also be used. In this application, we also utilize a more informal criterion where we take the minimum value of $K$, denoted $K_m$, that maximizes the number of non-null relationships. Table 2 of the supplementary material available at Biostatistics online (http://www.biostatistics.oxfordjournals.org) presents the results of our analysis for various choices of $K$. Using both criteria, the optimal choice is $K^* = K_m = 4$. 

Fig. 5. Case study. Posterior inferences on the network parameters when $K = 4$ using known clusters as obtained from Hopach (left panels) and when using our model to clustering and network inferences (right panels) under two priors. Under prior 2 the variances for $\beta$ and $\theta$ are five-fold higher than under prior 1. Top panels show the image plots of the posterior probabilities of network positivity (that is, $P(\beta_{ij} > 0|\text{Data})$). The darker the color, the higher the posterior probability. Bottom panels show the 90% posterior CIs. The vertical lines show the limits of the intervals and the circles denote the posterior medians.
Biological networks. To interpret our results with respect to the available biological knowledge, we performed a Gene Ontology Analysis using combinations of clusters that showed significant network relationships (that is, clusters 1 and 3 from our joint model using $K = 4$). The ontology analysis allows us to find which biological themes are enriched as compared to the background set of all genes in the microarray study. Considering the set of genes that showed significant enrichment in the ontology categories of interest (anti-apoptosis, angiogenesis, cytokine activity, lipid binding, chromatin binding, and others), we looked at gene network relationships that are supported by the literature and by our model. Specifically, we consider the special case of the model where each gene is its own cluster and no inference is required on the cluster membership.

The panel labeled “Data Network” in Figure 6 was obtained with Cytoscape and shows the network with genes that showed, according to our model, significant relationships (using posterior probability level of 80%). The network contains 26 genes/nodes and 36 edges/relationships. Over the same set of 26 genes, we created a “Literature Network” (see Figure 6) of co-cited and/or interacting genes using different tools (Ingenuity Pathways, PathwayStudio, Agilent Literature Search/Cytoscape plug-in, and CoPub Mapper). Using Merge Networks/Cytoscape plug-in we obtained the “Data/Literature Intersection
Fig. 6. Case study: Nodes/genes from clusters 1 and 3 are represented by squares and circles, respectively. In the data and in the intersection networks, positive or negative relationships are represented by continuous or dashed edges, respectively.

Network” representing the set of gene-to-gene interactions obtained from our analysis and that, according to the literature, are also involved in different biological systems.

As we look at Figure 6, some relationships are supported by both our model and the literature, and some are not. A few points are worth mentioning in this regard. On one hand, the identification of new relationships that are not mentioned in the current literature has its own merits in generating new biological hypotheses to be further tested in the laboratory. On the other hand, we should not expect relationships found in the literature to be necessarily supported by our model. One of the reasons is that some of the relationships supported by the literature are related to different biological systems. Moreover, the variation in the expression data can also limit our ability to detect true biological relationships.

Some interactions are worth discussing for their biological relevance. Vascular endothelial growth factor (VEGF) and BCL2 have been shown to act either synergistically or antagonistically in different systems. For example, in endothelial cells, VEGF acts as a survival and angiogenic factor by upregulating BCL2 expression (Nör and others, 2001). In prostate carcinoma cell lines that overexpress BCL2, there is an induction of VEGF under hypoxic conditions (Fernandez and others, 2001). In non-small-cell lung cancer, however, BCL2 plays a suppressive role on VEGF expression (Koukourakis and others, 1999) and there is an inverse relationship between BCL2 and VEGF immunoreactivity (Fontanini and others, 1998). In our Shionogi model system, we see a negative relationship between BCL2 (cluster 1) and VEGF (cluster 3), resembling the situation in non-small-cell lung cancer.
There is also support in the literature for the direct relationship between insulin-like growth factor binding protein 5 (IGFBP5) and secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T-lymphocyte activation 1) (SPP1) (Nam and others, 2000). SPP1 is involved in the progression of prostate cancer (Khodavirdi and others, 2006), and IGFBP5 is also an important factor in androgen-independent progression as well as in the Shionogi model (Gleave and Miyake, 2000). In our analysis, a positive relationship between IGFBP5 and SPP1, consistent with the \( 1 \rightarrow 1 \) association, is identified. The edge is not shown in the network due to the lower level of 70% posterior probability, but this is still suggestive of an association between these two factors.

5. Discussion

Clinical heterogeneity and progression of many diseases, such as cancer, reflect the activity of different genes that command biological processes at the cellular level. The ultimate goal in understanding gene regulation lies in the development of therapeutic interventions for treating diseases. The idea is somewhat simple in that if we can characterize gene regulation, then one could potentially design intervention strategies to modify the behavior of genes triggering the disease. The characterization of gene regulation through the investigation of the connectivity in the network is, however, a very complex problem and the usual approaches restrict the study to small set of genes.

Motivated by early experimental work suggesting that biological networks are modular and that time-delayed co-regulation is realistic, we proposed to study networks over sets of genes sharing the same profile of expression. Specifically, we proposed a model-based approach to simultaneously cluster time-course gene expression data and infer cluster networks. The key aspect of our work is that we target relationships between sets of genes (clusters) rather than gene-to-gene relationships. We note, however, that in the special case where each gene is in its own cluster, we can also look at gene-to-gene relationships. The basic premise of our model is that if genes share the same expression profile, then the relationships are captured by the cluster-specific profile. Our approach allows us to capture within- and between-cluster relationships. From a computational viewpoint, by looking at relationships defined at the cluster level, we reduce the number of relationships to be estimated in the network, making the problem of inferring networks more tractable when considering large sets of genes. Moreover, if one is interested in studying gene-to-gene relationships, our methods can still be utilized in a preliminary analysis to narrow down the number of genes fed into a gene-to-gene network analysis.

Our model can be extended in some directions. First, the sampling model, here assumed normal, could alternatively have a different distribution. Second, instead of using the diagonal representation with \( W = \text{diag}(W_1, \ldots, W_K) \), one could allow for non-null correlations between state parameters (a similar comment goes to \( V_I \)). Other possible extensions include allowing \( F \) and \( G \) to be time-dependent, or identifying networks at different time spacings (for example, second-order relationships between cluster profiles at time \( t \) and \( t + 2 \) and so on). However, we note that the ability to precisely estimate such models is affected by the number of measurements over time and expensive computing.

Acknowledgments

This work was partially supported by grants P50 CA 97186, 5 U01 CA 88160, and R01 CA 100778 from the National Cancer Institute. Lurdes Y. T. Inoue also acknowledges partial support from the Career Development Funding from the Department of Biostatistics, University of Washington. Lurdes Y. T. Inoue devised the statistical model, implemented the model and performed the statistical analysis on the networks. Mauricio Neira performed microarray experiments and the preliminary statistical analysis of the gene expression data. He also devised and wrote code for combining the results from the network analysis.
Cluster-based network model

with the bioinformatics analysis. Lurdes Y. T. Inoue, Mauricio Neira and Ruth Etzioni contributed to the writing of the article. Colleen Nelson and Martin Gleave directed the biological experiments of the Shionogi model. Conflict of Interest: None declared.

REFERENCES


*Received December 19, 2005; first revision August 4, 2006; second revision September 7, 2006; accepted for publication September 7, 2006*