

# Modelling biological responses using gene expression profiling and linear dynamical systems

Claudia Rangel  
Claremont Graduate  
University  
171 E. Tenth Street,  
Claremont, CA 91711, USA  
claudia.rangel@cgu.edu

David L. Wild<sup>\*</sup>  
Keck Graduate Institute  
535 Watson Drive  
Claremont, CA 91711, USA  
david.wild@kgi.edu

Francesco Falciani<sup>†</sup>  
Lorantis Limited  
Babraham Hall, Babraham  
Cambridge, CB2 4UL, UK  
francesco.falciani@lorantis.co.uk

## ABSTRACT

Linear dynamical systems are a subclass of dynamic Bayesian networks used for modeling time series data which assume the existence of a hidden state variable, from which we can make noisy measurements, which evolves with Markovian dynamics. We have applied these methods to the analysis highly replicated gene expression microarray time series data with the intention of building testable hypotheses about the causal influences between gene expression events involved in the activation of human T cells.

## 1. INTRODUCTION

The application of high-density DNA microarray technology to gene transcription analyses 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 the human 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 towards 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. Tools which have been applied to this problem include Boolean networks [1, 21, 29] and Bayesian networks [12, 15].

Our aim is to identify transcriptional networks based on large volumes of gene expression data. In order to test the validity of our approach we have chosen a well-established model of T cell activation and monitored a set of relevant genes across a time series. We have then applied linear dynamical systems modeling to reverse engineer genetic networks from expression profiling data. Linear Dynamical Systems (LDS) (also known as linear-Gaussian state-space models [25] or Kalman filter models [8]) are a subclass of dynamic Bayesian networks used for modeling time series data and have been used extensively in many areas of control and signal processing. LDS models have a number of features which make them attractive for modeling gene ex-

pression time series data. They assume the existence of a hidden state variable, from which we can make noisy 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 which cannot be measured in a gene expression profiling experiment, for example: the effects of genes which have not been included on the microarray, levels regulatory proteins, the effects of mRNA degradation etc. Our LDS models have produced testable hypotheses, which have the potential for rapid experimental validation.

### 1.1 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 signalling molecules, including kinases, phosphatases, and adaptor proteins that couple the stimulatory signal received from T cell receptor (TCR) to gene transcription events in the nucleus [18, 20]. The early signalling event most proximal to the TCR/CD3 complex is the activation of a number of protein tyrosine kinases (PTKs): Lck, Fyn, Yes [4, 27] and ZAP-70 [5, 13]. Activation of PTKs induces localization of adaptor proteins to the cytoplasmic membrane with consequent activation of the Ras/MAPK pathway that eventually transmits the stimulatory signal to the nucleus. Stimulation of PTKs is also coupled to the hydrolysis of PLC $\gamma$ 1, which results in a rise in intracellular Ca<sup>2+</sup>, and activation of PKC through IP<sub>3</sub> and DAG [7, 17]. The increase in intracellular level of free Ca<sup>2+</sup> and the activation of protein kinase C (PKC) are critical for TCR mediated T cell activation.

In this paper we describe the application of linear dynamical systems modeling to infer causal relationships between genes involved 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 with the calcium ionophore ionomycin and the PKC activator phorbol ester PMA [22].

<sup>\*</sup>Corresponding author

<sup>†</sup>Corresponding author - additional authors listed in section 6

This treatment bypasses the TCR requirement and thereby activates signalling transduction pathways [9, 30] leading to T cell activation. Although this is a well established model, there are clearly important differences between the activation of Jurkat cells and primary human T cells. The combined effect of PMA and ionomycin in Jurkat cells induces the expected activation markers (CD69, NF $\kappa$ B, etc) although the effect on proliferation is different. 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 widely used to study T-cell activation pathways.

## 1.2 Linear Dynamical Systems (State-Space Models)

In linear-Gaussian state-space models, a sequence of  $p$ -dimensional real-valued observation vectors  $\{y_1, \dots, y_T\}$ , is modeled by assuming that at each time step  $y_t$  was generated from a  $K$ -dimensional real-valued hidden state variable  $x_t$ , and that the sequence of  $x_t$ s define a first-order Markov process. Assuming that the initial state  $x_1$  and the noise vectors are independent and Gaussian distributed then the output of the system is also Gaussian. That is, all future hidden states  $x_t$  and observations  $y_t$  generated from those hidden states will be Gaussian distributed. Linear-Gaussian state space models 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 transition state matrix,  $C$  is the state to observation matrix and  $w_t$  and  $v_t$  are independent zero-mean random noise vectors.

## 1.3 State-Space Model with Inputs

Often, the observations can be divided into a set of input (or predictor) variables and a set of output (or response) variables. The equations describing the linear-Gaussian state space model then become:

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

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

where  $u_t$  is the input observation vector,  $A$  is the transition state matrix,  $B$  is the input to state matrix in the state transition equation (3),  $C$  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 vectors,  $w_t$  and  $v_t$  respectively, are random variables assumed to be Gaussian distributed, independent of each other and independent of the initial values of  $x$  and  $y$ . Noise vectors are also hidden variables. The conditional expectation of the states and the observables are given by:

$$P(x_{t+1}|x_t, u_t) \sim N(Ax_t + Bu_t, Q) \quad (5)$$

$$P(y_t|x_t, u_t) \sim N(Cx_t + Du_t, R) \quad (6)$$

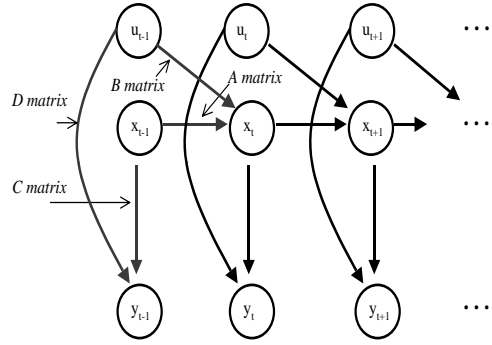


Figure 1: LDS model with inputs

where  $Q$  and  $R$  are the state and observation noise covariances respectively. Hence we have that

$$P(y_t|x_t, u_t) = \frac{\exp\{-\frac{1}{2}(y_t - Cx_t - Du_t)'R^{-1}(y_t - Cx_t - Du_t)\}}{(2\pi)^{p/2}|R|^{1/2}} \quad (7)$$

The Markov properties of the model and the Gaussian assumptions about the noise and initial distributions define the joint probability of a sequence of T states and outputs as follows:

$$P(\{x\}, \{y\}) = P(x_1) \prod_{t=1}^{T-1} P(x_{t+1}|x_t, u_t) \prod_{t=1}^T P(y_t|x_t, u_t) \quad (8)$$

The unknown parameters of the LDS model may be learnt from data using the Expectation-Maximization (EM) algorithm [3, 14] as described in the following section.

## 1.4 The EM Algorithm when $x$ is Hidden

The Expectation-Maximization (EM) algorithm for learning a linear dynamical system finds the maximum likelihood estimate  $\hat{\theta}$  of the parameters given a data set composed by observed data  $y$  and unobserved/missing/hidden data  $x$ . EM uses the solutions to the filtering and smoothing problem to estimate the unknown hidden states given the observations and the current model parameters. It then uses this "fictitious" complete data to solve for new model parameters. Note that the state estimate  $\hat{x}_t$ , differs from one computed in a Kalman filter in that it depends on past and future observations. The two sets of variables have joint distribution depending on  $\theta$ ,

$$P(x, y|\theta) = P(y|x, \theta)P(x|\theta) \quad (9)$$

where  $P(x|\theta)$  is the marginal density function of  $x$  given  $\theta$  and in  $P(y|x, \theta)$  we consider the values of  $x$  and  $\theta$  as guesses. Thus the likelihood function is given by

$$\mathcal{L}(\theta; x, y) \equiv P(x, y|\theta) \quad (10)$$

The idea of the maximum likelihood problem is to find the  $\theta$  that maximizes  $\mathcal{L}$ , that is we want to find  $\hat{\theta} = \arg_{\theta} \max \mathcal{L}(\theta; x, y)$ . If  $x$  is hidden, we maximize the marginal likelihood (integrating out  $x$ ). Maximizing the likelihood as a function of  $\theta$  is equivalent to maximizing the log-likelihood. Suppose the hidden variables have an arbitrary distribution  $Q$ , we can obtain a lower bound on the log marginal likelihood:

$$\begin{aligned}
\log \int_x P(x, y|\theta) dx &= \log \int_x Q(x) \left[ \frac{P(x, y|\theta)}{Q(x)} \right] dx \\
&= \log E_{x \sim Q} \left[ \frac{P(x, y|\theta)}{Q(x)} \right] \\
&\geq E_{x \sim Q} \log \left[ \frac{P(x, y|\theta)}{Q(x)} \right] \\
&= \int_x \log \left[ \frac{P(x, y|\theta)}{Q(x)} \right] Q(x) dx \\
&= \int_x Q(x) \log P(x, y|\theta) dx \\
&\quad - \int_x \log(Q(x)) Q(x) dx \quad (11) \\
&= \mathcal{F}(Q, \theta) \quad (12)
\end{aligned}$$

where the inequality in the third line follows from Jensen's inequality<sup>1</sup>

The EM algorithm alternates between maximizing  $\mathcal{F}$  with respect to  $Q$  and  $\theta$ , respectively, holding the other fixed.

$$\text{E step : } Q_{k+1} \leftarrow \arg_Q \max \mathcal{F}(Q, \theta_k) \quad (13)$$

$$\text{M step : } \theta_{k+1} \leftarrow \arg_{\theta} \max \mathcal{F}(Q_{k+1}, \theta) \quad (14)$$

In the E step, the  $Q_{k+1}$  that maximizes  $\mathcal{F}(Q, \theta)$  is  $Q_{k+1}(x) = P(x|y, \theta)$

$$\begin{aligned}
\mathcal{F}(Q, \theta) &= \int_x \log \left[ \frac{P(x, y|\theta)}{Q(x)} \right] Q(x) dx \\
&= - \int_x Q(x) \log \left[ \frac{Q(x)}{P(x, y|\theta)} \right] dx \\
&= - \int_x \left( \log \left[ \frac{P(x, y|\theta)}{Q(x)} \right] + \log(P(y|\theta)) \right) Q(x) dx
\end{aligned}$$

This last term holds by (9). Let  $H(Q, P)$  be the relative entropy of  $Q$  with respect to  $P$ . Recalling that the relative entropy of two probability distributions measures in some sense the dissimilarity between them [11] then we have,

$$\mathcal{F}(Q, \theta) = -H(Q, P(x|y, \theta)) + \log(P(y|\theta)) \quad (15)$$

and the maximum occurs when  $Q(x) = P(x|y, \theta)$  by the relative entropy result. This means that the E-step must lead to  $Q_{k+1}(x) = P(x|y, \theta)$ . Therefore, for this choice of  $Q_{k+1}$ ,  $\mathcal{F}(Q, \theta)$  is the likelihood of  $\theta$ .

## 1.5 EM Applied to LDS with Inputs

### 1.5.1 E step:

The maximization of the expectation of the joint probability in (8) is equivalent to the minimization of the expectation

<sup>1</sup>Jensen's inequality: If  $f$  is convex, then  $f(E(x)) \geq E(f(x))$

of:

$$\begin{aligned}
-2 \log P(\{x\}, \{y\}) &= -2 \log P(x_1) \quad (16) \\
&+ \sum_{t=1}^{T-1} (x_{t+1} - Ax_t - Bu_t)' Q^{-1} (x_{t+1} - Ax_t - Bu_t) \\
&+ \sum_{t=1}^T (y_t - Cx_t - Du_t)' R^{-1} (y_t - Cx_t - Du_t) \\
&+ (T-1) \log |Q| + T \log |R| + Tp \log(2\pi) \\
&+ (t-1)K \log(2\pi) \quad (17)
\end{aligned}$$

If  $P(x_1)$  is Gaussian  $(\mu_1, Q_1)$ , then (16) equals

$$\begin{aligned}
-2 \log P(\{x\}, \{y\}) &= \sum_{t=1}^{T-1} (x_{t+1} - Ax_t - Bu_t)' \times \\
&\quad Q^{-1} (x_{t+1} - Ax_t - Bu_t) \\
&+ \sum_{t=1}^T (y_t - Cx_t - Du_t)' R^{-1} (y_t - Cx_t - Du_t) \\
&+ (x_1 - \mu_1)' Q^{-1} (x_1 - \mu_1) + (T-1) \log |Q| \\
&+ T \log |R| + T(p-K) \log(2\pi) + \log |Q_1| \quad (18)
\end{aligned}$$

For linear Gaussian models, the EM learning procedure involves minimizing quadratic forms as above, and this can be done by linear regression. This process is repeated using these new model parameters to infer the hidden states again, and so on.

### 1.5.2 M step:

Here, we solve for A,B,C,D,Q, and R. Each of these is re-estimated by taking the corresponding partial derivative of the expected log likelihood, setting to zero, and solving. The maximum in the M step is obtained by maximizing the first term in (11), since the entropy of  $Q$  does not depend on  $\theta$ :

$$\text{M step : } \theta_{k+1} \leftarrow \arg_{\theta} \max \int_x P(x|y, \theta_k) \log P(x, y|\theta) dx \quad (19)$$

Since  $\mathcal{F}$  is equal to the log marginal likelihood at the beginning of each M step, and since the E step does not change  $\theta$ , we are guaranteed not to decrease the likelihood after each combined EM step. The ML parameters could be solved by minimizing the expectation of (16) if all the random variables were observed. Since the states are in fact hidden we use expected values wherever we don't have access to the actual observed values. For instance, if we want to solve for the matrix C we set the partial derivative of the expectation of (16) equal to zero and solve. Since only the third term in (16) depends on C, we only need to solve for:

$$\begin{aligned}
&\frac{\partial [-2 \log P(\{x\}, \{y\})]}{\partial C} \\
&= \frac{\partial \left[ \sum_{t=1}^T (y_t - Cx_t - Du_t)' R^{-1} (y_t - Cx_t - Du_t) \right]}{\partial C} \\
&= \sum_{t=1}^T y_t x_t' - C \sum_{t=1}^T x_t x_t' - D \sum_{t=1}^T u_t x_t' = 0 \quad (20)
\end{aligned}$$

Taking expectations with respect to the hidden states, the M step for C is:

$$C = \left( \sum_{t=1}^T y_t \hat{x}_t' - D \sum_{t=1}^T u_t \hat{x}_t' \right) \left( \sum_{t=1}^T x_t \hat{x}_t' \right)^{-1} \quad (21)$$

In general, we require all terms of the kind:

$$\hat{x}_t = E[x_t | \{y\}] \quad (22)$$

$$P_t = E[x_t x_t' | \{y\}] \quad (23)$$

$$P_{t+1,t} = E[x_t x_{t+1}' | \{y\}] \quad (24)$$

These terms can be computed using the Kalman smoothing algorithm.

## 1.6 Kalman Smoothing

The Kalman smoother solves the problem of estimating the state at time  $t$  of a linear-Gaussian state-space model given the model parameters and a sequence of observations  $\{y_1, \dots, y_T\}$ . It consists of two parts: a forward recursion which uses the observations from  $y_1$  to  $y_T$ , known as the Kalman filter, and a backward recursion which uses the observations from  $y_T$  to  $y_{t+1}$ . The forward and backward recursions together are also known as the Rauch-Tung-Streifel (RTS) smoother. Treatments of Kalman filtering and smoothing can be found in [8].

The Gaussian marginal density of the hidden state vector is completely specified by its mean and covariance matrix. It is useful to define the quantities  $x_t^T$  and  $V_t^T$  as the mean vector and covariance matrix of  $x_t$ , respectively, given observations  $\{y_1, \dots, y_T\}$ . The Kalman filter consists of the following forward recursions:

$$x_t^{t-1} = Ax_{t-1}^{t-1} + Bu_t^{t-1} \quad (25)$$

$$V_t^{t-1} = AV_{t-1}^{t-1}A' + Q \quad (26)$$

$$K_t = V_t^{t-1}C'(CV_t^{t-1}C' + R)^{-1} \quad (27)$$

$$x_t^t = x_t^{t-1} + K_t(y_t - Cx_t^{t-1} - Du_t^{t-1}) \quad (28)$$

$$V_t^t = (I - K_tC)V_t^{t-1} \quad (29)$$

where  $x_1^0$  and  $V_1^0$  are the prior mean and covariance of the state, which are model parameters. Equations (25) and (26) describe the forward propagation of the state mean and variance before having accounted for the observation at time  $t$ . The mean evolves according to the known dynamics  $A$  which also affects the variance. In addition the variance also increases by  $Q$ , the state noise. The observation  $y_t$  has the effect of shifting the mean by an amount proportional to the prediction error  $y_t - Cx_t^{t-1} - Du_t^{t-1}$ , where the proportionality term  $K_t$  is known as the Kalman gain matrix. Observing  $y_t$  also has the effect of reducing the variance of  $x_t$ . These equations can all be derived (perhaps laboriously) by analytically evaluating the Gaussian integrals that result when belief propagation is applied to the Bayesian network corresponding to state-space models.

At the end of the forward recursions we have the values for  $x_T^T$  and  $V_T^T$ . We now need to proceed backwards and evaluate the influence of future observations on our estimate

of states in the past:

$$J_{t-1} = V_{t-1}^{t-1}A(V_{t-1}^{t-1})^{-1} \quad (30)$$

$$x_{t-1}^T = x_{t-1}^{t-1} + J_{t-1}(x_t^T - Ax_{t-1}^{t-1}) \quad (31)$$

$$V_{t-1}^T = V_{t-1}^{t-1} + J_{t-1}(V_t^T - V_{t-1}^{t-1})J_{t-1}' \quad (32)$$

where  $J_t$  is a gain matrix with a similar role to the Kalman gain matrix. Again, equation (31) shifts the mean by an amount proportional to the prediction error  $x_t^T - Ax_{t-1}^{t-1}$ . We can also recursively compute the covariance across two time steps

$$V_{t,t-1}^T = V_t^T J_{t-1}' + J_t(V_{t+1,t}^T - AV_t^T)J_{t-1}' \quad (33)$$

which is initialized

$$V_{T,T-1}^T = (I - K_T C)AV_{T-1}^{T-1} \quad (34)$$

The expectations (22),(23),(24) required for EM can now be readily computed

## 2. METHODS

### 2.1 Cell culture, treatments and RNA extraction

Jurkat cells were cultured in RPMI 1640 (GibcoBRL) supplemented with 2mM Glutamine (GibcoBRL), Penicillin-Streptomycin 50 units/ml(GibcoBRL) and with 10% Fetal Bovine Serum (FBS)(Biocrom KG). When the culture reached the density of  $10^6$  cells/ml cells were treated with 50ng/ml of Phorbol ester PMA (Sigma) plus 1 $\mu$ g/ml of ionomycin (Sigma). Cells were collected in 300  $\mu$ l of RTL lysing solution (Quiagen) at the following times after treatment (0, 2, 4, 6, 8, 18, 24, 32, 48, 72 hours). In order to ensure the efficacy of the stimulation, cells were tested for the correct expression of T cell and activation markers using FACS analysis. The cells used in this experiment were all expressing the T cell receptor (detected with anti CD3 antibodies) and after 24 hours of stimulation strongly upregulate CD69, an early surface activation marker. RNA was then extracted using RNA easy miniprep kit (Quiagen) according to the manufacturer's instructions.

### 2.2 Gene expression profiling

Microarrays were manufactured by spotting purified PCR products on amino-modified glass slides [16] using a Microgrid II spotter (Biorobotics, Cambridge, UK). Every PCR product was spotted 34 times on every array. Microarray probes were prepared by labelling 40 micrograms of total RNA by a reverse transcriptase reaction incorporating dCTP-Cy3 labelled nucleotide. Probe labelling and purification was then performed as described in [16]. Purified probes were then hybridised on the arrays for two days at 42°C in a 25% formaldehyde, 5X SSC, 0.1% SDS solution. Slides were washed twice in 2X SSC, 0.2 SDS for 5 minutes at 42°C, twice in 0.1X SSC, 0.2 SDS 5 minutes at room temperature and finally once 2X SSC, 0.2 SDS for 5 minutes at room temperature. Once dried, the slides were scanned on a GSI Lumonics confocal scanner at 100% laser power and 70% photo-multiplier 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 (GSI Lumonics). Numeric values for the gene expression intensities were

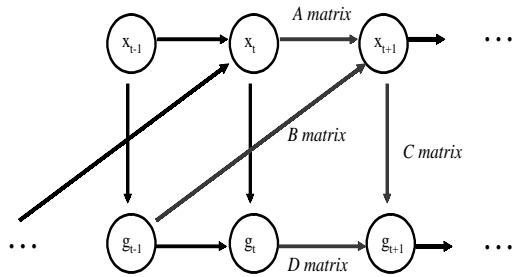


Figure 2: The LDS model for gene expression

calculated using the histogram method implemented in the same application. Values were calculated as integrals of the pixel signal distribution associated to each spot and local background values subtracted.

### 2.3 Data pre-processing

Genes whose expression values in all the time points were below 20000 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 filtering the non-expressed genes, 88 genes were left as training data for the LDS modeling algorithms. Gene identities were associated to spot coordinates using an Access database.

### 2.4 The LDS model for gene expression

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 LDS model with inputs described in the previous section as follows. We let the observations  $y_t^i = g_t^i$ , the expression level of gene  $i$  at time point  $t$ , and the inputs  $u_t = g_{t-1}$  to give the model shown in Figure 2. This model is described by the following equations:

$$x_{t+1} = Ax_t + Bg_{t-1} + w_t \quad (35)$$

$$g_t = Cx_t + Dg_{t-1} + v_t \quad (36)$$

Here matrix D captures gene-gene expression level influences at consecutive time points whilst the matrix C captures the influence of the hidden variables on gene expression level at each time point. This LDS model was trained, using EM learning, on the data described above, a ten point time series for each of 88 genes, replicated 34 times. Prior to training, the data were normalised so that each time series had a mean expression level of 1. The optimal number of

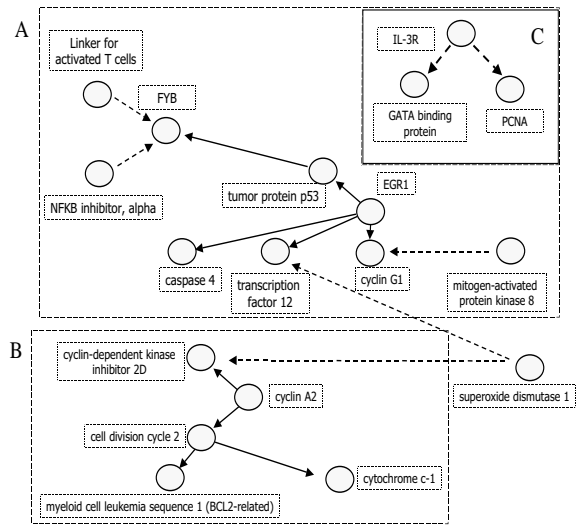


Figure 3: Directed acyclic graph representing the elements of the D matrix  $D_{ij}$  with Z-scores  $> 3$  and connectivity  $\geq 2$ . Panels A and panel B are functional sub-elements of a 15 gene network. Panel A represents the influence of the Early Growth Response gene 1 (EGR-1) on apoptotic and cell cycle genes. Panel B represents a network of proliferation related genes and Panel C represents an isolated 3 gene network centering on the IL-3 receptor gene. Positive weights are shown as solid arrows, negative weights as dotted lines

hidden states for this amount of data was chosen by comparing validation set likelihoods in a 34-way cross-validation experiment and was found to be 2. To evaluate the robustness of our model, we utilized the bootstrap method [10] to generate 300 “perturbed” versions of our data set and learnt a separate model from each. For each of the 300 bootstrap samples we generated a “new” data set of the same size as the original data set by re-sampling with replacement 34 instances of the time series for each of the 88 genes. From the resulting 300 models we calculate the mean and standard deviation of the elements of the D matrix in equation (36) and convert these to Z-scores (mean/standard deviation). By considering only those values of the D matrix ( $D_{ij}$ ) with Z-scores  $> 3$  we obtain an 88x88 adjacency matrix which gives us an estimate of the gene-gene influences (or interactions) which are reproduced with high confidence in the models learnt from the 300 bootstrap samples, together with a numerical value for the coefficients  $D_{ij}$  which expresses the “weight” and “direction” (positive or negative) of the influence of the expression level of one gene on another at a subsequent time point. Part of this adjacency matrix is shown as a directed acyclic graph (DAG) in Figure 3. Arrows indicate the direction of the influence, whilst positive coefficients  $D_{ij}$  are shown as solid lines and negative coefficients as dotted lines. Figure 4 shows the top subgraph with representative expression profiles for the genes involved.

Matrix C comprises 2 88-dimensional vectors whose ele-

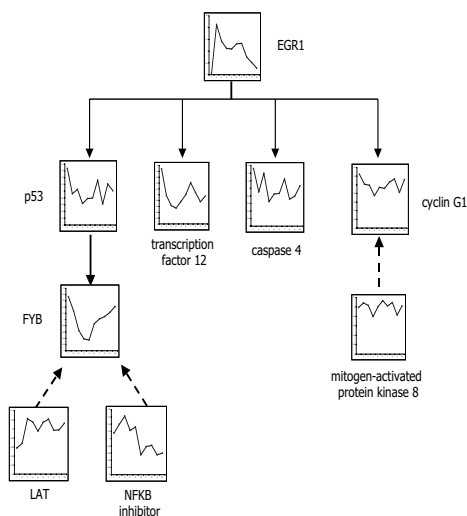


Figure 4: Diagram representing the elements of the D matrix  $D_{ij}$  with Z-scores  $> 3$  that are highlighted in panel A of Figure 3. Individual representative gene expression profiles are shown in a plot alongside the gene names. Positive weights are represented with solid arrows whereas negative weights are represented by dotted lines

ments represent the influence of the hidden variables on gene expression at each time step. We calculate an average value for matrix C over our 300 bootstrap models as follows: we normalise to unit vectors by dividing by the vector norm and calculate the average of  $CC^T$  over our 300 models. The 2 eigenvalues of the resulting matrix are plotted in Figure 4. We note that the first hidden state (upper plot in Figure 4) appears to effect most genes in the same “direction”, with the exception of about 10 genes on which it has either no effect or an effect in the opposite direction (eg: genes 43 and 49), whereas the second hidden state (lower plot in Figure 4) effects genes differentially. We do not currently have a biological explanation for what these hidden variables are modeling but speculate that they are capturing the effects of a complex mixture of factors. However, we note that Miskin [23] observed a similar pattern to our first hidden state in one of the latent variables in an independent components analysis model of static microarray data, which was dubbed a “housekeeping gene”. By applying the Kalman smoother algorithm described above to each of the 300 bootstrap models we may investigate the dynamics of the hidden variables. This work is currently in progress.

### 3. RESULTS AND DISCUSSION

#### 3.1 Single gene-gene interactions

Table 1 lists individual gene-gene “interactions” that the DAG produced from the D-matrix with Z-scores  $> 3$  has revealed. The column labelled “Influence” indicates whether the weight  $D_{ij}$  is positive or negative. For many of these hypothetical interactions there are no experimental data to support or invalidate our predictions. However, we have been able to identify some interactions that are supported

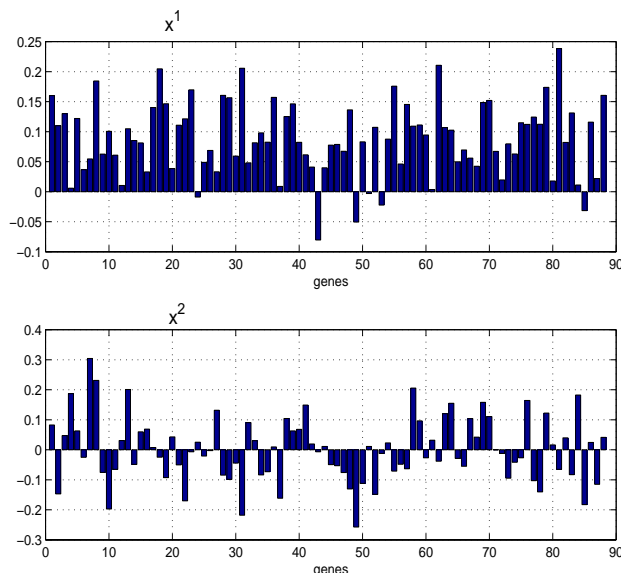


Figure 5: The magnitudes of the two dominant eigenvalues of the averaged  $CC^T$  matrix representing the effects of the two hidden variables on the observed gene expression levels

by existing literature. The relationship between myeloid differentiation primary response gene 88(Myd 88) and caspase 6 is one of these. Myd 88 is an adaptor protein involved in transducing the signals from Toll-like receptors and is important in the induction of the caspase cascade in programmed cell death. It has been reported that Myd 88 mediates both apoptosis and nuclear factor-kappaB (NF $\kappa$ B) activation by the Bacterial Lipoprotein (BLP) stimulated Toll like receptor gene (TLR2). This receptor signals apoptosis through Myd 88 via a pathway involving Fas-associated death domain protein (FADD) and caspase 8, a protease involved in the activation of the caspase cascade initiating programmed cell death [2]. Although current experimental evidence does not directly relate Myd 88 to the upregulation of caspase 6 it is reasonable to hypothesise that the activation of the caspase cascade may result in the transcriptional activation of members of this protease cascade. Interestingly caspase 8 and caspase 6 transcriptional profiles are highly correlated ( $r=0.75$ ), suggesting that the influence we have detected may extend to other members of this protein family.

#### 3.2 Multiple gene-gene interactions

A total of 15 genes were found to be connected in a single network of positive and negative interactions in the DAG produced from the D-matrix with Z-scores  $> 3$  (Figure 3 A,B). This network can be functionally subdivided into two small subgraphs. The first (Figure 3A and Figure 4) is characterised by the positive “causal” influence of the early growth response gene (EGR-1) on p53, caspase 4, cyclin G1 and transcription factor gene 12. Both p53 and caspase 4 genes are involved in the activation of programmed cell death, whereas cyclin G1 is a key gene involved in cell cycle progression [19]. The model also predicts a positive influence of p53 and a negative influence of NF $\kappa$ B inhibitor and Linker of Activated T cells (LAT) genes on the expression of

**Table 1: Table 1: The elements of D matrix  $D_{ij}$  with Z-scores  $> 3$  and connectivity = 1. The direction of the influence is from Gene A to Gene B. The last column indicates if the value of  $D_{ij}$  is positive (+) or negative (-)**

Gene A	Gene B	Influence
matrix metalloproteinase 7	Src-like-adaptor	+
zinc finger protein,subfamily 1A,1(lkaros)	transcription factor 8	-
mitogen-activated protein kinase kinase 4	inhibitor of DNA binding 3	+
cyclin-dependent kinase 6	jun D proto-oncogene	+
caspase 8, apoptosis-related cysteine protease	LPS-induced TNF-alpha factor	+
jun B proto-oncogene	inhibitor of DNA binder 1	+
interleukin 2 receptor, gamma	pyruvate dehydrogenase kinase,isoenzyme 2	-
cytochrome P450,subfamily XIX	chemokine (C-X3-C) receptor 1	+
myeloid differentiation primary response gene (Myd 88)	caspase 6	+

the FYB gene. Both LAT and FYB are accessory molecules involved in the transduction of the T cell receptor signal [26]. The second subgraph (Figure 3B) is characterised by the interaction of a number of genes involved in cell cycle progression and cytochrome c1. Both pathways are connected by the gene superoxide dismutase (SOD) which negatively influences the expression of cyclin dependent kinase inhibitor 2D (CDKN2D) and the transcription factor 12 gene. There are some experimental data supporting this model. The Early Growth Response gene (EGR-1) was first identified as an immediate-early-response gene transcriptionally activated by mitogenic stimulation [28]. EGR-1 is a transcription factor whose function is in the control of cell proliferation and apoptosis. Our model fits well with the experimental evidence available on EGR-1. Nair et al. [24] have demonstrated that EGR-1 binding directly activates p53. Such activation is necessary for the activation of an apoptotic response involving the activation of caspase genes. Moreover it has been demonstrated that direct binding of EGR-1 is necessary for the promoter sequence of cyclin D1 to respond to TGF $\alpha$  [31]. This has important mechanistic implications for the transcriptional regulation of cyclin D1 and cell cycle progression by an essential pro-proliferative growth factor and cell cycle progression. The positive effect of EGR-1 on cyclin G described in our model may be indirect and may be a consequence of cyclin D activation.

The relationship between NF $\kappa$ B, FYB and LAT is intriguing (Figure 3A). These genes encode molecules that are involved in the early events of T cell activation and FYB and LAT are both accessory molecules involved in transducing the signal from the T cell receptor. This association is consistent with the need to regulate the expression of members of the T cell receptor complex during T cell activation. This speculation will need to be verified experimentally.

In our model the expression of the SOD gene negatively influences the expression of the CDKN2D gene. SOD1 is a scavenger for superoxide radicals and converts  $O_2^{\cdot-}$  into  $H_2O_2$ . The SOD gene has been shown to have anti-proliferative effects that seem to be associated with the accumulation of  $H_2O_2$  in the cell [6]. This fact apparently contradicts the predicted negative influence on the CDKN2D gene, an inhibitor of cell cycle progression. In our model there are also some evident inconsistencies. The negative influence of IL-3R on PCNA (Figure 3C), for example, is contradicted by the demonstrated role of IL-3 as a growth factor. At

present we do not have any rational explanation for this. A certain number of artefacts and inconsistencies are, however, to be expected in any evolving model. More data need to be collected and our modeling techniques modified after the necessary experimental feedback has been obtained.

## 4. CONCLUSIONS

In conclusion, we have demonstrated how the application of linear dynamical systems modeling on highly replicated gene expression microarray data is providing a useful tool for investigating the causal influences between gene expression events. Experimental verification of these results is needed to fully validate the approach and improve the model. Over-expression and gene knock-out experiments will be useful approaches to test the validity of our predictions. In addition, we are aware that many of the assumptions inherent in LDS models (eg: linear Gaussian dynamics and measurements and Gaussian noise) may be unrealistic for gene expression time series data. Further work will investigate these limitations and propose nonlinear, and, if necessary, non-Gaussian extensions to our model.

## 5. ACKNOWLEDGMENTS

The authors would like to thank John Angus (Claremont Graduate University) for helpful discussions, Alpan Raval (Keck Graduate Institute) for valuable assistance with  $\LaTeX$ , Elizabeth Sotheran (Lorantis Ltd) for help with the microarray experiments and Brian Champion (Lorantis Ltd) for useful comments and for his enthusiastic support of the project.

## 6. ADDITIONAL AUTHORS

Zoubin Ghahramani (Gatsby Unit for Computational Neuroscience, University College, London, WC1N 3AR, UK email [zoubin@gatsby.ucl.ac.uk](mailto:zoubin@gatsby.ucl.ac.uk)) and Alessia Gaiba (Lorantis Limited, Babraham Hall, Babraham, Cambridge, CB2 4UL, UK email: [alessia.gaiba@lorantis.co.uk](mailto:alessia.gaiba@lorantis.co.uk))

## 7. REFERENCES

- [1] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pac. Symp. Biocomput.*, pages 17–28, 1999.
- [2] A. AO, Y. RB, W. DS, G. P, and Z. A. The apoptotic signaling pathway activated by toll-like receptor-2. *EMBO J.*, 19(13):3325–3336, July 2000.

- [3] N. L. A.P. Dempster and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.
- [4] M. Appleby, J. Gross, M. Cooke, S. Levin, X. Qian, and R. Perlmutter. Defective t cell receptor signalling in mice lacking the thymic isoform of p59 (csk). *Cell*, 70:751–763, 1992.
- [5] E. Arpaia, M. Shahar, H. Dadi, A. Cohen, and C. Roifman. Defective t cell receptor signalling and cd8 (+) thymic selection in humans lacking zap-70 kinase. *Cell*, 76:947–958, 1994.
- [6] D. Bernard, B. Quatannens, A. Begue, B. Vandebunder, and C. Abbadie. Antiproliferative and antiapoptotic effects of cre1 may occur within the same cells via the up-regulation of manganese superoxide dismutase. *Cancer Res.*, 61(6):2656–2664, March 2001.
- [7] M. Berridge and R. Irvine. Inositol phosphates and releases calcium ions from intracellular stores. *Nature*, 341:197–205, 1989.
- [8] R. G. Brown and P. Y. Hwang. *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley and Sons, New York, 1997.
- [9] M. Castagna, Y. Takai, K. Kaibuchi, K. Sano, U. Kikkawa, and U. Nishizuka. Direct activation of calcium-activated, phospholipid-dependent protein kinase by tumor promoting phorbol esters. *J. Biol. Chem.*, 257:7847–7851, 1982.
- [10] B. Efron and R. J. Tibshirani. *An introduction to the Bootstrap*. Chapman and Hall, New York, 1993.
- [11] W. J. Ewens and G. R. Grant. *Statistical Methods in Bioinformatics*. Springer-Verlag, New York, 2001.
- [12] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *J. Comput. Biol.*, 7:601–620, 2000.
- [13] E. Gelfand, K. Weinberg, B. Mazer, T. Kadlecsek, and A. Weiss. Absence of zap-70 prevents signalling through the antigen receptor on peripheral blood t cells but not on thymocytes. *J. Exp. Med.*, 182:1057–1065, 1995.
- [14] Z. Ghahramani and G. E. Hinton. Parameter estimation for linear dynamical systems. *Technical Report CRG-TR-96-2*, 1996.
- [15] A. Hartemink, D. Gifford, T. Jaakkola, and R.A.Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.*, pages 422–433, 2001.
- [16] P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Hughes, E. Snosrud, N. Lee, and J. Quackenbush. A concise guide to cDNA microarray analysis. *Biotechniques*, 29(3):548–550, 552–554, 556 passim, September 2000.
- [17] J. Imboden and J. Stobo. Transmembrane signalling by t cell antigen receptor: perturbation of the t3-antigen receptor complex generates inositol phosphates and releases calcium ions from intracellular stores. *J. Exp. Med.*, 161:446–456, 1985.
- [18] M. Iwashima, B. Irving, N. V. Oers, A.C.Chan, and A. Weiss. Sequential interactions of the tcr with two distinct cytoplasmic tyrosine kinases. *Science*, 263:1136–1139, 1994.
- [19] S. Kimura, M. Ikawa, A. Ito, M. Okabe, and H. Nojima. Cyclin g1 is involved in g2/m arrest in response to dna damage and in growth control after damage recovery. *Oncogene*, 20(25):3290–3300, May 2001.
- [20] S. Ley, A. Davies, B. Druker, and M. Crumpton. 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, 1991.
- [21] S. Liang, S. Fuhrman, and R. Somogyi. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pac. Symp. Biocomput.*, pages 18–29, 1998.
- [22] B. Manger, A. Weiss, J. Imboden, T. Laing, and J. Stobo. 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, 1987.
- [23] J. Miskin. *D. Phil Thesis*. University of Cambridge, 2001.
- [24] P. Nair, Muthukkumar, S. Sells, S. Han, V. Sukhatme, and V. Rangnekar. Early growth response-1-dependent apoptosis is mediated by p53. *Journal of biological chemistry*, 272:20131–20138, 1997.
- [25] S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neural Computation*, 11:305–345, 1999.
- [26] A. J. D. Silva, L. Zhuwen, C. D. Vera, C. E. P. Findell, and C. E. Rudd. Cloning of a novel t-cell protein fyb that binds fyn and sh2-domain-containing leukocyte protein 76 and modulates interleukin 2 production. *Proc. Nat. Acad. Sci. USA*, 94:7493–7498, 1997.
- [27] P. Stein, H.-M. Lee, S. Rich, and P. Soriano. pp59fyn mutant mice display differential signalling in thymocytes and peripheral t cells. *Cell*, 70:741–750, 1992.
- [28] V. Sukhatme, X. Cao, L. Chang, C. Tsai-Morris, D. Stamenkovich, P. Ferreira, D. Cohen, S.A.Edwards, T. Shows, T. C. T, M. Lebeau, and E. Adamson. A zinc finger-encoding gene coregulated with c-fos during growth and differentiation, and after cellular depolarization. *Cell*, 53(1):37–43, April 1988.
- [29] R. Thomas. Boolean formalization of genetic control circuits. *J Theor Biol*, 42(3):563–586, 1973.

- [30] A. Weiss, J. Imboden, R. Wiskocil, and J. Stobo. The role of t3 in the activation of human t cells. *J. Clin. Immunol*, 4:165–173, 1984.
- [31] Y. Yan, H. Nakagawa, M. Lee, and A. Rustgi. Transforming growth factor-alpha enhances cyclin d1 transcription through the binding of early growth response protein to a cis-regulatory element in the cyclin d1 promoter. *J. Biol. Chem.*, 272(52):33181–33190, December 1997.