BIOINFORMATICS

Vol. 20 no. 9 2004, pages 1361–1372
DOI: 10.1093/bioinformatics/bth093

Modeling T-cell activation using gene expression profiling and state-space models

Claudia Rangel[1,5], John Angus[1], Zoubin Ghahramani[2], Maria Lioumi[3], Elizabeth Sotheran[3], Alessia Gaiba[4], David L. Wild[5,*] and Francesco Falciani[6]

[1]School of Mathematical Sciences, Claremont Graduate University, 121 E. Tenth St., Claremont, CA 91711, USA, [2]Gatsby Computational Neuroscience Unit, University College London, 17 Queen Square, London, WC1N 3AR, UK, [3]Lorantis Limited, 307 Cambridge Science Park, Cambridge, CB4 OWG, UK, [4]Department of Oncology, University of Bologna, Bellaria Hospital, Bologna, Italy, [5]Keck Graduate Institute of Applied Life Sciences, 535 Watson Drive, Claremont, CA 91171, USA and [6]School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

Received on August 25, 2003; revised on November 20, 2003; accepted on December 18, 2003
Advance Access publication February 12, 2004

ABSTRACT

Motivation: We have used state-space models to reverse engineer transcriptional networks from highly replicated gene expression profiling time series data obtained from a well-established model of T-cell activation. State space models are a class of dynamic Bayesian networks that assume that the observed measurements depend on some hidden state variables that evolve according to Markovian dynamics. These hidden variables can capture effects that cannot be measured in a gene expression profiling experiment, e.g. genes that have not been included in the microarray, levels of regulatory proteins, the effects of messenger RNA and protein degradation, etc.

Results: Bootstrap confidence intervals are developed for parameters representing 'gene–gene' interactions over time. Our models represent the dynamics of T-cell activation and provide a methodology for the development of rational and experimentally testable hypotheses.

Availability: Supplementary data and Matlab computer source code will be made available on the web at the URL given below.

Contact: david_wild@kgi.edu; f.falciani@bham.ac.uk

Supplementary information: http://public.kgi.edu/~wild/LDS/index.htm

INTRODUCTION

The application of high-density DNA microarray technology to gene transcription analysis has been responsible for a real paradigm shift in biology. The majority of research groups now have the ability to measure the expression of a significant proportion of an organism's genome in a single experiment, resulting in an unprecedented volume of data being made available to the scientific community. This has in turn stimulated the development of algorithms to classify and describe the complexity of the transcriptional response of a biological system, but efforts toward developing the analytical tools necessary to exploit this information for revealing interactions between the components of a cellular system are still in their early stages. The availability of such tools would allow a large-scale systematic approach to pathway reconstruction in a large spectrum of organisms. The popular use of clustering techniques, reviewed in Dopazo *et al*. (2001), while providing putative classes and allowing qualitative inferences about the co-regulation of certain genes to be made, does not provide models of the underlying transcriptional networks that lend themselves to statistical hypothesis testing.

Many of the tools that have been applied in an exploratory way to the problem of reverse engineering genetic regulatory networks from gene expression data have been recently reviewed by van Someren *et al*. (2002). These include Boolean networks (Akutsu *et al*., 1999; Liang *et al*., 1998; Thomas, 1973), time-lagged cross-correlation functions (Arkin *et al*., 1997), differential equation models (Kholodenko *et al*., 2002) and linear and non-linear autoregression models (D'Haeseleer *et al*., 1999; van Someren *et al*., 2000; Holter *et al*., 2001; Weaver *et al*., 1999). Murphy and Mian (1999) have shown that many of these published models can be considered special cases of a general class of graphical models known as dynamic Bayesian networks (DBNs). Bayesian networks have a number of features that make them attractive candidates for modeling gene expression data, such as their ability to handle noisy or missing data, to handle hidden variables such as protein levels that may have an effect on messenger RNA (mRNA) expression levels and to describe locally interacting processes and the possibility of making causal

*To whom correspondence should be addressed.

inferences from the derived models. Following the pioneering work of Friedman *et al*. (2000), a number of other authors have described Bayesian network models of gene expression data. Although microarray technologies have made it possible to measure time series of the expression level of many genes simultaneously, we cannot hope to measure all possible factors contributing to genetic regulatory interactions, and the ability of Bayesian networks to handle such hidden variables would appear to be one of their main advantages as a modeling tool. However, most published work to date has only considered either static Bayesian networks with fully observed data (Pe'er *et al*., 2001) or static Bayesian networks that model discretized data but incorporate hidden variables (Cooper and Herskovits, 1992; Yoo *et al*., 2002). Ong *et al*. (2002) have described an DBN model for *Escherichia coli* that explicitly includes operons as hidden variables but again uses discretized gene expression measurements. There appears to be the need, therefore, for a dynamic modeling approach that can both accomodate gene expression measurements as continuous, rather than discrete, variables and that can model unknown factors as hidden variables.

We have applied linear state-space modeling to reverse engineer transcriptional networks from highly replicated expression profiling data obtained from a well-established model of T-cell activation in which we have monitored a set of relevant genes across a time series (Rangel *et al*., 2001, 2004). Linear-Gaussian state-space models (SSMs), also known as linear dynamical systems (Roweis and Ghahramani, 1999) or Kalman filter models (Brown and Hwang, 1997), are a subclass of DBNs used for modeling time series data and have been used extensively in many areas of control and signal processing. SSM models have a number of features that make them attractive for modeling gene expression time series data. They assume the existence of a hidden state variable from which we can make noisy continuous measurements, which evolves with Markovian dynamics. In our application, the noisy measurements are the observed gene expression levels at each time point, and we assume that the hidden variables are modeling effects that cannot be measured in a gene expression profiling experiment, e.g. the effects of genes that have not been included on the microarray, levels of regulatory proteins, the effects of mRNA and protein degradation, etc. Our SSMs have produced testable hypotheses that have the potential for rapid experimental validation.

SYSTEMS AND METHODS

The biological system

The central event in the generation of an immune response is the activation of T-lymphocytes. Activated T-cells proliferate and produce cytokines involved in the regulation of effector cells (i.e. B cells and macrophages), which are the primary mediators of the immune response. T-cell activation is initiated by the interaction between the T-cell receptor (TCR) complex and the antigenic peptide presented on the surface of an antigen-presenting cell. This event triggers a network of signaling molecules, including kinases, phosphatases and adaptor proteins that couple the stimulatory signal received from the TCR to gene transcription events in the nucleus (Iwashima *et al*., 1994; Ley *et al*., 1991).

Activation leads to the transcription of a number of target genes. Immediate genes, such as the transcription factors *c*-Fos, *c*-myc, *c*-jun, NF-AT and NF-κB are activated within the first 0.5 h after TCR stimulation. Early genes such as interleukins (e.g. IL-2, IL-2R, IL-3, IL-6, IFN-γ) are activated within the first 2 h. IL-2 is the paradigm of a pro-inflammatory cytokine. Once secreted, it acts as a powerful proliferation stimulus and induces the expression of a number of effector genes. Days after the activation event, various adhesion molecules begin to be expressed. These influence the migratory and adhesion properties of activated lymphocytes (Iwashima, 2003).

In this paper, we describe the application of linear state-space modeling to identifying genetic regulatory networks in the activation of T-cells. We have used a well-established model of T-cell activation based on the stimulation of a lymphoblast cell line (Jurkat) with the calcium ionophore ionomycin and the PKC activator phorbol ester PMA (Manger *et al*., 1987). This treatment bypasses the TCR requirement and thereby activates signaling transduction pathways (Castagna *et al*., 1982) leading to T-cell activation.

SSMs (linear dynamical systems)

In linear SSMs, a sequence of p-dimensional observation vectors $\{y_1, \ldots, y_T\}$ is modeled by assuming that at each time step, y_t was generated from a K-dimensional hidden-state variable that we denote by x_t and that the sequence $\{x_1, \ldots, x_T\}$ defines a first-order Markov process. The most basic linear SSM can be described by the following two equations:

$$x_{t+1} = Ax_t + w_t, \quad (1)$$

$$y_t = Cx_t + v_t, \quad (2)$$

where A is the state dynamics matrix, C is the state to observation matrix and $\{w_t\}$ and $\{v_t\}$ are uncorrelated white noise sequences.

SSM with inputs Often, the observations can be divided into a set of input (or exogenous) variables and a set of output (or response) variables. Allowing inputs to both the state and observation equations, the equations describing the linear SSM then become

$$x_{t+1} = Ax_t + Bh_t + w_t, \quad (3)$$

$$y_t = Cx_t + Du_t + v_t, \quad (4)$$

where h_t, u_t are the inputs to the state and observation vectors, A is the state dynamics matrix, B is the input to state matrix,

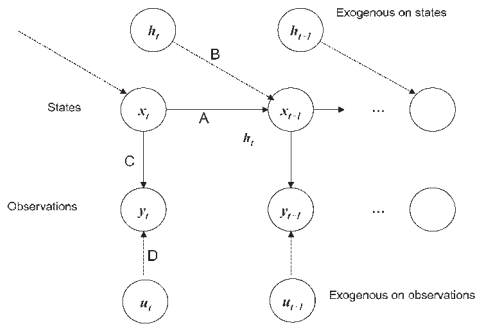

Fig. 1. SSM with inputs.

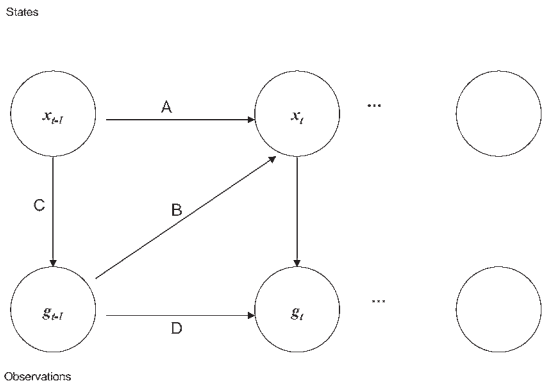

Fig. 2. Bayesian network representation of the model for gene expression.

C is the state to observation matrix and D is the input to observation matrix. A Bayesian network representation of this model is shown in Figure 1.

The state and observation noise sequences, $\{w_t\}$ and $\{v_t\}$, respectively, are generally taken to be white noise sequences, with $\{w_t\}$ and $\{v_t\}$ orthogonal to one another. Note that the noise vectors may also be considered hidden variables. The unknown parameters of the SSM may be estimated or learned from data using the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977; Shumway and Stoffer, 1982; Ghahramani and Hinton, 1996; Rangel *et al.*, 2001, 2004). In the application of the EM algorithm to the SSM, there is little harm in making the additional assumption that the noise sequences are Gaussian distributed and independent of the initial values of x and y. If there are no extreme outliers, this leads to fairly robust parameter estimates that are maximum-likelihood estimates if the Gaussian assumption is reasonable and weighted least squares estimates otherwise. We test the validity of both Gaussian and independent and identically distributed (iid) assumptions by examining residuals as described in the Implementation section.

SSM for gene expression The fluorescent intensities measured in a microarray experiment are noisy measures of gene expression levels. Values of some of these variables influence the values of others through the regulatory proteins they express, including the possibility that the expression of a gene at one time point may, in various circumstances, influence the expression of the same or other genes at a later time point. The time steps in the model do not have to correspond with a fixed unit of real time, and we have chosen to model each sample in the experimental time series as a single step in the SSM.

To model the effects of the influence of the expression of one gene at a previous time point on another gene and its associated hidden variables, we modified the SSM with inputs (3, 4) described above as follows. Letting g_t be the (suitably transformed[1]) vector of gene expression levels measured at time t, we take $y_t = g_t$ and the inputs $h_t = g_t$ and $u_t = g_{t-1}$ to give the model shown in Figure 2.

[1] We use log transformation and normalization as described.

This model is described by the following equations:

$$x_{t+1} = Ax_t + Bg_t + w_t. \quad (5)$$

$$g_t = Cx_t + Dg_{t-1} + v_t. \quad (6)$$

Here, the matrix D in the observation equation captures gene–gene expression level influences at consecutive time points while the matrix C captures the influence of the hidden variables on gene expression level at each time point. Matrix B models the influence of gene expression values from previous time points on the hidden states, and A is the state dynamics matrix. However, our interests focus on $CB + D$, which not only captures the direct gene-to-gene interaction but also the gene-to-gene interactions 'through' the hidden states over time. This is the matrix we will concentrate our analysis on since it captures all the information related to gene–gene interaction over one time step. We have also shown that if the gene expression model is stable, controllable and observable, then the $CB + D$ matrix remains invariant to any coordinate transformations of the state and is, therefore, identifiable (Rangel *et al.*, 2004). The identifiability property is important, for without it, it would be possible for different values of the SSM parameters (and hence, different values of $CB + D$) to give rise to identically distributed observables, making the statistical problem of estimation ill-posed.

Cell culture, treatments and RNA extraction

The data used in this paper are the results of two experiments that we have performed to characterize the response of a human T-cell line (Jurkat) to PMA and ionomycin treatment. In the first experiment, we monitored the expression of 88 genes using cDNA array technology across 10 time points. In the second experiment, an identical experimental protocol was used, but additional genes were added to the arrays. Data were combined, and genes with high-experimental variation were eliminated from the data set as described below. Jurkat cells were cultured in RPMI 1640 (GibcoBRL) supplemented with 2 mM glutamine (GibcoBRL) and

penicillin–streptomycin 50 units/ml (GibcoBRL) and with 10% fetal bovine serum (FBS; Biochrom KG). When the culture reached the density of 10^6 cells/ml, cells were treated with 50 ng/ml of Phorbol ester PMA (Sigma) plus 1 μg/ml of ionomycin (Sigma). Cells were collected in 300 μl of RTL lysing solution (Qiagen) at the following times after treatment: 0, 2, 4, 6, 8, 18, 24, 32, 48, 72 h. In order to ensure the efficacy of the stimulation, cells were tested for the correct expression of T-cell and activation markers using Fluorescence-activated cell scanning (FACS) analysis. The cells used in this experiment were all expressing the T-cell receptor (detected with anti CD3 antibodies) and after 24 h of stimulation strongly upregulate CD69, an early surface activation marker. RNA was then extracted using an RNA easy miniprep kit (Qiagen) according to the manufacturer's instructions.

Gene expression profiling

Microarrays were manufactured by spotting purified polymerase chain reaction (PCR) products on amino-modified glass slides (Hegde *et al*., 2000) using a Microgrid II spotter (Biorobotics, Cambridge, UK). The two replicated experiments were hybridized on two sets of arrays. For the first experiment, microarrays representing 34 replications of each gene were manufactured. The second experiment employed arrays with each gene replicated 10 times. Microarray probes were prepared by labeling 40 μg of total RNA by a reverse transcriptase reaction incorporating dCTP–Cy3 labeled nucleotide. Probe labeling and purification was then performed as described in previous sections. Purified probes were then hybridized on the arrays for 2 days at 42°C in a 25% formaldehyde, 5× SSC, 0.1% sodium dodecyl sulfate (SDS) solution. Slides were washed twice in 2× SSC, 0.2% SDS for 5 min at room temperature and finally once in 2× SSC, 0.2% SDS for 5 min at room temperature. Once dried, the slides were scanned on a GSI lumonics confocal scanner at 100% laser power and 70% photomultiplier tube efficiency.

Slide images were processed as follows. Array spots representing the signal associated with individual spotted clones were identified and quantified using the quantarray application (Packard). Numeric values for the gene expression intensities were calculated using the histogram method implemented in the same application. Values were calculated as integrals of the pixel signal distribution associated with each spot, and local background values were subtracted.

Data pre-processing

In this work, we have pre-selected genes that are all modulated in response to activation. Genes whose expression values in all the time points were below a defined value were filtered out of the analysis. This threshold was estimated as being associated with a 99% probability that a signal corresponded to an expressed gene. The figure was derived by estimating the signal probability distribution from 250 negative control spots in the experimental slides after 500 bootstrap replications. After this step, genes that displayed very poor reproducibility between the two experiments were removed, leaving 58 genes.

Normalization methods aim at removing systematic variation due to experimental artifacts or at least minimizing this variability. With two 'biological' replicates of the experiment and several 'technical' replicates of each measurement, it was necessary for all replicates of the expression profiles of the same genes to be normalized or scaled together. Two color normalization methods (Yang *et al*., 2002) could not be used because the data were generated using a single dye.

After log transformation, expression profiles for the same gene in the two experiments were scaled together using a variant of the Quantile Normalization method of (Bolstad *et al*., 2002). As published, this method is based on the assumption that there is an underlying common distribution of intensities across arrays. This method was adapted to our data with the assumption that all 44 replicates have a similar underlying distribution.

Distributions of the 44 replicates of all genes, in the form of boxplots, and gene expression profiles before and after quantile normalization are shown in the supplementary information on the associated website (http://public.kgi.edu/∼wild/LDS/index.htm).

IMPLEMENTATION

Determining state dimensions by cross-validation

The first parameter to estimate for the SSM described by (5–6) is the optimal number of hidden states. This can be determined by a cross-validation experiment in which we increment the number of hidden states and monitor the predictive likelihood using a portion of the data set that has not been used to train the model. A special case of cross-validation was implemented, the so called leave-one-out method, which is a general method to estimate the predictive accuracy of the learning algorithm. In general, the cross-validation analysis consists of four steps:

(1) Begin with $K = 1$, where K is the number of hidden states.

(2) Split the data into two parts, an evaluation set $\mathcal{E}$ and a training set $\mathcal{T} = \text{Data} - \mathcal{E}$, where $\mathcal{E}$ is a set of one replicate of the complete time series for all genes.

(3) An SSM is trained on $\mathcal{T}$ and then the likelihood is evaluated on both the training data, $\mathcal{T}$, and the evaluation data, $\mathcal{E}$.

(4) Increase K, go to step 2.

Figure 3 shows the behavior of the likelihood for both the training data and the evaluation data. As expected, the likelihood for the training data continues to increase as the number of hidden states increases since the model fits the data better and better as the number of parameters (in this case, hidden

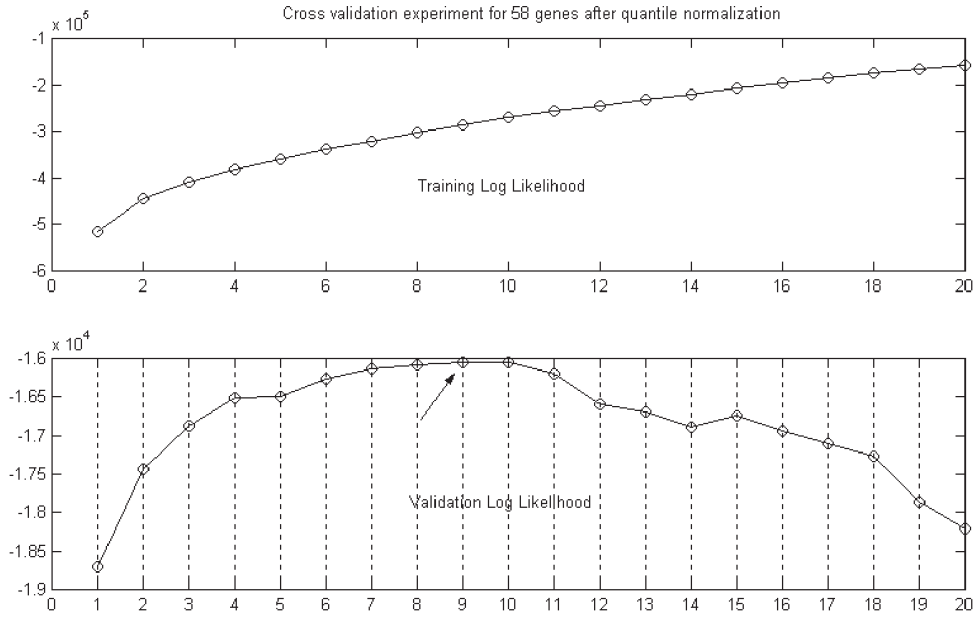

Fig. 3. Cross-validation experiment to determine the number of hidden states.

states) increases. Over-fitting and under-fitting are avoided by choosing the number of hidden states at which the likelihood of the evaluation data (not used in training) reaches its maximum. The bottom plot shows this optimum number of hidden states to be $K = 9$.

Bootstrap analysis

For the SSM defined by the two Equations (5) and (6), estimates of the structural parameters for this model $[\hat{A}, \hat{B}, \hat{C}, \hat{D}]$, as well as estimates of the noise covariances $\hat{Q}, \hat{R}$, are computed using the EM algorithm as described in Rangel *et al.* (2001, 2004).

In this research, we collected replicated sequences of observations of the gene expression vector $g_t, t = 1, 2, \ldots, T$. The key idea in the bootstrap procedure is to resample with replacement the replicates within the original data. By resampling from the replicates N_B times (where the value N_B is a large number, say 200 or 300), we can estimate, among other things, the sampling distributions of the estimators of the elements of $CB + D$, which is the identifiable gene–gene interaction matrix in the gene expression model (5) and (6). In general, once we have estimates of these distributions, we can make statistical inferences about those underlying parameters (in particular, confidence intervals and hypothesis tests).

Each replicate represents a reproduction of the same experiment under the same circumstances and assumptions. Hence replicates are assumed to be iid with unknown (multivariate) cumulative probability distribution F_0. That is, the i-th replicate consists of a time series $Y_i = \left(g_1^i, g_2^i, \ldots, g_T^i\right)$ with each g_t^i a p-dimensional vector (one component for each gene). Thus, the collection $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]$ can be viewed as a sequence of N iid random matrices, each with cumulative distribution F_0. Under this assumption, a bootstrap sample $\mathbf{Y}^* = [Y_1^*, Y_2^*, \ldots, Y_N^*]$ is obtained by selecting at random with replacement, N elements from $[Y_1, Y_2, \ldots, Y_N]$.

The following is the bootstrap procedure for the model (5, 6) with data collected as described above. We denote a generic element of the matrix $CB + D$ by θ. The following steps lead to a bootstrap confidence interval for θ using the percentile method.

(1) Calculate estimates for the unknown matrices A, B, C, D from the full data set with replicates using the EM algorithm. From the estimates $\hat{B}, \hat{C}, \hat{D}$, compute $\hat{\theta}$, the estimate of the given element of $CB + D$.

(2) Generate N_B independent bootstrap samples $\mathbf{Y}_1^*$, $\mathbf{Y}_2^*, \ldots, \mathbf{Y}_{N_B}^*$ from the original data.

(3) For each bootstrap sample compute bootstrap replicates of the parameters. This is done using the EM algorithm on each bootstrap sample $\mathbf{Y}_i^*, i = 1, 2, \ldots, N_B$. This yields bootstrap estimates of the parameters $\{\hat{A}_1^*, \hat{B}_1^*, \hat{C}_1^*, \hat{D}_1^*\}, \{\hat{A}_2^*, \hat{B}_2^*, \hat{C}_2^*, \hat{D}_2^*\}, \ldots, \{\hat{A}_{N_B}^*, \hat{B}_{N_B}^*, \hat{C}_{N_B}^*, \hat{D}_{N_B}^*\}$.

(4) From $\{\hat{B}_1^*, \hat{C}_1^*, \hat{D}_1^*\}, \{\hat{B}_2^*, \hat{C}_2^*, \hat{D}_2^*\}, \ldots, \{\hat{B}_{N_B}^*, \hat{C}_{N_B}^*, \hat{D}_{N_B}^*\}$, compute the corresponding bootstrap estimates of the parameter of interest, leading to $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_{N_B}^*$. For the given parameter θ, estimate the distribution of $\hat{\theta} - \theta$ by the empirical distribution of the values

$$\left\{\hat{\theta}_j^* - \hat{\theta} \colon j = 1, 2, \ldots, N_B\right\}.$$

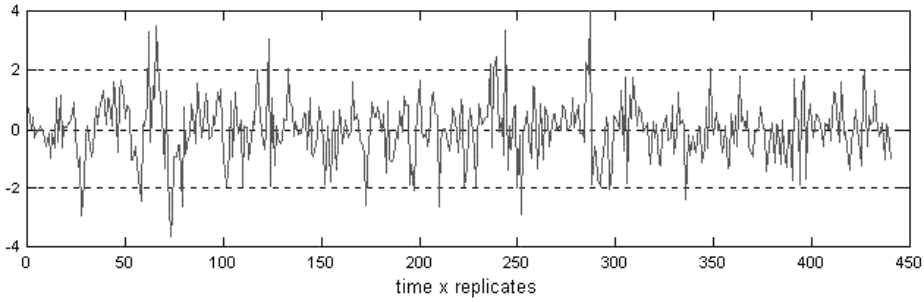

Fig. 4. Standardized innovations for a randomly selected gene.

Using quantiles of this latter empirical distribution to approximate corresponding quantiles of the distribution of $\hat{\theta} - \theta$, compute an estimated confidence interval on the parameter θ.

(5) Test the null hypothesis that the selected parameter is 0 by rejecting the null hypothesis if the confidence interval computed in step 4 does not contain the value 0.

(6) Repeat steps 4 and 5 for each element of $CB + D$. Elements for which zero is in between the upper and lower bounds will take the value zero. By setting the other non-zero entries to be 1, we obtain a network connectivity matrix in which 0s indicate the absence of a connection, and 1s indicate the presence of a connection.

The advantage of using the bootstrap procedure, instead of the asymptotic Gaussian distributions or approximations that would depend on the Gaussian assumptions for the SSM noise terms, is that bootstrapping is robust to deviations from the Gaussian assumption and can capture higher-order properties (e.g. skewness and kurtosis) that would not be estimated correctly in small samples by using the asymptotic Gaussian distributions.

Diagnostic checking

Diagnostic checking provides a means to assess how well the model represents the data (Durbin and Koopman, 2001). Diagnostics for fitting the SSM are based on estimated forecast errors, also called innovations. Innovations, v_t, represent the part of the observations, y_t, that cannot be predicted from the past.

Basic diagnostics on the innovations that examined correlation and distribution were performed. Innovation sequences should be approximately uncorrelated if the parameter estimates are accurate and the model fits well, so that standardized innovations should appear approximately as either white noise or iid with the identity matrix as the common covariance matrix. If, in addition, the innovations appear Gaussian, this would support the assumption that the noise sequences in the SSM are Gaussian. As pointed out earlier, however, the inferences drawn using the bootstrap analysis above are robust to deviations from the Gaussian assumption.

For the gene expression model described above, the innovations are given by

$$v_t = g_t - C\hat{x}_t^- - Dg_{t-1},$$

where $\hat{x}_t^-$ is the Kalman filter estimate of x_t, given the observations $g_1, g_2, \ldots, g_{t-1}$. The variance–covariance matrix of v_t is

$$\mathrm{Var}(v_t) = C\mathrm{Var}(x_t - \hat{x}_t^-)C' + R, \qquad (7)$$

which is not diagonal, indicating that there is correlation between the elements. The innovation components can be transformed in a way that they will become uncorrelated by applying the transformation suggested in Durbin and Koopman (2001), namely the inverse square root of the variance–covariance matrix (7). This gives the standardized innovations, which should appear as white noise with unit variance over both time and components. The innovations and their variance–covariance matrices can be estimated from the fitted SSM by substituting parameter estimates for C, D and R. The model will pass this test if these estimated standardized innovations appear to be consistent with white noise over all time and components. The plot in Figure 4 appears to show in a satisfactory way that the standardized innovations fluctuate without any apparent pattern. Additional plots are shown on the website containing the supplementary material. Histograms of the estimated innovations for some selected genes are plotted in Figure 5, and additional plots are shown on the website containing the supplementary material. The solid curve is an estimated Gaussian density in each case.

It turns out that in all cases, apart from occasional outliers, the distributions appear consistent with the Gaussian assumptions. The occasional outliers in the standardized innovations correspond to certain outlying replicates in the normalized gene expression profiles shown on the supplementary website. The Q–Q plots for selected genes shown in Figure 6 and in the supplementary website confirm that indeed the innovations are approximately Gaussian. However, we are mostly interested

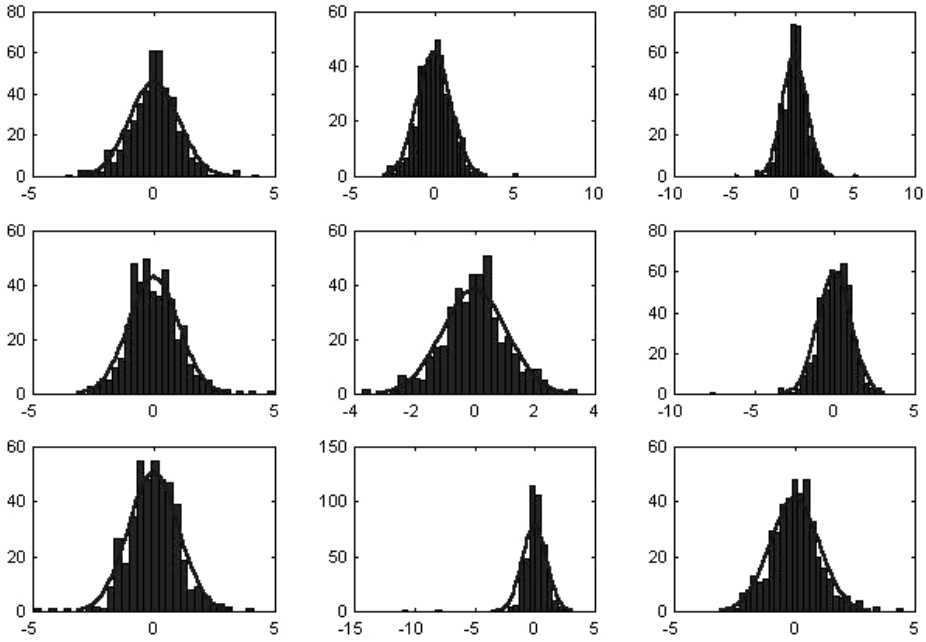

Fig. 5. Histograms of estimated innovations with a superimposed estimated density curve.

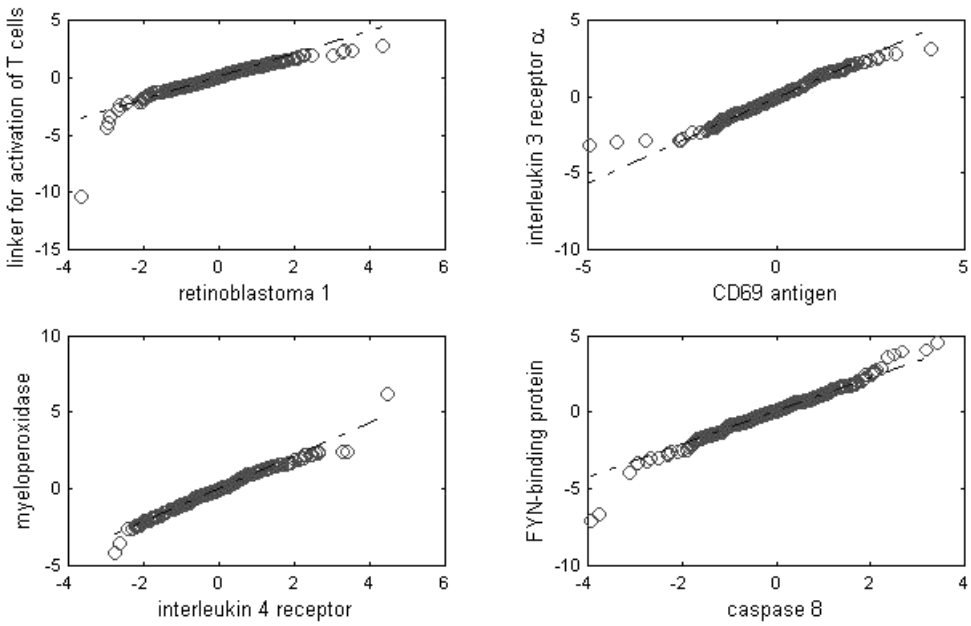

Fig. 6. Q–Q plot of ordered standardized innovations.

in verifying that the standardized innovations appear to show no pattern. Figure 4 seems consistent with this.

RESULTS

We applied the bootstrap procedure described in the Implementation section to identify 'high-probablity' gene–gene interaction networks that are shared by a significant number of sub-models built from randomly resampled data sets. In our procedure, we use bootstrap methods to find confidence intervals for the parameters defining the gene–gene interaction networks (i.e. the elements of $CB + D$), and so we can eliminate those that are not significantly different from zero. Thresholding the elements of the matrix $CB + D$ using

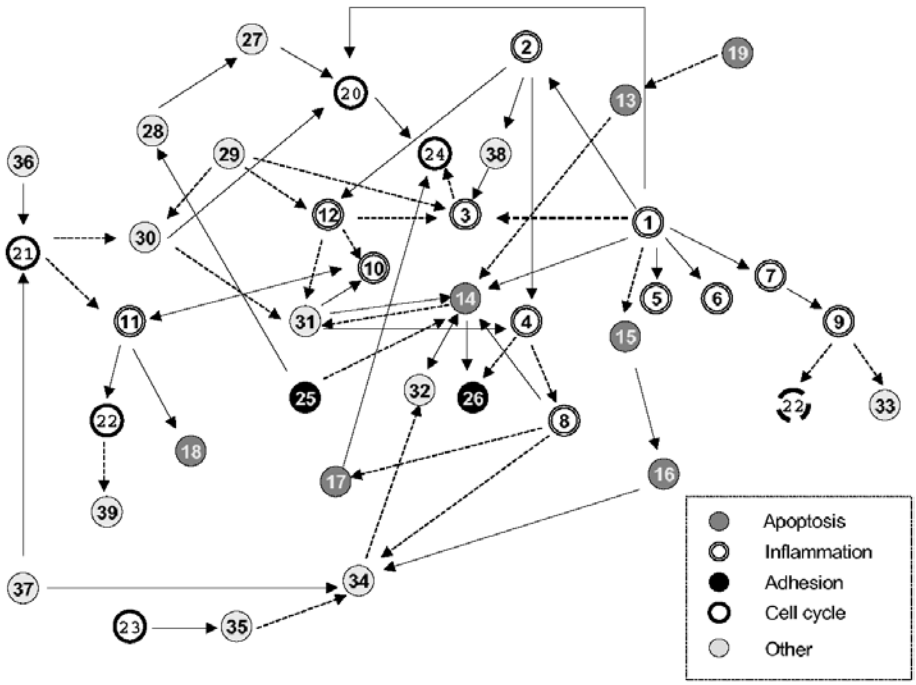

Fig. 7. Directed graph representing the elements of the $CB + D$ matrix. The main functional categories involved in T-lymphocyte response (cytokines, proliferation and apoptosis) are marked in different shades. Positive coefficients are represented by solid arrows; negative coefficients are represented by dotted arrows. Numbers refer to genes. The key to gene numbers is given on the supplementary website. Key genes mentioned in the discussion are FYB (gene 1), IL3Rα (gene 2), CD 69 (gene 3), TRAF5 (gene 4), IL4Rα (gene 5), GATA-binding protein 3 (gene 6), IL-2Rγ (gene 7), chemokine receptor CX3CR1 (gene 9), IL-16 (gene 11), Jun B (gene 13), caspase 8 (gene 14), clusterin (gene 15), caspase 7 (gene 18), survival of motoneuron 1 (gene 19), cyclin A2 (gene 20), CDC2 (gene 21), PCNA (gene 22), integrin alpha-M (gene 26) and MCL-1 (gene 31).

these confidence levels, we can obtain a connectivity matrix that describes all gene–gene interactions over successive time points. Our experiments in reconstucting networks from simulated data, generated from the gene expression model (5–6), indicate that, if it is desired to have a high percentage of overall correctness in the graph that is identified, then it is advisable to set the confidence level high on testing individual connections in a large, sparsely connected graph (Rangel *et al.*, 2004). The output from this procedure is a directed graph in which arrows are drawn from a gene expression variable at a given time t to another gene variable whose expression it influences at the next time point, $t + 1$. In addition, the non-zero entries in $CB + D$ represent the strength of the connection or the strength with which gene i influences gene j at consecutive time points. These values can be either positive or negative, indicating up- or down-regulation. The directed graph produced by this process with a confidence level on individual connections equal to 99.66% is shown in Figure 7.

DISCUSSION

Our analysis identifies a network of 39 genes out of the 58 that have interactions significant at the 99.66% confidence level. From a strictly topological point of view, the gene FYB (gene 1) occupies a crucial position in the graph since it has the highest number of outward connections. In order to interpret further the results of our analysis, we have mapped genes according to the main cellular functions modulated during T-cell activation (cytokine production, apoptosis, cell cycle and adhesion) and explored the network for evident functional groupings. Interestingly, the majority of the genes that are directly related to the inflammation response are directly connected to or located in close proximity to FYB (Fig. 7). These two observations fit well with the known role of FYB in T-cell activation. FYB is an important adaptor molecule in the T-cell receptor signaling machinery (Silva *et al.*, 1994) and is, therefore, very high in the hierarchy of events downstream of cell activation. Cells defective in this component have a severely impaired proliferation and migratory response and have reduced IL-2 secretion (Burack *et al.*, 2002). In our model, FYB influences the expression of eight genes. Of these, six have been reported as inducible in response to IL-2. These are the following: three interleukin receptor genes [IL-2Rγ (gene 7), IL4Rα (gene 5), IL3Rα (gene 2)], two apoptosis related genes [clusterin (gene 15) and caspase 8 (gene 14)] (Rosenberg and Silkensen, 1995),

a proliferation gene [cyclin A2 (gene 20)], an early T-cell activation marker [CD 69 (gene 3)] (Cambiaggi *et al*., 1992) and GATA-binding protein 3 (gene 6), a member of a GATA family of zinc-finger transcription factors involved in T-cell antigen regulation (Zheng and Flavell, 1997).

The three IL receptor genes encode for the IL-4 receptor (formed by the IL-4 receptor alpha subunit and by the promiscuous IL-2 receptor gamma signaling subunit), for the binding subunit of the IL-3 receptor and for the signaling subunit of the IL-2 receptor. The cytokines associated with these receptors all function as proliferation signals in T-cells. In particular, IL-2 is an antigen-unspecific proliferation factor that induces cell cycle progression in resting cells and thus allows clonal expansion of activated T-lymphocytes. Due to its effects on T-cells and B-cells, IL-2 is a central regulator of immune responses. IL-3 is also an important signal that controls the viability and the function of several hematopoietic cells (Ihle, 1992). IL-4 has additional roles in regulating antibody production, hematopoiesis and inflammation and the development of effector T-cell responses (Boulay and Paul, 1992). CD-69 is the earliest inducible cell surface glycoprotein acquired during lymphoid activation. It is involved in lymphocyte proliferation and functions as a signal-transmitting receptor in lymphocytes, natural killer (NK) cells and platelets (Testi *et al*., 1994).

In addition to the ability of regulating cytokine production, FYB also stimulates adhesion through direct interaction with the LFA-1 integrin (Peterson *et al*., 2001). In our model, FYB is connected to integrin alpha-M (gene 26) through IL3Rα (gene 2) and TRAF5 (gene 4), a gene activated by granulocyte-macrophage colony-stimulating factor (GM-CSF) and IL-3 signaling pathways. Although these connections do not reflect the direct post-transcriptional nature of the known FYB–integrin interaction, it is interesting and encouraging that our model implies that FYB mRNA levels are predictive of the level of expression of a member of a functionally and structurally related gene family of integrins (Corbi *et al*., 1988).

Other examples of genes with correlated functions that appear linked in our graph are survival of motoneuron 1 SMN1, (gene 19), Jun B (gene 13) and caspase 8 (gene 14). These genes are involved to different degrees in programmed cell death. In our model, the gene SMN1 influences negatively the expression of JunB, a pro-apoptotic gene (Weitzman, 2001). This fits well with the finding that SMN1 has been described as inhibiting the onset of apoptosis in PC12 cells by preventing cytochrome *c* release and caspase-3 activation (Vyas *et al*., 2002). A number of specific connections in the graph are supported by the published literature. The chemokine receptor CX3CR1 (gene 9) mediates both adhesive and migratory functions. It functions as a chemotactic receptor with the soluble form of Fractalkine and as an adhesion molecule with membrane-bound Fractalkine. The receptor is expressed in neutrophils, monocytes, T-lymphocytes, and in several solid organs. In our model, the gene encoding for this receptor is directly downstream of IL-2 receptor gamma (gene 7). This prediction is consistent with the finding that CX3CR1 is up-regulated in response to stimulation with IL-2 in a different cell type (Inngjerdingen *et al*., 2001). Our model also predicts IL-16 (gene 11) to be linked to two key cell cycle genes: PCNA (gene 22) and CDC2 (gene 21). IL-16 is a ligand and a chemotactic factor for CD4+ T cells. IL-16 is generally thought to inhibit CD3-mediated lymphocyte activation and proliferation. However, the effects of IL-16 on the target cells are dependent on the cell type and the presence of co-activators. Zhang and Xu (2002) tested the activity of IL-16 on Jurkat T leukemia cells and discovered that the IL-16 stimulated proliferation at low doses but inhibited the growth of the cells at higher concentrations. In accordance with our model IL-16 (gene 11) has been proven to directly activate caspase 7 (gene 18) (a key gene in the apoptotic pathway). In our model, the gene MCL-1 (gene 31) is downstream of the IL-3 receptor (gene 2). This is well supported by the finding that MCL-1 is an immediate-early gene activated by the GM-CSF and IL-3 signaling pathways (Wang *et al*., 1999).

In interpreting the model, we need to ask if increased levels of mRNA for a given gene are likely to result in a functional protein that is able to influence the transcription of downstream genes. Unless direct evidence exists, these interactions should not be interpreted as causal but rather representing direct or indirect mechanisms of action. In the case of FYB, it has been demonstrated that its over-expression results in a potentiation of T-cell receptor-mediated IL-2 production (Silva *et al*., 1994). A large proportion of the genes downstream of FYB in our graph are known targets of IL-2. This would suggest that the clustering of inflammation-related genes downstream of FYB (as predicted by our model) could be explained via an IL-2-dependent mechanism (Fig. 8). Is this interpretation realistic considering that we are stimulating lymphocytes with PMA and ionomycin? This treatment bypasses T-cell receptor stimulation and may not effectively trigger mechanisms involving FYB. From careful analysis of the data in the literature (Silva *et al*., 1994; Veale *et al*., 1999) it appears that PMA may be able to synergize with FYB in transfection experiments. Although the effect is small compared with combined T-cell receptor stimulation, the levels of IL-2 expression could be sufficiently high to induce a biological effect. We propose that during activation with PMA and ionomycin the level of IL-2 expression could be influenced by the available levels of the FYB protein. In agreement with the known function of FYB, our model also predicts the expression levels of FYB to influence the expression of cyclin A2. The protein encoded by this gene binds and activates CDC2 or CDK2 kinases and thus promotes both cell cycle G1/S and G2/M transitions (Faivre *et al*., 2001).

Interestingly, the expression levels of cyclin A2 and other cell cycle genes decrease in Jurkat cells after stimulation

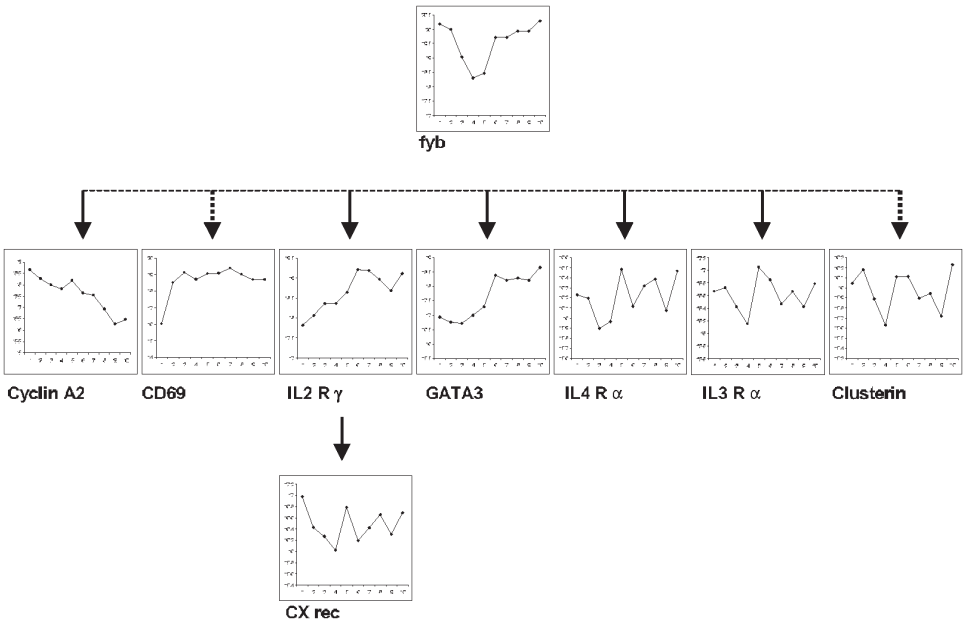

Fig. 8. Diagram representing genes downstream of FYB. Individual gene expression profiles are represented by plots of average expression profiles. Gene identities are reported alongside the plots. Positive coefficients are represented by solid arrows; negative coefficients are represented by dotted arrows.

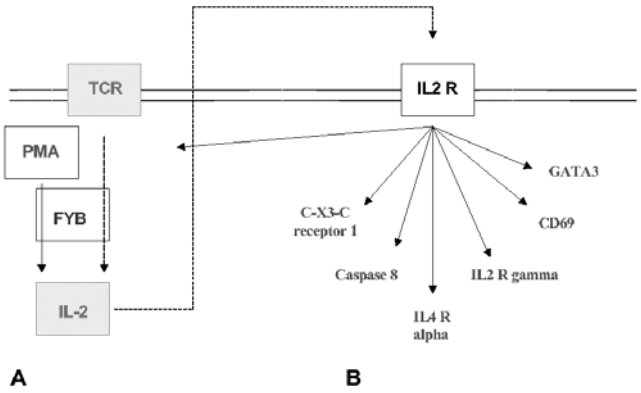

Fig. 9. FYB influences the activation of IL-2 target genes. The figure represents, in a schematic format, our interpretation of the predicted influence of FYB on the expression of IL-2 target genes. (**A**) The level of IL-2 expression increases in response to PMA and ionomycin stimulation and is influenced by the amount of FYB. (**B**) Once functional, IL-2 is secreted and binds its receptor so that target genes are activated. Since IL-2 was not included in the data set, the model could infer a direct link between FYB and the IL-2 target genes.

(Fig. 9). This unusual response to stimulation is one of the main differences between our biological model and primary CD4+ human T-lymphocytes. Unlike primary T-cells, Jurkat T-cells proliferate spontaneously, and PMA and ionomycin treatment will, in fact, result in reduced proliferation (due to cell cycle arrest and apoptosis). Despite these differences, the model has been used widely to study T-cell activation pathways.

This provides an excellent example of the type of hypothesis that can be generated using reverse engineering approaches. Obviously, interactions for which we do not find support in the current literature represent novel hypotheses. A detailed investigation of these predicted interactions is one focus of our current and future research since they provide an opportunity to validate experimentally or redefine the model. Despite the linear assumptions inherent in our SSMs, we have shown that our model reflects many of the dynamics of an activated T-cell. In particular, it reveals the integrated activation of cytokines, proliferation and adhesion following activation. However, further experimental work would be required to identify novel causal interactions. The application of this methodology to more physiological models (e.g. TCR-mediated activation of primary human T-lymphocytes) would be the logical next step.

Further improvements may also be made to the modeling procedure. Our experiments with simulated data (Rangel *et al.*, 2004) indicate that improved performance in the fidelity of network reconstruction should be obtained from experimental data sets containing more replicates and additional time points. We did not find a one-to-one correspondence between the nine hidden variables and known biological effects or unmeasured regulatory genes. This is not surprising, given that although the direct gene–gene interations (in the $CB + D$ matrix) are identifiable, the hidden variables are in general not identifiable. That is, two models can have equivalent gene–gene interactions but different implementations

of those in terms of hidden variables. The hidden variables were, however, important in practice since they played a large role in mediating the gene–gene interactions over time. In our model, the hidden variables are likely to represent a combination of complex molecular events (such as a combination of genes and possibly entire pathways) linking two genes. In this scenario, allowing hidden factors is an essential part of our overall goal of developing biologically realistic models. With larger data sets, we would also expect to be able to learn models with a larger number of hidden variables, which may then have a clearer biological interpretation.

Future work will include investigating Bayesian approaches to model selection using Markov chain Monte Carlo methods to sample from the full Bayesian posterior distributions of all unknown quantities. This approach will also allow us to examine the robustness of the inferences with respect to choices in the prior distribution over parameters and to study different choices for the priors. One attraction of this approach is that it is possible to incorporate priors in the form of known connections supported by the literature, including constraints with regard to the sign of the interaction (i.e. negative—inhibition or positive—activation). An alternative approach will explore the use of variational Bayesian methods for model selection. The theory of variational Bayesian learning has been successfully applied to learning non-trivial SSM structures in other application domains (Ghahramani and Beal, 2000, 2001), which suggests that it will provide good solutions in the case of modeling genetic regulatory networks, where one is typically working with data sets that are small compared with the number of parameters that need to be estimated. Our initial experiments with linear dynamics also pave the way for future work on models with non-linear dynamics.

ACKNOWLEDGEMENTS

The authors would like to thank Terry Speed (Berkeley) and Nathalie Thorne (Melbourne, Australia) for advice and code relating to quantile normalization, Nick Davies (Birmingham, UK) for helpful discussions and Brian Champion (Lorantis Ltd, UK) for his enthusiastic support of the project. C.R. acknowledges support from the Keck Graduate Institute of Applied Life Sciences.

REFERENCES

Akutsu,T., Miyano,S. and Kuhara,S. (1999) Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pac. Symp. Biocomput.*, 17–28.

Arkin,A., Shen,P. and Ross,J. (1997) A test case of correlation metric construction of a reaction pathway from measurements. *Science*, **277**, 1275–1279.

Bolstad,B., Irizarry,R., Astrand,M. and Speed,T. (2002) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185–193.

Boulay,J. and Paul,W. (1992) The interleukin-4 family of lymphokines. *Curr. Opin. Immunol.*, **4**, 294–298.

Brown,R.G. and Hwang,P.Y. (1997) *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley and Sons, New York.

Burack,W., Cheng,A. and Shaw,A. (2002) Scaffolds, adaptors and linkers of TCR signaling: theory and practice. *Curr. Opin. Immunol.*, **14**, 312–316.

Cambiaggi,C., Scupoli,M., Cestari,T., Gerosa,F., Carra,G., Tridente,G. and Accolla,R. (1992) Constitutive expression of CD69 in interspecies T-cell hybrids and locus assignment to human chromosome 12. *Immunogenetics*, **36**, 117–120.

Castagna,M., Takai,Y., Kaibuchi,K., Sano,K., Kikkawa,U. and Nishizuka,U. (1982) Direct activation of calcium-activated, phospholipid-dependent protein kinase by tumor promoting phorbol esters. *J. Biol. Chem.*, **257**, 7847–7851.

Cooper,G. and Herskovits,E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347.

Corbi,A., Larson,R., Kishimoto,T., Springer,T. and Morton,C. (1988) Chromosomal location of the genes encoding the leukocyte adhesion receptors lfa-1, mac-1 and p150,95: identification of a gene cluster involved in cell adhesion. *J. Exp. Med.*, **167**, 1597–1607.

Dempster,A., Laird,N. and Rubin,D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, **39**, 1–38.

D'Haeseleer,P., Wen,X., Fuhrman,S. and Somogyi,R. (1999) Linear modeling of mRNA expression levels during CNS development and injury. *Pac. Symp. Biocomput.*, **3**, 41–52.

Dopazo,J., Zanders,E., Dragoni,I., Amphlett,G. and Falciani,F. (2001) Methods and approaches in the analysis of gene expression data. *J. Immunol. Meth.*, **250**, 93–112.

Durbin,J. and Koopman,S. (2001) *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Faivre,J., Frank-Vaillant,M., Poulhe,R., Mouly,H., Brechot,C., Sobczak-Thepot,J. and Jessus,C. (2001) Membrane-anchored cyclin A2 triggers activation in *Xenopus* oocyte. *Bioinformatics*, **506**, 243–248.

Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Ghahramani,Z. and Beal,M. (2000) Variational inference for Bayesian mixture of factor analysers. *Adv. Neural Inform. Process. Syst.*, **12**, 449–455.

Ghahramani,Z. and Beal,M. (2001) Propagation algorithms for variational Bayesian learning. *Adv. Neural Inform. Process. Syst.*, **13**, 507–513.

Ghahramani,Z. and Hinton,G.E. (1996) Parameter estimation for linear dynamical systems. *Technical Report*, University of Toronto.

Hegde,P., Qi,R., Abernathy,K., Gay,C., Dharap,S., Gaspard,R., Hughes,J., Snesrud,E., Lee,N. and Quackenbush,J. (2000) A concise guide to cDNA microarray analysis. *Biotechniques*, **29**, 548–550, 552–554, 556 passim.

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

Ihle,J.N. (1992) Interleukin-3 and hematopoiesis. *Chem. Immunol.*, **51**, 65–106.

Inngjerdingen,M., Damaj,B. and Maghazachi,A. (2001) Expression and regulation of chemokine receptors in human natural killer cells. *Blood*, **97**, 367–375.

Iwashima,M. (2003) Kinetic prospectives of T cell antigen receptor signaling. *Immunol. Rev.*, **191**, 196–210.

Iwashima,M., Irving,B., Oers,N.V., Chan,A.C. and Weiss,A. (1994) Sequential interactions of the TCR with two distinct cytoplasmic tyrosine kinases. *Science*, **263**, 1136–1139.

Kholodenko,B., Kiyatkin,A., Bruggeman,F., Sontag,E., Westerhoff,H. and Hoek,J. (2002) Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc. Natl Acad. Sci., USA*, **99**, 12841–12846.

Ley,S., Davies,A., Druker,B. and Crumpton,M. (1991) The T cell receptor/CD3 complex and CD2 stimulate the tyrosine phosphorylation of indistinguishable patterns of polypeptides in the human T leukemic cell line jurkat. *Eur. J. Immunol.*, **21**, 2203–2209.

Liang,S., Fuhrman,S. and Somogyi,R. (1998) Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pac. Symp. Biocomput.*, 18–29.

Manger,B., Weiss,A., Imboden,J., Laing,T. and Stobo,J. (1987) The role of protein kinase C in transmembrane signalling by the T cell antigen receptor complex: effect of stimulation with soluble or immobilized CD3 antibodies. *J. Immunol.*, **139**, 2755–2760.

Murphy,K. and Mian,S. (1999) Modelling gene expression data using Dynamic Bayesian Networks. *Technical Report*, University of California, Berkeley.

Ong,I., Glasner,J. and Page,D. (2002) Modelling regulatory pathways in *E.coli* from time series expression profiles. *Bioinformatics*, **18**, S241–S248.

Pe'er,D., Regev,A., Elidan,G. and Friedman,N. (2001) Inferring subnetworks from perturbed expression profiles. *Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology* (*ISMB*), July 21–25, Copenhagen.

Peterson,E., Woods,M., Dmowski,S., Derimanov,G., Jordan,M., Wu,J., Myung,P., Liu,Q., Pribila,J., Freedman,B., Shimizu,Y. and Koretzky,G. (2001) Coupling of the TCR to integrin activation by slap-130/fyb. *Science*, **293**, 2263–2265.

Rangel,C., Angus,J., Ghahramani,Z. and Wild,D.L. (2004) Modeling genetic regulatory networks using gene expression profiling and state space models. In Husmeier,D., Roberts,S. and Dybowski,R. (eds), *Probabilistic Modelling in Bioinformatics and Medical Informatics*. Springer-Verlag (in press).

Rangel,C., Wild,D.L., Falciani,F., Ghahramani,Z. and Gaiba,A. (2001) Modelling biological responses using gene expression profiling and linear dynamical systems. *Proceedings of the 2nd International Conference on Systems Biology*, Omipress, Madison, WI, pp. 248–256.

Rosenberg,M. and Silkensen,J. (1995) Clusterin: physiologic and pathophysiologic considerations. *Int. J. Biochem. Cell Biol.*, **27**, 633–645.

Roweis,S. and Ghahramani,Z. (1999) A unifying review of linear Gaussian models. *Neural Comput.*, **11**, 305–345.

Shumway,R. and Stoffer,D. (1982) An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Analysis*, **3**, 253–264.

Silva,A.J.D., Zhuwen,L., Vera,C.D., Findell,C.E.P. and Rudd,C.E. (1994) Cloning of a novel T-cell protein FYB that binds FYN and SH2-domain-containing leukocyte protein 76 and modulates interleukin 2 production. *Proc. Natl Acad. Sci., USA*, **94**, 7493–7498.

Testi,R., D'Ambrosio,D., Maria,R.D. and Santoni,A. (1994) The cd69 receptor: a multipurpose cell-surface trigger for hematopoietic cells. *Immunol. Today*, **15**, 479.

Thomas,R. (1973) Boolean formalization of genetic control circuits. *J. Theor. Biol.*, **42**, 563–586.

van Someren,L.F., Wessels,E. and Reinders,M. (2000) Linear modeling of genetic networks from experimental data. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology* (*ISMB*). pp. 355–366, August 16–23, La Jolla.

van Someren,E., Wessels,L., Backer,E. and Reinders,M. (2002) Genetic network modeling. *Pharmacogenomics*, **3**, 507–525.

Veale,M., Raab,M., Li,Z., da Silva,A.J., Kraefti,S.-K., Weremowicz,S., Morton,C.C. and Rudd,C.E. (1999) Novel isoform of lymphoid adaptor FYN-T-binding protein (FYB-130) interacts with slp-76 and up-regulates interleukin 2 production. *J. Biol. Chem.*, **274**, 28427–28435.

Vyas,S., Bechade,C., Riveau,B., Downward,J. and Triller,A. (2002) Involvement of survival motor neuron (SMN) protein in cell death. *Hum. Mol. Genet.*, **11**, 2751–2764.

Wang,J.-M., Chao,J.-R., Chen,W., Kuo,M.-L., Yen,J. and Yang-Yen,H.-F. (1999) The antiapoptotic gene mcl-1 is up-regulated by the phosphatidylinositol 3-kinase AKT signaling pathway through a transcription factor complex containing CREB. *Mol. Cell. Biol.*, **19**, 6195–6206.

Weaver,D., Workman,C. and Stormo,G. (1999) Modeling regulatory networks with weight matrices. *Pac. Symp. Biocomput.*, **4**, 112–123.

Weitzman,J. (2001) Life and death in the jungle. *Trends Mol. Med.*, **7**, 141–142.

Yang,Y., Dudoit,S., Luu,P., Lin,D., Peng,V., Ngai,J. and Speed,T. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Yoo,C., Thorsson,V. and Cooper,G. (2002) Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. *Pac. Symp. Biocomput.*, 422–433.

Zhang,X.M. and Xu,Y.H. (2002) The associated regulators and signal in r-IL-16/CD4 mediated growth regulation in Jurkat cells. *Cell Res.*, **12**, 363–372.

Zheng,W. and Flavell,R.A. (1997) The transcription factor gata-3 is necessary and sufficient for th2 cytokine gene expression in CD4 T cells. *Cell*, **89**, 587–596.