Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 51947, 12 pages
doi:10.1155/2007/51947

Research Article

Inferring Time-Varying Network Topologies from Gene Expression Data

Arvind Rao,[1, 2] Alfred O. Hero III,[1, 2] David J. States,[2, 3] and James Douglas Engel[4]

[1] *Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122, USA*
[2] *Bioinformatics Graduate Program, Center for Computational Medicine and Biology, School of Medicine, University of Michigan, Ann Arbor, MI 48109-2218, USA*
[3] *Department of Human Genetics, School of Medicine, University of Michigan, Ann Arbor, MI 48109-0618, USA*
[4] *Department of Cell and Developmental Biology, School of Medicine, University of Michigan, Ann Arbor, MI 48109-2200, USA*

Received 24 June 2006; Revised 4 December 2006; Accepted 17 February 2007

Recommended by Edward R. Dougherty

Most current methods for gene regulatory network identification lead to the inference of steady-state networks, that is, networks prevalent over all times, a hypothesis which has been challenged. There has been a need to infer and represent networks in a dynamic, that is, time-varying fashion, in order to account for different cellular states affecting the interactions amongst genes. In this work, we present an approach, *regime-SSM*, to understand gene regulatory networks within such a dynamic setting. The approach uses a clustering method based on these underlying dynamics, followed by system identification using a state-space model for each learnt cluster—to infer a network adjacency matrix. We finally indicate our results on the mouse embryonic kidney dataset as well as the T-cell activation-based expression dataset and demonstrate conformity with reported experimental evidence.

1. INTRODUCTION

Most methods of graph inference work very well on stationary time-series data, in that the generating structure for the time series does not exhibit switching. In [1, 2], some useful method to learn network topologies using linear *state-space models (SSM)*, from T-cell gene expression data, has been presented. However, it is known that regulatory pathways do not persist over all time. An important recent finding in which the above is seen to be true is following examination of regulatory networks during the yeast cell cycle [3], wherein topologies change depending on underlying (endogeneous or exogeneous) cell condition. This brings out a need to identify the variation of the "hidden states" regulating gene network topologies and incorporating them into their network inference framework [4]. This hidden state at time t (denoted by x_t) might be related to the level of some key metabolite(s) governing the activity (g_t) of the gene(s). These present a notion of condition specificity which influence the dynamics of various genes active during that regime (condition). From time-series microarray data, we aim to partition each gene's expression profile into such regimes of expression, during which the underlying dynamics of the gene's controlling state (x_t) can be assumed to be stationary. In [5], the powerful notion of context sensitive boolean networks for gene relationships has been presented. However, at least for short time-series data, such a boolean characterization of gene state requires a one-bit quantization of the continuous state, which is difficult without expert biological knowledge of the activation threshold and knowledge of the precise evolution of gene expression. Here, we work with gene profiles as continuous variables conditioned on the regime of expression. Each regime is related to the state of a state-space model that is estimated from the data.

Our method (*regime-SSM*) examines three components: to find the switch in gene dynamics, we use a change-point detection (CPD) approach using singular spectrum analysis (SSA). Following the hypothesis that the mechanism causing the genes to switch at the same time came from a common underlying input [3, 6], we group genes having similar change points. This clustering borrows from a mixture of Gaussian (MoG) model [7]. The inference of the network adjacency matrix follows from a state-space representation of expression dynamics among these coclustered genes [1, 2]. Finally, we present analyses on the publicly available embryonic kidney gene expression dataset [8] and the T-cell activation dataset [1], using a combination of the above developed methods and we validate our findings with previously published literature as well as experimental data.

For the embryonic kidney dataset, the biological problem motivating our network inference approach is one of identifying gene interactions during mammalian nephrogenesis (kidney formation). Nephrogenesis, like several other developmental processes, involves the precise temporal interaction of several growth factors, differentiation signals, and transcription factors for the generation and maturation of progenitor cells. One such key set of transcription factors is the GATA family, comprising six members, all containing the (–GATA–) binding domain. Among these, *Gata2* and *Gata3* have been shown to play a functional role [8, 9] in nephric development between days 10–12 after fertilization. From a set of differentially expressed genes pertinent to this time window (identified from microarray data), our goal is to prospectively discover regulatory interactions between them and the *Gata2/3* genes. These interactions can then be further resolved into transcriptional, or signaling interactions on the basis of additional biological information.

In the T-cell activation dataset, the question is if events downstream of T-cell activation can be partitioned into early and late response behaviors, and if so, which genes are active in a particular phase. Finally, can a network-level influence be inferred among the genes of each phase and do they correlate with known data? We note here that we are not looking for the behavior of any particular gene, but only interested in genes from each phase.

As will be shown in this paper, *regime-SSM* generates biologically relevant hypotheses regarding time-varying gene interactions during nephric development and T-cell activation. Several interesting transcripts are seen to be involved in the process and the influence network hereby generated resolves cyclic dependencies.

The main assumption for the formulation of a linear state-space model to examine the possibility of gene-gene interactions is that gene expression is a function of the underlying cell state and the expression of other genes at the previous time step. If longer-range dependencies are to be considered, the complexity of the model would increase. Another criticism of the model might be that nonlinear interactions cannot be adequately modeled by such a framework. However, around the equilibrium point (steady state), we can recover a locally linearized version of this nonlinear behavior.

2. SSA AND CHANGE-POINT DETECTION

First we introduce some notations. Consider N gene expression profiles, $g^{(1)}, g^{(2)}, \dots, g^{(N)} \in \mathbb{R}^{\mathbb{T}}$, T being the length of each gene's temporal expression profile (as obtained from microarray expression). The jth time instant of gene i's expression profile will be denoted by $g_j^{(i)}$.

State-space partitioning is done using *singular spectrum analysis* [10] (SSA). SSA identifies structural change points in time-series data using a sequential procedure [11]. We will briefly review this method.

Consider the "windowed" (width N_W) time-series data given by $\{g_1^{(i)}, g_2^{(i)}, \dots, g_{N_W}^{(i)}\}$, with M ($M \leq N_W/2$) as some integer-valued lag parameter, and a replication parameter $K = N_W - M + 1$. The SSA procedure in CPD involves the following.

(i) Construction of an l-dimensional subspace: here, a "trajectory matrix" for the time series, over the interval $[n+1, n+T]$ is constructed,

$$\mathbf{G}_B^{i,(n)} = \begin{pmatrix} g_{n+1}^{(i)} & g_{n+2}^{(i)} & g_{n+3}^{(i)} & \cdots & g_{n+K}^{(i)} \\ g_{n+2}^{(i)} & g_{n+3}^{(i)} & g_{n+4}^{(i)} & \cdots & g_{n+K+1}^{(i)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_{n+M}^{(i)} & g_{n+M+1}^{(i)} & g_{n+M+2}^{(i)} & \cdots & g_{n+N_W}^{(i)} \end{pmatrix}, \quad (1)$$

where $K = N_W - M + 1$. The columns of the matrix $\mathbf{G}_B^{i,(n)}$ are the vectors $G_j^{i,(n)} = (g_{n+j}^{(i)}, \dots, g_{n+j+M-1}^{(i)})^T$, with $j = 1, \dots, K$.

(ii) Singular vector decomposition of the lag covariance matrix $\mathbf{R}^{i,n} = \mathbf{G}_B^{i,(n)}(\mathbf{G}_B^{i,(n)})^T$ yields a collection of singular vectors—a grouping of l of these Singular vectors, corresponding to the l highest eigenvalues—denoted by $I = \{1, \dots, l\}$, establishes a subspace $\mathcal{L}_{n,I}$ of $\mathbb{R}^M$.

(iii) Construction of the *test matrix*: use $\mathbf{G}_{\text{test}}^{i,(n)}$ defined by

$$\mathbf{G}_{\text{test}}^{i,(n)} = \begin{pmatrix} g_{n+p+1}^{(i)} & g_{n+p+2}^{(i)} & \cdots & g_{n+q}^{(i)} \\ g_{n+p+2}^{(i)} & g_{n+p+3}^{(i)} & \cdots & g_{n+q+1}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n+p+M}^{(i)} & g_{n+p+M+1}^{(i)} & \cdots & g_{n+q+M-1}^{(i)} \end{pmatrix}. \quad (2)$$

Here, we use the length (p) and location (q) of test sample. We choose $p \geq K$, with $K = N_W - M + 1$. Also $q > p$, here we take $q = p + 1$. From this construction, the matrix columns are the vectors $G_j^{i,(n)}$, $j = p+1, \dots, q$. The matrix has dimension $M \times Q$, $Q = (q - p) = 1$.

(iv) Computation of the detection statistic: the detection statistics used in the CPD are

(a) the normed Euclidean distance between the column span of the *test matrix*, that is, $G_j^{i,(n)}$ and the l-dimensional subspace $\mathcal{L}_{n,I}$ of $\mathbb{R}^M$. This is denoted by $\mathcal{D}_{n,I,p,q}$;

(b) the normalized sum of squares of distances, denoted by $S_n = \mathcal{D}_{n,I,p,q}/MQ\mu_{n,I}$, with $\mu_{n,I} = \mathcal{D}_{m,I,0,K}$, where m is the largest value of $m \leq n$ so that the hypothesis of no change is accepted;

(c) a cumulative sum- (CUSUM-) type statistic $W_1 = S_1$, $W_{n+1} = \max\{(W_n + S_{n+1} - S_n - 1/3MQ), 0\}$, $n \geq 1$.

The CPD procedure declares a structural change in the time series dynamics if for some time instant n, we observe $W_n > h$ with the threshold $h = (2t_\alpha/(MQ))\sqrt{(1/3)q(3MQ - Q^2 + 1)}$, t_α being the $(1-\alpha)$ quantile of the standard normal distribution.

(v) Choice of algorithm parameters:

(a) window width (N_W): here, we choose $N_W \simeq T/5$, T being the length of the original time series, the algorithm provides a reliable method of extracting most structural changes. As opposed to choosing a much smaller N_W, this might lead to some outliers being classified as potential change points, but in our set-up this is preferred in contrast to losing genuine structural changes based on choosing larger N_W;

(b) choice of lag M: in most cases, choose $M = N_W/2$.

3. MIXTURE-OF-GAUSSIANS (MoG) CLUSTERING

Having found change points (and thus, regimes) from the gene trajectories of the differentially expressed genes, our goal is to now group (cluster) genes with similar temporal profiles within each regime. In this section, we derive the parameter update equations for a *mixture-of-Gaussian* clustering paradigm. As will be seen later, the Gaussian assumptions on the gene expression permit the use of coclustered genes for the SSM-based network parameter estimation.

We now consider the group of gene expression profiles $\mathcal{G} = \{\mathbf{g}^{(\mathbf{1})}, \mathbf{g}^{(\mathbf{2})}, \dots, \mathbf{g}^{(\mathbf{n})}\}$, all of which share a common change point (time of switch)—c_1. Consider gene profile i, $\mathbf{g}^{(\mathbf{i})} = [g_1^{(i)}, g_2^{(i)}, \dots, g_{T_{c_1}}^{(i)}]^T$, a T_{c_1}-dimensional random vector which follows a k-component finite mixture distribution described by

$$p(\mathbf{g} \mid \boldsymbol{\theta}) = \sum_{m=1}^{k} \alpha_m p(\mathbf{g} \mid \boldsymbol{\phi}_{\boldsymbol{m}}), \tag{3}$$

where $\alpha_1, \dots, \alpha_k$ are the mixing probabilities, each ϕ_m is the set of parameters defining the mth component, and $\boldsymbol{\theta} \equiv \{\phi_1, \dots, \phi_k, \alpha_1, \dots, \alpha_k\}$ is the set of complete parameters needed to specify the mixture. We have

$$\alpha_m \geq 0, \quad m = 1, \dots, k, \quad \sum_{m=1}^{k} \alpha_m = 1. \tag{4}$$

For a set of n independently and identically distributed samples,

$$\mathcal{G} = \{\mathbf{g}^{(\mathbf{1})}, \mathbf{g}^{(\mathbf{2})}, \dots, \mathbf{g}^{(\mathbf{n})}\}, \tag{5}$$

the log-likelihood of a k-component mixture is given by

$$\begin{aligned} \log p(\mathcal{G} \mid \boldsymbol{\theta}) &= \log \prod_{i=1}^{n} p(\mathbf{g}^{(\mathbf{i})} \mid \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log \sum_{m=1}^{k} \alpha_m p(\mathbf{g}^{(\mathbf{i})} \mid \phi_m). \end{aligned} \tag{6}$$

(i) Treat the labels, $\mathcal{Z} = \{\mathbf{z}^{(\mathbf{1})}, \dots, \mathbf{z}^{(\mathbf{n})}\}$, associated with the n samples—as missing data. Each label is a binary vector $\mathbf{z}^{(\mathbf{i})} = [z_1^{(i)}, \dots, z_k^{(i)}]$, where $z_m^{(i)} = 1$ and $z_p^{(i)} = 0$, for $p \neq m$ indicate that sample $\mathbf{g}^{(\mathbf{i})}$ was produced by the mth component.

In this setting, the *expectation maximization* algorithm can be used to derive the cluster parameter $(\boldsymbol{\theta})$ update equations.

In the *E-step* of the *EM algorithm*, the function $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(t)) \equiv E[\log p(\mathcal{G}, \mathcal{Z} \mid \boldsymbol{\theta}) \mid \mathcal{G}, \hat{\boldsymbol{\theta}}(t)]$ is computed. This yields

$$w_m^{(i)} \equiv E[z_m^{(i)} \mid \mathcal{G}, \hat{\theta}_t] = \frac{\hat{\alpha}_m(t) p(\mathbf{g}^{(\mathbf{i})} \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}(t))}{\sum_{j=1}^{k} \hat{\alpha}_j(t) p(\mathbf{g}^{(\mathbf{i})} \mid \hat{\boldsymbol{\theta}}_{\boldsymbol{j}}(t))}, \tag{7}$$

where $w_m^{(i)}$ is the posterior probability of the event $z_m^{(i)} = 1$, on observing $g_m^{(i)}$.

The estimate of the number of components (k) is chosen using a minimum message length (MML) criterion [7]. The MML criterion borrows from algorithmic information theory and serves to select models of lowest complexity to explain the data. As can be seen below, this complexity has two components: the first encodes the observed data as a function of the model and the second encodes the model itself. Hence, the MML criterion in our setup becomes,

$$\hat{k}_{\text{MML}} = \arg\min_k \left\{ -\log p(\mathcal{G} \mid \hat{\boldsymbol{\theta}}(k)) + \frac{k(N_p + 1)}{2} \log n \right\}, \tag{8}$$

N_p is number of parameters per component in the k component mixture, given the number of clusters $k_{\min} \leq k \leq k_{\max}$. In the *M-step*, for $m = 0, 1, \dots, k$, $\hat{\theta}_m(t+1) = \arg\max_{\phi_m} Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(t))$, for $m : \hat{\alpha}_m(t+1) > 0$, the elements $\hat{\phi}$'s of the parameter vector estimate $\hat{\boldsymbol{\theta}}$ are typically not closed form and depend on the specific parametrization of the densities in the mixture, that is, $p(\mathbf{g}^{(\mathbf{i})} \mid \boldsymbol{\phi}_{\boldsymbol{m}})$. If $p(\mathbf{g}^{(i)} \mid \phi_m)$ belongs to the Gaussian density $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ class, we have, $\phi = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and EM updates yield [7]

$$\begin{aligned} \hat{\alpha}_m(t+1) &= \frac{\sum_{i=1}^{n} w_m^{(i)}}{n}, \\ \boldsymbol{\mu}_m(t+1) &= \frac{\sum_{i=1}^{n} w_m^{(i)} \mathbf{g}^{(i)}}{\sum_{i=1}^{n} w_m^{(i)}}, \\ \boldsymbol{\Sigma}_m(t+1) &= \frac{\sum_{i=1}^{n} w_m^{(i)} (\mathbf{g}^{(i)} - \mu_m(t+1))(\mathbf{g}^{(i)} - \mu_m(t+1))^T}{\sum_{i=1}^{n} w_m^{(i)}}. \end{aligned} \tag{9}$$

Equations (7) and (9) are the parameter update equations for each of the $m = 1, \dots, k$ cluster components.

For the kidney expression data, since we are interested in the role of *Gata2* and *Gata3* during early kidney development, we consider all the genes which have similar change points as the *Gata2* and *Gata3* genes, respectively. We perform an MoG clustering within such genes and look at those coclustered with *Gata2* or *Gata3*. Coclustering within a regime potentially suggests that the governing dynamics are the same, even to the extent of coregulation. We note that just because a gene is coclustered with *Gata2* in one regime, it does not mean that it will cocluster in a different regime. This approach suggests a way to localize regimes of correlation instead of the traditional global correlation measure that can mask transient and condition-specific dynamics. For this gene expression data, the MML penalized criterion indicates that an adequate number of clusters to describe this data is two ($k = 2$). In Tables 1 and 2, we indicate some of the genes with similar coexpression dynamics as *Gata2/Gata3* and a cluster assignment of such genes. We observe that this clustering corresponds to the first phase of embryonic development (days 10–12 dpc), the phase where *Gata2* and *Gata3* are perhaps most relevant to kidney development [12–15].

A word about Table 1 is in order. The entries in each column of a row (gene) indicate the change points (as found by the SSA-CPD procedure) in the time series of the interpolated gene expression profile. Our simulation studies with the T-cell data indicate that the SSM and CoD performance is not much worse with the interpolated data compared to the original time series (Table 7). We note that because of the present choice of parameters N_W, we might have the detection of some false positive change points, but this is preferable to the loss of genuine change points. An examination of the change points of the various genes in Table 1 indicates three regimes—between points approximately 1–5, 5–11 and 12–20. The missing entries mean that there was no change point identified for a certain regime and are thus treated as such. Since our focus is early *Gata3* behavior, we are interested in time points 1–12, and hence we examine the evolution of network-level interactions over the first two regimes for the genes coclustered in these regimes.

To clarify the validity of the presented approach, we present a similar analysis on another data set—the T-cell expression data presented in [1]. This data looks at the expression of various genes after T-cell activation using stimulation with phorbolester PMA and ionomycin [16]. This data has the profiles of about 58 genes over 10 time points with $44(34+10)$ replicate measurements for each time point. Since here we have no specific gene in mind (unlike earlier where we were particularly interested in *Gata3* behavior), the change point procedure (CPD) yields two distinct regimes—one from time points 1 to 4 and the other from time points 5 to 10. Following the MoG clustering procedure yields the optimal number of clusters to be 1 (from MML) in each regime. We therefore call these two clusters "early response" and "late response" genes and then proceed to learn a network relationship amongst them, within each cluster. The CPD and cluster information for the early and late responses are summarized in Table 3.

4. STATE-SPACE MODEL

For a given regime, we treat gene expression as an observation related to an underlying hidden cell state (x_t), which is assumed to govern regime-specific gene expression dynamics for that biological process, globally within the cell. Suppose there are N genes whose expression is related to a single process. The ith gene's expression vector is denoted as $g_t^{(i)}$, $t = 1,\dots T$, where T is the number of time points for which the data is available. The *state-space model (SSM)* is used to model the gene expression ($g_t^{(i)}$, $i = 1, 2, \dots, N$ and $t = 1, 2, \dots, T$) as a function of this underlying cell state (x_t) as well as some external inputs. A notion of influence among genes can be integrated into this model by considering the SSM inputs to be the gene expression values at the previous time step. The state and observation equations of the *state-space model* [17] are

(i) state equation:

$$\mathbf{x_{t+1}} = A\mathbf{x_t} + B\mathbf{g_t} + \mathbf{e_{s,t}}; \quad \mathbf{e_{s,t}} \sim \mathcal{N}(0, Q), \\ i = 1, \dots, N; \ t = 1, \dots, T; \tag{10}$$

(ii) observation equation:

$$\mathbf{g_t} = C\mathbf{x_t} + D\mathbf{g_{t-1}} + \mathbf{e_{o,t}}; \quad \mathbf{e_{o,t}} \sim \mathcal{N}(0, R), \tag{11}$$

Table 1: Change-point analysis of some key genes, prior to clustering (annotations in Table 8). The numbers indicate the time points at which regime changes occur for each gene.

Gene symbol	Change point I	Change point II	Change point III
Bmp7	6	10	12
Rara	5	11	16
Pax2	6	12	15
Gata3	5	9	12
Gata2	—	—	18
Gdf11	—	10	20
Npnt	—	12	16
Cd44	5	11	15
Pgf	5	11	—
Pbx1	5	12	20
Ret	—	10	—

Table 2: Some of the genes coclustered with *Gata2* and *Gata3* after MoG clustering (annotations in Table 8).

Genes with the same dynamics as *Gata3*	Genes with the same dynamics as *Gata2*
Bmp7	*Lamc2*
Nrtn	*Cldn3*
Pax2	*Ros1*
Ros1	*Ptprd*
Pbx1	*Npnt*
Rara	*Cdh16*
Gdf11	*Cldn4*

Table 3: Some of the genes related to early and late responses in T-cell activation (annotations in Table 9).

Genes related to early response (time points: 1–4)	Genes related to late response (time points: 5–10)
CD69	*CCNA2*
Mcp1	*CDC2*
Mcl1	*EGR1*
EGR1	*IL2r* gamma
JunD	*IL6*
CKR1	—

Table 4: Assumptions and log-likelihood calculations in the state-space model. The (≡) symbol indicates a definition.

Symbol	Interpretation	Expression
T	Number of time points	—
R_g	Number of replicates	—
$P(\mathbf{g_t} \mid \mathbf{x_t})$	≡	$\prod_{t=2}^{T} \{e^{-1/2[\mathbf{g_t}-C\mathbf{x_t}-D\mathbf{g_{t-1}}]' R^{-1} [\mathbf{g_t}-C\mathbf{x_t}-D\mathbf{g_{t-1}}]}\} \cdot (2\pi)^{-p/2} \det(R)^{-1/2}$
$P(\mathbf{x_t} \mid \mathbf{x_{t-1}})$	—	$\prod_{t=2}^{T} \{e^{-1/2[\mathbf{x_t}-A\mathbf{x_{t-1}}-B\mathbf{g_{t-1}}]' Q^{-1} [\mathbf{x_t}-A\mathbf{x_{t-1}}-B\mathbf{g_{t-1}}]}\} \cdot (2\pi)^{-k/2} \det(Q)^{-1/2}$
$P(\mathbf{x_1})$	Initial state density assumption	$e^{-1/2[\mathbf{x_1}-\boldsymbol{\pi}_1]' V_1 [\mathbf{x_1}-\boldsymbol{\pi}_1]} \cdot (2\pi)^{-k/2} \det\left(V_1\right)^{-1/2}$
$P(\{\mathbf{x}\}, \{\mathbf{g}\})$	Markov property	$\prod_{i=1}^{R_g} \left(P(\mathbf{x_1}^{(i)}) \prod_{t=2}^{T} P(\mathbf{x_t}^{(i)} \mid \mathbf{x_{t-1}}^{(i)}, \mathbf{g_{t-1}}^{(i)}) \cdot \prod_{t=1}^{T} P(\mathbf{g_t}^{(i)} \mid \mathbf{x_t}^{(i)}, \mathbf{g_{t-1}}^{(i)}) \right)$
$\log P(\{\mathbf{x}\}, \{\mathbf{g}\})$	Joint log probability	$\begin{aligned} -\sum_{i=1}^{R_g} \Bigg\{ & \sum_{t=2}^{T} \left(\frac{1}{2} [\mathbf{g_t}^{(i)} - C\mathbf{x_t}^{(i)} - D\mathbf{g_{t-1}}^{(i)}]' R^{-1} [\mathbf{g_t}^{(i)} - C\mathbf{x_t}^{(i)} - D\mathbf{g_{t-1}}^{(i)}] \right) - \left(\frac{T}{2} \right) \log(\det(R)) \\ & - \sum_{t=1}^{T} \left(\frac{1}{2} [\mathbf{x_t}^{(i)} - A\mathbf{x_{t-1}}^{(i)} - B\mathbf{g_{t-1}}^{(i)}]' Q^{-1} [\mathbf{x_t}^{(i)} - A\mathbf{x_{t-1}}^{(i)} - B\mathbf{g_{t-1}}^{(i)}] \right) \\ & - \frac{T-1}{2} \log(\det(Q)) - \frac{1}{2} [\mathbf{x_1} - \boldsymbol{\pi}_1] V_1^{-1} [\mathbf{x_1} - \boldsymbol{\pi}_1] - \frac{1}{2} \log(\det(V_1)) \\ & - \frac{T(p+k)}{2} \log(2\pi) \Bigg\} \end{aligned}$

with $\mathbf{x_t} = [x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(K)}]^T$ and $\mathbf{g_t} = [g_t^{(1)}, g_t^{(2)}, \ldots, g_t^{(N)}]^T$. A likelihood method [1] is used to estimate the state dimension K. The noise vectors $\mathbf{e_{s,t}}$ and $\mathbf{e_{o,t}}$ are Gaussian distributed with mean 0 and covariance matrices Q and R, respectively.

From the state and observation equations (10) and (11), we notice that the matrix-valued parameter $D = [D_{i,j}]_{i=1,\ldots,N}^{j=1,\ldots,N}$ quantifies the influence among genes i and j from one time instant to the next, within a specific regime. To infer a biological network using D, we use bootstrapping to estimate the distribution of the strength of association estimates amongst genes and infer network linkage for those associations that are observed to be significant.

Within this proposed framework, we segment the overall gene expression time trajectories into smaller, approximately stationary, gene expression regimes. We note that the MoG clustering framework is a nonlinear one in that the regime-specific state space is partitioned into clusters. These cluster assignments of correlated gene expression vectors can change with regime, allowing us to capture the sets of genes that interact under changing cell condition.

5. SYSTEM IDENTIFICATION

We consider the case where we have $R_g = B \times P$ realizations of expression data for each gene available. Arguably, mRNA level is a measure of gene expression, $B(= 2)$ denotes the number of biological replicates, and $P(= 16$ perfect match probes) denotes the number of probes per gene transcript. Each of these R_g realizations is T-time-point long and is obtained from Affymetrix U74Av2 murine microarray raw CEL files. In the section below, we derive the update equations for maximum-likelihood estimates of the parameters A, B, C, D, Q and R (in (10) and (11)) using an EM algorithm, based on [17, 18]. The assumptions underlying this model are outlined in Table 4. A sequence of T output vectors $(\mathbf{g_1}, \mathbf{g_2}, \ldots, \mathbf{g_T})$ is denoted by $\{\mathbf{g}\}$, and a subsequence $\{\mathbf{g_{t_0}}, \mathbf{g_{t_0+1}}, \ldots, \mathbf{g_{t_1}}\}$ by $\{\mathbf{g}\}_{t_0}^{t_1}$. We treat the $(\mathbf{x_t}, \mathbf{g_t})$ vector as the complete data and find the log-likelihood $\log P(\{\mathbf{x}\}, \{\mathbf{g}\})$ under the above assumptions. The complete E-and M-steps involved in the parameter update steps are outlined in Tables 5 and 6.

6. BOOTSTRAPPED CONFIDENCE INTERVALS

As suggested above, the entries of the D matrix indicate the strength of influence among the genes, from one time step to the next (within each regime). We use bootstrapping to find confidence intervals for each entry in the D matrix and if it is significant, we assign a positive or negative direction (+1 or −1) to this influence.

The bootstrapping procedure [19] is adapted to our situation as follows.

Table 5: M-step of the EM algorithm for state-space parameter estimation. The (≡) symbol indicates a definition.

Matrix symbol	Interpretation	Expression
M-Step		
$\boldsymbol{\pi}_{\mathbf{1}}{}^{\text{new}}$	Initial state mean	$\widehat{\mathbf{x_1}}$
V_1^{new}	Initial state covariance	$P_1 - \widehat{\mathbf{x_1}}\widehat{\mathbf{x_1}}' + \dfrac{1}{R_g}\sum_{i=1}^{R_g}[\widehat{\mathbf{x_1}}'^{(i)} - \overline{\widehat{\mathbf{x_1}}}][\widehat{\mathbf{x_1}}'^{(i)} - \overline{\widehat{\mathbf{x_1}}}]'$
C^{new}	Output matrix	$\left(\sum_{i=1}^{R_g}\sum_{t=1}^{T}\mathbf{g_t}^{(i)}\hat{\mathbf{x}}_{\mathbf{t}}' - D\sum_{i=1}^{R_g}\sum_{t=1}^{T}\hat{\mathbf{x}}_{\mathbf{t}}{}^{(i)}\mathbf{g_{t-1}}'^{(i)}\right)\cdot\left(\sum_{i=1}^{R_g}\sum_{t=1}^{T}P_t^{(i)}\right)^{-1}$
R^{new}	Output noise covariance	$\dfrac{1}{R_g \times T}\left[\sum_{i=1}^{R_g}\sum_{t=1}^{T}(\mathbf{g_t}^{(i)}\mathbf{g_t}'^{(i)}) - C^{\text{new}}(\hat{\mathbf{x}}_{\mathbf{t}}{}^{(i)}\mathbf{g_t}'^{(i)}) - D^{\text{new}}\mathbf{g_{t-1}}^{(i)}\mathbf{g_t}'^{(i)}\right]$
A^{new}	State dynamics matrix	$\sum_{i=1}^{R_g}\sum_{t=2}^{T}[P_{t,t-1}^{(i)} - B\hat{\mathbf{x}}_{\mathbf{t}}{}^{(i)}\mathbf{g_{t-1}}'^{(i)}]\cdot\left(\sum_{i=1}^{R_g}\sum_{t=2}^{T}P_{t-1}^{(i)}\right)^{-1}$
D^{new}	Input to observation	$\sum_{i=1}^{R_g}\sum_{t=1}^{T}\left[\mathbf{g_t}^{(i)}\mathbf{g_{t-1}}'^{(i)} - \mathbf{g_t}^{(i)}\hat{\mathbf{x}}_{\mathbf{t}}'^{(i)}\left(\sum_{i=1}^{R_g}\sum_{t=1}^{T}P_t^{(i)}\right)^{-1}\hat{\mathbf{x}}_{\mathbf{t}}{}^{(i)}\mathbf{g_{t-1}}'^{(i)}\right]$ $\cdot\left[\sum_{i=1}^{R_g}\sum_{t=1}^{T}\left(\mathbf{g_{t-1}}^{(i)}\mathbf{g_{t-1}}'^{(i)} - \mathbf{g_{t-1}}^{(i)}\hat{\mathbf{x}}_{\mathbf{t}}'^{(i)}\cdot\left(\sum_{i=1}^{R_g}\sum_{t=1}^{T}P_t^{(i)}\right)^{-1}\hat{\mathbf{x}}_{\mathbf{t}}{}^{(i)}\mathbf{g_{t-1}}'^{(i)}\right)\right]^{-1}$
B^{new}	Input to state matrix	$\sum_{i=1}^{R_g}\sum_{t=2}^{T}\left[P_{t,t-1}^{(i)}\left(\sum_{i=1}^{R_g}\sum_{t=2}^{T}P_t^{(i)}\right)^{-1}\hat{\mathbf{x}}_{\mathbf{t}}{}^{(i)}\mathbf{g_{t-1}}'^{(i)} - \hat{\mathbf{x}}_{\mathbf{t}}{}^{(i)}\mathbf{g_{t-1}}'^{(i)}\right]$ $\cdot\left[\sum_{i=1}^{R_g}\sum_{t=2}^{T}\mathbf{g_{t-1}}^{(i)}\mathbf{x_t}'^{(i)}\left(\sum_{i=1}^{R_g}\sum_{t=2}^{T}P_t^{(i)}\right)^{-1}\cdot\mathbf{x_t}^{(i)}\mathbf{g_{t-1}}'^{(i)} - \mathbf{g_{t-1}}\mathbf{g_{t-1}}'^{(i)}\right]^{-1}$
Q^{new}	State noise covariance	$\dfrac{1}{R_g \times (T-1)}\left(\sum_{i=1}^{R_g}\sum_{t=2}^{T}P_t^{(i)} - A^{\text{new}}\sum_{i=1}^{R_g}\sum_{t=2}^{T}P_{t-1,t}^{(i)} - B\sum_{i=1}^{R_g}\sum_{t=2}^{T}\mathbf{g_{t-1}}^{(i)}\hat{\mathbf{x}}_{\mathbf{t}}'^{(i)}\right)$

(i) Suppose there are R regimes in the data with change points $(c_1, c_2, \ldots, c_R)$ identified from SSA. For the rth regime, generate B independent bootstrap samples of size N (the original number of genes under consideration), -$(\mathbf{Y}_1^*, \mathbf{Y}_2^*, \ldots, \mathbf{Y}_B^*)$ from original data, by random resampling from $\mathbf{g}^{(\mathbf{i})} = [g_{c_r}^{(i)}, \ldots, g_{c_{r+1}}^{(i)}]^T$.

(ii) Using the EM algorithm for parameter estimation, estimate the value of D (the influence parameter). Denote the estimate of D for the ith bootstrap sample by D_i^*.

(iii) Compute the sample mean and sample variance of the estimates of D over all the B bootstrap samples. That is,

$$\text{mean} = \overline{D}^* = \frac{1}{B}\sum_{i=1}^{B}(D_i^*),$$
$$\text{variance} = \frac{1}{B-1}\sum_{i=1}^{B}(D_i^* - \overline{D}^*)^2. \qquad (12)$$

(iv) Using the above obtained sample mean and variance, estimate confidence intervals for the elements of D. If D lies in this *bootstrapped* confidence interval, we infer a potential influence and if not, we discard it. Note that even though we write D, we carry out this hypothesis test for each $D_{i,j}$, $i = 1, \ldots, n$; $j = 1, \ldots, n$; for each of the n genes under consideration in every regime.

7. SUMMARY OF ALGORITHM

Within each regime identified by CPD, we model gene expression as Gaussian distributed vectors. We cluster the genes using a *mixture-of-Gaussians (MoG)* clustering algorithm [7] to identify sets of genes which have similar "dynamics of expression"—in that they are correlated within that regime. We then proceed to learn the dynamic system parameters (matrices A, B, C, D, Q, and R) for the state-space model (SSM) underlying each of the clusters. We note two important ideas:

(i) we might obtain different cluster assignments for the genes depending on the regime;

(ii) since all these genes (across clusters within a regime) are still related to the same biological process, the hidden state $\mathbf{x_t}$ is shared among these clusters.

Therefore, we learn the SSM parameters in an alternating manner by updating the estimates from cluster to cluster

Table 6: E-step of the EM algorithm for state-space parameter estimation.

E-Step		
Forward		
$\mathbf{x_1}^0$	$\equiv$	$\boldsymbol{\pi}_\mathbf{1}$
V_1^0	$\equiv$	V_1
$\mathbf{x_t}^{t-1}$	Update	$A\mathbf{x_{t-1}}^{t-1} + B\mathbf{g_{t-1}}$
V_t^{t-1}	Update	$AV_{t-1}^{t-1}A' + Q$
K_t	Update	$V_t^{t-1}C'\left(CV_t^{t-1}C' + R\right)^{-1}$
$\mathbf{x_t}^t$	Update	$\mathbf{x_t}^{t-1} + K_t(\mathbf{g_t} - C\mathbf{x_t}^{t-1} - D\mathbf{g_{t-1}})$
V_t^t	Update	$V_t^{t-1} - K_tCV_t^{t-1}$
Backward		
$V_{T,T-1}^T$	*Initialization*	$(I - K_TC)AV_{T-1}^{T-1}$
$\hat{\mathbf{x}}_\mathbf{t}$	$\equiv$	$\mathbf{x_t}^\tau$
P_t	$\equiv$	$V_t^T + \mathbf{x_t}^T\mathbf{x_t}^{T'}$
J_{t-1}	Update	$V_t^{t-1}A'\left(V_t^{t-1}\right)^{-1}$
$\mathbf{x_{t-1}}^T$	Update	$\mathbf{x_{t-1}}^{t-1} + J_{t-1}(\mathbf{x_1}^T - A\mathbf{x_{t-1}}^{t-1} - Bg_{t-2})$
V_t^T	Update	$V_{t-1}^{t-1} + J_{t-1}(V_t^T - V_t^{t-1})J_{t-1}'$
$P_{t,t-1}$	$\equiv$	$V_{t,t-1}^T + \mathbf{x_t}^T\mathbf{x_{t-1}}^{T'}$
$V_{t-1,t-2}^T$	Update	$V_{t-1}^{t-1}J_{t-2}' + J_{t-1}(V_{t,t-1}^T - AV_{t-1}^{t-1})J_{t-2}'$

while still retaining the form of the state vector $\mathbf{x_t}$. The learning is done using an *expectation-maximization*-type algorithm. The number of components during regime-specific clustering is estimated using a *minimum message length* criterion. Typically, $O(N)$ iterations suffice to infer the mixture model in each regime with N genes under consideration. Thus, our proposed approach is as follows.

(i) Identify the N key genes based on required phenotypical characteristic using fold change studies. Preprocess the gene expression profiles by standardization and cubic spline interpolation.

(ii) Segment each gene's expression profile into a sequence of state-dependent trajectories (regime change points), from underlying dynamics, using SSA.

(iii) For each regime (as identified in step 2),

> cluster genes using an MoG model so that genes with correlated expression trajectories cluster together. Learn an SSM [17, 18] for each cluster (from (10) and (11) for estimation of the mean and covariance matrices of the state vector) within that regime. The input to observation matrix (D) is indicative of the topology of the network in that regime.

(iv) Examine the network matrices D (by bootstrapping to find thresholds on strength of influence estimates) across all regimes to build the time-varying network.

The discussion of the network inference procedure would be incomplete in the absence of any other algorithms for comparison. For this purpose, we implement the CoD- (coefficient-of-determination-) based approach [20, 21] along with the models proposed in [1] (SSM) and [22] (GGM). The CoD method allows us to determine the association between two genes within a regime via an R^2 goodness of fit statistic. The methods of [1, 22] are implemented on the time-series data (with regard to underlying regime). Such a study would be useful to determine the relative merits of each approach. We believe that no one procedure can work for every application and the choice of an appropriate procedure would be governed by the biological question under investigation. Each of these methods use some underlying assumptions and if these are consistent with the question that we ask, then that method has great utility. These individual results, their evaluation, and their comparison are summarized in Section 8.

8. RESULTS

8.1. Application to the GATA pathway

To illustrate our approach (*regime-SSM*), we consider the embryonic kidney gene expression dataset [8] and study the set of genes known to have a possible role in early nephric development. An interruption of any gene in this signaling cascade potentially leads to early embryonic lethality or abnormal organ development. An influence network among these genes would reveal which genes (and their products) become important at a certain phase of nephric development. The choice of the $N(= 47)$ genes is done using FDR fold change studies [23] between ureteric bud and metanephric mesenchyme tissue types, since this spatial tissue expression is of relevance during early embryonic development. The dataset is obtained by daily sampling of the mRNA expression ranging from 11.5–16.5 days post coitus (dpc). Detailed studies of the phenotypes characterizing each of these days is available from the Mouse Genome Informatics Database at http://www.informatics.jax.org/. We follow [24] and use interpolated expression data pre-processing for cluster analysis. We resample this interpolated profile to obtain twenty points per gene expression profile. Two key aspects were confirmed after interpolation [24, 25]: (1) there were no negative expression values introduced, (2) the differences in fold change were not smoothed out.

Initial experimental studies have suggested that the 10.5–12.5 dpc are relatively more important in determination of the course of metanephric development. We chose to explore which genes (out of the 47 considered) might be relevant in this specific time window. The SSA-CPD procedure identified several genes which exhibit similar dynamics (have approximately same change points, for any given regime) in the early phase and distinctly different dynamics in later phases (Table 1).

Our approach to influence determination using the state-space model yields up to three distinct regimes of expression over all the 47 genes identified from fold change studies between bud and mesenchyme. MoG clustering followed by

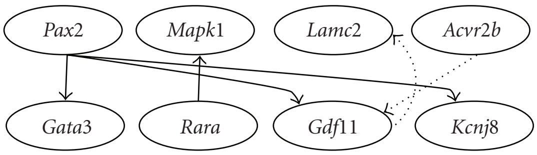

Figure 1: Network topology over regimes (solid lines represent the first regime, and the dotted lines indicate the second regime).

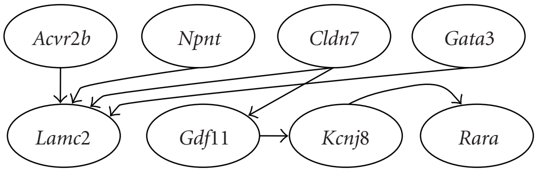

Figure 2: Steady-state network inferred over all time, using [1].

state-space modeling yield three regime topologies of which we are interested in the early regime (days 10.5–12.5). This influence topology is shown in Figure 1.

We compare our obtained network (using *regime-SSM*) with the one obtained using the approach outlined in [1], shown in Figure 2. We note that the network presented in Figure 2 extends over all time, that is, days 10.5–16.5 for which basal influences are represented but transient and condition-specific influences may be missed. Some of these transient influences are recaptured in our method (Figure 1) and are in conformity (lower false positives in network connectivity) with pathway entries in *Entrez Gene* [15] as well as in recent reviews on kidney expression [8, 12] (also, see Table 8). For example, the *Mapk1-Rara* [26] or the *Pax2-Gdf11* [27] interactions are completely missed in Figure 2—this is seen to be the case since these interactions only occur during the 10.5–12.5 dpc regime. We also see that the *Acvr2b-Lamc2* [28] interaction is observed in the steady state but not in the first regime. This interaction becomes active in the second regime (first via the *Acvr2b-Gdf11* and then via the *Gdf11-Lamc2*), indicating that it might not have particular relevance in the day 10.5–12.5 dpc stage. Several of these predicted interactions need to be experimentally characterized in the laboratory. It is especially interesting to see the *Rara* gene in this network, because it is known that *Gata3* [29, 30] has tissue-specific expression in some cells of the developing eye. Also *Gdf11* exhibits growth factor activity and is extremely important during organ formation.

In Figure 3, we give the results of the CoD approach of network inference. Here the *Gata3-Pax2* interaction seems reversed and counterintuitive. As can be seen, some of the interactions (e.g., *Pax2-Gata3*) can be seen here (via other nodes: *Mapk1-Wnt11*), but there is a need to resolve cycles (*Ros1–Wnt11-Mapk1*) and feedback/feedforward loops (*Bmp7-Gata3-Wnt11*). Both of these topologies can convey potentially useful information about nephric development. Thus a potentially useful way to combine these two methods is to "seed" the network using CoD and then try to resolve cycles using *regime-SSM*.

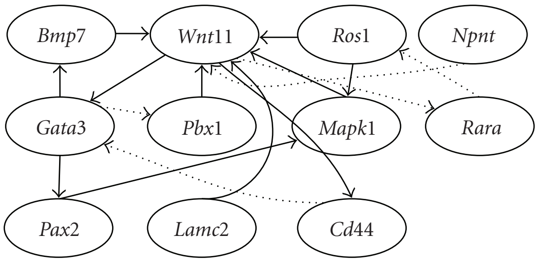

Figure 3: Steady-state network inferred using CoD (solid lines represent the first regime, and the dotted lines indicate the second regime).

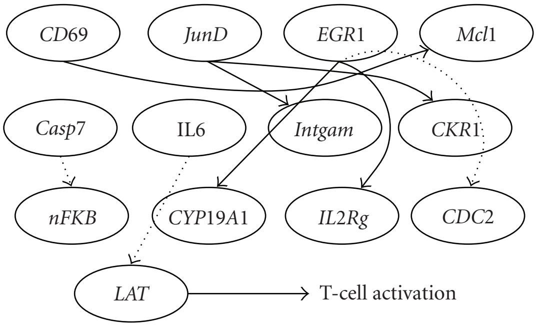

Figure 4: Steady-state network inferred using SSM (solid lines represent the first regime, and the dotted lines indicate the second regime).

8.2. T-cell activation

The *regime-SSM* network is shown in Figure 4. The corresponding network learnt in each regime using CoD is also shown (Figure 5). The study of this network using GGM (for the whole time-series data) is already available in [22]. Though there are several interactions of interest discovered in both the SSM and CoD procedures, we point out a few of interest. It is already known that synergistic interactions between IL-6 and IL-1 are involved in T-cell activation [31]. IL-2 receptor transcription is affected by EGR1 [32]. An examination of the topology of these two networks (CoD and SSM) would indicate some matches and is worth pursuing for experimental investigation. However, as already alluded to above, we have to find a way to resolve cycles from the CoD network [33]. Several of these match the interactions reported in [1, 22]. However, the additional information that we can glean is that some of the key interactions occur during "early response" to stimulation and some occur subsequently (interleukin-6 mediated T-cell activation) in the "late phase."

An examination of the gene ontology (GO) terms represented in each cluster as well as the functional annotations in *Entrez Gene* shows concordance with literature findings (Table 9). Because this dataset has been the subject of several interesting investigations, it would be ideal to ask other questions related to network inference procedures, for the purpose of comparison. One of the primary questions we seek

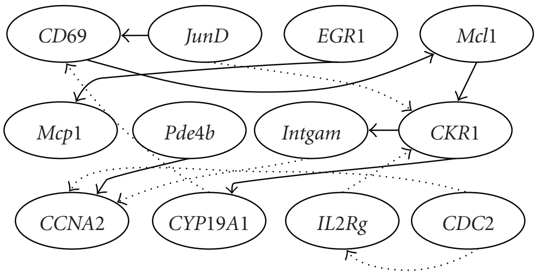

Figure 5: Steady-state network inferred using CoD (solid lines represent the first regime, and the dotted lines indicate the second regime).

to answer is what is the performance of the network inference procedure if a subsampled trajectory is used instead?

In Table 7, the performances of the CoD and SSM algorithms are summarized. Using the T-cell (10 points, 44 replicates) data, we infer a network using the SSM procedure. With the identified edges as the gold standard for comparison, we now use SSM network inference on an undersampled version of this time series (5 points, 44 replicates) and check for any new edges (f_{new}) or deletion of edges (f_{lost}). Ideally, we would want both these numbers to be zero. f_{new} is the fraction of new edges added to the original set and f_{lost} is number of edges lost from the original data network over both regimes. Further, we now interpolate this undersampled data to 10 points and carry out network inference. This is done for each of the identified regimes. The same is done for the CoD method. We note that this is not a comparison between SSM and CoD (both work with very different assumptions), but of the effect of undersampling the data and subsequently interpolating this undersampled data to the original data length (via resampling). Table 7 suggests that as expected, there is degradation in performance (SSM/CoD) in the absence of all the available information. However, it is preferred to infer some false positives rather than lose true positive edges. This also indicates that interpolated data does not do worse than the undersampled data in terms of true positives (f_{lost}).

We make three observations regarding this method of network inference.

(i) It is not necessary for the target gene (*Gata2/Gata3*) to be present as part of the inferred network. We can obtain insight into the mechanisms underlying transcription in each regime even if some of the genes with *similar coexpression dynamics* as the target gene(s) are present in the inferred network.

(ii) Probe-level observations from a small number of biological replicates seem to be very informative for network inference. This is because the LDS parameter estimation algorithm uses these *multiple* expression realizations to iteratively estimate the state mean, covariance and other parameters, notably D [17]. Hence inspite of few time points, we can use multiple measurements (biological, technical, and probe-level replicates) for reliable network inference. This follows similar observations in [34] that probe-level replicates are very useful for understanding intergene relationships.

(iii) Following [24], it would seem that several network hypotheses can individually explain the time evolution behavior captured by the expression data. The LDS parameter estimation procedure seeks to find a maximum-likelihood (ML) estimate of the system parameters A, B, C, and D and then finally uses bootstrapping to only infer high confidence interactions. This ML estimation of the parameters uses an EM algorithm with multiple starts to avoid initialization-related issues [17], and thus finds the "most consistent" hypothesis which would explain the evolution of expression data. It is this network hypothesis that we investigate. Since this network already contains our gene of interest *Gata3*, we can proceed to verify these interactions from literature and experimentally.

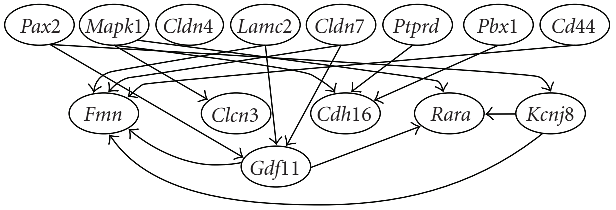

Figure 6: Steady-state network inferred using GGMs.

9. DISCUSSION

One of the primary motivations for computational inference of state specific gene influence networks is the understanding of transcriptional regulatory mechanisms [36]. The networks inferred via this approach are fairly general, and thus there is a need to "decompose" these networks into transcriptional, signal transduction or metabolic using a combination of biological knowledge and chemical kinetics. Depending on the insights expected, the tools for dissection of these predicted influences might vary.

For comparison, we additionally investigated a *graphical Gaussian model (GGM)* approach as suggested in [35] using partial correlation as a metric to quantify influence (Figure 6). This method works for short time-series data but we could not find a way to incorporate previous expression values as inputs to the evolution of state or individual observations—something we could explicitly do in the state-space approach. However, we are now in the process of examining the networks inferred by the GGM approach over the regimes that we have identified from SSA. Again, we observe that the network connections reflect a steady-state behavior and that transient (state-specific) changes in influence are not fully revealed. The same is observed in the case of the T-cell data, from the results reported in [22]. A comparison of all the presented methods, along with *regime-SSM,* has been presented in Table 10. The comparisons are based

Table 7: Functional annotations (*Entrez Gene*) of some of the genes coclustered with *Gata2* and *Gata3*.

Gene symbol	Gene name	Possible role in nephrogenesis (function)
Bmp7	Bone morphogenetic protein	Cell signaling
Rara	Retinoic acid receptor	Retinoic acid pathway, related to eye phenotype
Gata2	GATA binding protein 2	Hematopoiesis, urogenital development
Gata3	GATA binding protein 3	Hematopoiesis, urogenital development
Pax2	Paired homeobox-2	Direct target of *Gata2*
Lamc2	Laminin	Cell adhesion molecule
Npnt	Nephronectin	Cell adhesion molecule
Ros1	Ros1 proto-oncogene	Signaling epithelial differentiation
Ptprd	protein tyrosine phosphatase	Cell adhesion
Ret-Gdnf	Ret proto-oncogene, Glial neutrophic factor	Metanephros development
Gdf11	Growth development factor	Cell-cell signaling and adhesion
Mapk1	Mitogen-activated protein kinase 1	Role in growth factor activity, cell adhesion
Kcnj8	potassium inwardly rectifying channel, subfamily J, member 8	Potassium ion transport
Acvr2b	Activin receptor IIB	Transforming growth factor-beta receptor activity

Table 8: Functional annotations of some of the coclustered genes (early and late responses) following T-cell activation.

Gene symbol	Gene name	Possible role in T-cell activation (function)
CD69	*CD69* antigen	Early T-cell activation antigen
Mcl1	Myeloid cell leukemia sequence 1 (BCL2-related)	Mediates cell proliferation and survival
IL6	Interleukin 6	Accessory factor signal
LAT	Linker for activation of T cells	Membrane adapter protein involved in T-cell activation
EGR1	Early growth response gene 1	activates nFKB signaling
CDC2	Cell division control protein 2	Involved in cell-cycle control
Casp7	Caspase 7	Involved in apoptosis
JunD	Jun D proto-oncogene	Regulatory role in T lymphocyte proliferation and T_h cell differentiation
CKR1	Chemokine receptor 1	negative regulator of the antiviral CD8+ T-cell response
CYP19A1	Cytochrome P450, member 19	cell proliferation
Intgam	Integrin alpha M	Mediates phagocytosis-induced apoptosis
nFKB	nFKB protein	Signaling transduction activity
IL2Rg	Interleukin-2 receptor gamma	Signaling activity
Pde4b	Phosphodiesterase 4B, cAMP-specific	Mediator of cellular response to extracellular signal
Mcp1	Monocyte chemotactic protein 1	Cytokine gene involved in immunoregulation
CCNA2	Cyclin A2	Involved in cell-cycle control

Table 9: Results of network inference on original, subsampled, and interpolated data.

Method (T-cell data)	Edges inferred	f_{new}	f_{lost}
SSM on original data	14	—	—
SSM on undersampled data	—	3	3
SSM on interpolated data	—	4	2
CoD on original data	12	—	—
CoD on undersampled data	—	3	2
CoD on interpolated data	—	4	2

on whether these frameworks permit the inference of directional influences, regime specificity, resolution of cycles, and modeling of higher lags.

10. CONCLUSIONS

In this work, we have developed an approach (*regime-SSM*) to infer the time-varying nature of gene influence network topologies, using gene expression data. The proposed approach integrates change-point detection to delineate phases

Table 10: Comparison of various network inference methods (Y: Yes, N: No).

Method	Direction	Regime-specific	Resolve cycles	Higher lags (> 1)	Nonlinear/locally linear
CoD [20, 21]	Y	Y	N	N	Y
GGM [35]	Y	N	N	N	Y
SSM [1]	Y	N	Y	Y	Y
Regime-SSM	Y	Y	Y	Y	Y

of gene coexpression, MoG clustering implying possible coregulation, and network inference amongst the regime-specific coclustered genes using a state-space framework. We can thus incorporate condition specificity of gene expression dynamics for understanding gene influences. Comparison of the proposed approach with other current procedures like GGM or CoD reveals some strengths and would very well complement existing approaches (Table 10). We believe that this approach, in conjunction with sequence and transcription factor binding information, can give very valuable clues to understand the mechanisms of transcriptional regulation in higher eukaryotes.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the NIH under Award 5R01-GM028896-21 (JDE). The authors also thank the three anonymous reviewers for constructive comments to improve this manuscript. The material in this paper was presented in part at the IEEE International Workshop on Genomic Signal Processing and Statistics 2005 (GENSIPS05).

REFERENCES

[1] C. Rangel, J. Angus, Z. Ghahramani, et al., “Modeling T-cell activation using gene expression profiling and state-space models,” *Bioinformatics*, vol. 20, no. 9, pp. 1361–1372, 2004.

[2] B.-E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. D’Alché-Buc, “Gene networks inference using dynamic Bayesian networks,” *Bioinformatics*, vol. 19, supplement 2, pp. II138–II148, 2003.

[3] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, “Genomic analysis of regulatory network dynamics reveals large topological changes,” *Nature*, vol. 431, no. 7006, pp. 308–312, 2004.

[4] E. Sontag, A. Kiyatkin, and B. N. Kholodenko, “Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data,” *Bioinformatics*, vol. 20, no. 12, pp. 1877–1886, 2004.

[5] S. Kim, H. Li, D. Russ, et al., “Context-sensitive probabilistic Boolean networks to mimic biological regulation,” in *Proceedings of Oncogenomics*, Phoenix, Ariz, USA, January-February 2003.

[6] H. Li, C. L. Wood, Y. Liu, T. V. Getchell, M. L. Getchell, and A. J. Stromberg, “Identification of gene expression patterns using planned linear contrasts,” *BMC Bioinformatics*, vol. 7, p. 245, 2006.

[7] M. A. T. Figueiredo and A. K. Jain, “Unsupervised learning of finite mixture models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[8] R. O. Stuart, K. T. Bush, and S. K. Nigam, “Changes in gene expression patterns in the ureteric bud and metanephric mesenchyme in models of kidney development,” *Kidney International*, vol. 64, no. 6, pp. 1997–2008, 2003.

[9] M. Khandekar, N. Suzuki, J. Lewton, M. Yamamoto, and J. D. Engel, “Multiple, distant Gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system,” *Molecular and Cellular Biology*, vol. 24, no. 23, pp. 10263–10276, 2004.

[10] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky, *Analysis of Time Series Structure—SSA and Related Techniques*, Chapman & Hall/CRC, New York, NY, USA, 2001.

[11] V. Moskvina and A. Zhigljavsky, “An algorithm based on singular spectrum analysis for change-point detection,” *Communications in Statistics Part B: Simulation and Computation*, vol. 32, no. 2, pp. 319–352, 2003.

[12] K. Schwab, L. T. Patterson, B. J. Aronow, R. Luckas, H.-C. Liang, and S. S. Potter, “A catalogue of gene expression in the developing kidney,” *Kidney International*, vol. 64, no. 5, pp. 1588–1604, 2003.

[13] Y. Zhou, K.-C. Lim, K. Onodera, et al., “Rescue of the embryonic lethal hematopoietic defect reveals a critical role for GATA-2 in urogenital development,” *The EMBO Journal*, vol. 17, no. 22, pp. 6689–6700, 1998.

[14] G. A. Challen, G. Martinez, M. J. Davis, et al., “Identifying the molecular phenotype of renal progenitor cells,” *Journal of the American Society of Nephrology*, vol. 15, no. 9, pp. 2344–2357, 2004.

[15] NCBI Pubmed, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.

[16] H. H. Zadeh, S. Tanavoli, D. D. Haines, and D. L. Kreutzer, “Despite large-scale T cell activation, only a minor subset of T cells responding *in vitro* to *Actinobacillus actinomycetemcomitans* differentiate into effector T cells,” *Journal of Periodontal Research*, vol. 35, no. 3, pp. 127–136, 2000.

[17] Z. Ghahramani and G. E. Hinton, “Parameter estimation for linear dynamical systems,” Tech. Rep., University of Toronto, Toronto, Ontario, Canada, 1996.

[18] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Applications*, Springer Texts in Statistics, Springer, New York, NY, USA, 2000.

[19] B. Effron, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York, NY, USA, 1993.

[20] E. R. Dougherty, S. Kim, and Y. Chen, “Coefficient of determination in nonlinear signal processing,” *Signal Processing*, vol. 80, no. 10, pp. 2219–2235, 2000.

[21] S. Kim, E. R. Dougherty, M. L. Bittner, et al., “General nonlinear framework for the analysis of gene interaction via multivariate expression arrays,” *Journal of Biomedical Optics*, vol. 5, no. 4, pp. 411–424, 2000.

[22] R. Opgen-Rhein and K. Strimmer, “Using regularized dynamic correlation to infer gene dependency networks from

time-series microarray data," in *Proceedings of the 4th International Workshop on Computational Systems Biology (WCSB '06)*, Tampere, Finland, June 2006.

[23] A. O. Hero III, G. Fleury, A. J. Mears, and A. Swaroop, "Multicriteria gene screening for analysis of differential expression with DNA microarrays," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 1, pp. 43–52, 2004, special issue on genomic signal processing.

[24] Z. Bar-Joseph, "Analyzing time series gene expression data," *Bioinformatics*, vol. 20, no. 16, pp. 2493–2503, 2004.

[25] A. Kundaje, O. Antar, T. Jebara, and C. Leslie, "Learning regulatory networks from sparsely sampled time series expression data," Tech. Rep., Columbia University, New York, NY, USA, 2002.

[26] J. E. Balmer and R. Blomhoff, "Gene expression regulation by retinoic acid," *Journal of Lipid Research*, vol. 43, no. 11, pp. 1773–1808, 2002.

[27] A. F. Esquela and S. E.-J. Lee, "Regulation of metanephric kidney development by growth/differentiation factor 11," *Developmental Biology*, vol. 257, no. 2, pp. 356–370, 2003.

[28] A. Maeshima, S. Yamashita, K. Maeshima, I. Kojima, and Y. Nojima, "Activin a produced by ureteric bud is a differentiation factor for metanephric mesenchyme," *Journal of the American Society of Nephrology*, vol. 14, no. 6, pp. 1523–1534, 2003.

[29] M. Mori, N. B. Ghyselinck, P. Chambon, and M. Mark, "Systematic immunolocalization of retinoid receptors in developing and adult mouse eyes," *Investigative Ophthalmology and Visual Science*, vol. 42, no. 6, pp. 1312–1318, 2001.

[30] K.-C. Lim, G. Lakshmanan, S. E. Crawford, Y. Gu, F. Grosveld, and J. D. Engel, "Gata3 loss leads to embryonic lethality due to noradrenaline deficiency of the sympathetic nervous system," *Nature Genetics*, vol. 25, no. 2, pp. 209–212, 2000.

[31] H. Mizutani, L. T. May, P. B. Sehgal, and T. S. Kupper, "Synergistic interactions of IL-1 and IL-6 in T cell activation. Mitogen but not antigen receptor-induced proliferation of a cloned T helper cell line is enhanced by exogenous IL-6," *Journal of Immunology*, vol. 143, no. 3, pp. 896–901, 1989.

[32] J.-X. Lin and W. J. Leonard, "The immediate-early gene product Egr-1 regulates the human interleukin- 2 receptor β-chain promoter through noncanonical Egr and Sp1 binding sites," *Molecular and Cellular Biology*, vol. 17, no. 7, pp. 3714–3722, 1997.

[33] M. J. Herrgård, M. W. Covert, and B. Ø. Palsson, "Reconciling gene expression data with known genome-scale regulatory network structures," *Genome Research*, vol. 13, no. 11, pp. 2423–2434, 2003.

[34] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 1, pp. 31–36, 2001.

[35] J. Schäfer and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks," *Bioinformatics*, vol. 21, no. 6, pp. 754–764, 2005.

[36] A. Rao, A. O. Hero III, D. J. States, and J. D. Engel, "Inference of biologically relevant gene influence networks using the directed information criterion," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 2, pp. 1028–1031, Toulouse, France, May 2006.