BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 13 2010, pages 1623–1629
doi:10.1093/bioinformatics/btq244

Gene expression

Advance Access publication May 5, 2010

Learning combinatorial transcriptional dynamics from gene expression data

Manfred Opper[1] and Guido Sanguinetti[2,*]

[1]Department of Computer Science, Technische Universität Berlin D-10587 Berlin, Germany and [2]School of Informatics, University of Edinburgh, 10 Crichton St, Edinburgh EH8 9AB, UK

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: mRNA transcriptional dynamics is governed by a complex network of transcription factor (TF) proteins. Experimental and theoretical analysis of this process is hindered by the fact that measurements of TF activity *in vivo* is very challenging. Current models that jointly infer TF activities and model parameters rely on either of the two main simplifying assumptions: either the dynamics is simplified (e.g. assuming quasi-steady state) or the interactions between TFs are ignored, resulting in models accounting for a single TF.

Results: We present a novel approach to reverse engineer the dynamics of multiple TFs jointly regulating the expression of a set of genes. The model relies on a continuous time, differential equation description of transcriptional dynamics where TFs are treated as latent on/off variables and are modelled using a switching stochastic process (telegraph process). The model can not only incorporate both activation and repression, but allows any non-trivial interaction between TFs, including AND and OR gates. By using a factorization assumption within a variational Bayesian treatment we formulate a framework that can reconstruct both the activity profiles of the TFs and the type of regulation from time series gene expression data. We demonstrate the identifiability of the model on a simple but non-trivial synthetic example, and then use it to formulate non-trivial predictions about transcriptional control during yeast metabolism.

Availability: `http://homepages.inf.ed.ac.uk/gsanguin/`

Contact: `g.sanguinetti@ed.ac.uk`

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 17, 2009; revised on April 22, 2010; accepted on May 2, 2010

1 INTRODUCTION

Understanding the mechanisms by which cells alter gene expression in response to external stimuli is a fundamental challenge in systems biology. Perhaps the most critical level at which this control is exerted is in the regulation of mRNA transcription. A common mechanism for achieving this is by structural or chemical modification of transcription factor (TF) proteins: the environmental signal (typically a change in the concentration of some small molecule) causes a change in state in the TF, which enables the protein to bind target promoters and hence modulate the rate of recruitment of RNA polymerase and consequently transcription (Ptashne and Gann, 2002). This view lends itself to a straightforward mathematical formulation in terms of ordinary differential equations (ODEs), and has played a prominent role in the success of kinetic models of regulatory networks, enabling the construction of detailed models of many important processes (Demin and Goryanin, 2008).

Despite the success of kinetic modelling, a major obstacle to its wider application is its reliance on an accurate knowledge of a number of model parameters that are often difficult to experimentally measure. In particular, TFs' activity states are often difficult to assay in a dynamical fashion due to several technical limitations: TFs are often present at low intercellular concentrations and the changes in their activity state can occur rapidly due to post-translational modifications. This has led to considerable activity in the statistics and machine learning community to develop computational methods to infer TF activity profiles from time-course measurements of mRNA expression levels. Early TF inference models exploited the availability of maps of the architecture of the regulatory network (obtained e.g. through high-throughput ChIP-on-chip experiments; Lee *et al*., 2002) to simultaneously infer the activity profile of a large number of TFs. This approach was pioneered by the network component analysis model of Liao *et al*. (2003), which was subsequently cast in a probabilistic framework (Sabatti and James, 2006; Sanguinetti *et al*., 2006). While the predictions of such models have provided valuable biological insights (Partridge *et al*., 2007), the large number of variables means that only extremely simplified models of transcriptional control can be considered to ensure model identifiability. To have more realistic predictions, some authors have recently proposed inferential approaches based on realistic, ODE-based models of transcription (Barenco *et al*., 2006; Khanin *et al*., 2007; Lawrence *et al*., 2006; Rogers *et al*., 2007; Sanguinetti *et al*., 2009). Thus, gene expression dynamics is governed by an ODE with a driving force represented by the TF activity. Placing a suitable prior distribution over the TF activity, posterior estimation can be performed, along with estimation of model parameters such as mRNA decay and synthesis rates. While there is much promise in employing more realistic models, the mathematical difficulties encountered in solving this inverse problem are significant, and the inference procedure cannot handle models of the complexity routinely used in kinetic modelling. To our knowledge, all existing methods restrict themselves to the single-input motif (SIM) scenario, where a module of genes is controlled by a single TF.

In this contribution, we build on our previous model of SIM dynamics (Sanguinetti *et al*., 2009) and extend the results to simultaneously infer the activities of multiple interacting TFs.

*To whom correspondence should be addressed.

This removes in principle one of the major obstacles to the use of statistical, ODE-based models in large-scale systems biology applications. We demonstrate the identifiability of the model on a simple yet non-trivial synthetic example, where three genes are controlled by two TFs that interact non-trivially at one of the genes' promoter. The model successfully reconstructs both the TF activities and the interaction parameters. We then apply the model to a microarray time series of yeast respiration (Tu *et al*., 2005), focusing on two important regulators of yeast metabolism, FHL1 and RAP1. While the main contribution of this article is methodological, the results on both the synthetic and the real data demonstrate that the model is capable of making strong predictions potentially leading to new biological insights.

2 MODEL AND METHODS

We first describe the model of transcription regulation, which we will employ. The starting point for the derivation of our model is the classical Michaelis–Menten model of transcriptional control

$$\frac{dx}{dt}=\frac{Af(t)}{\kappa_1+f(t)}+b-\lambda x, \tag{1}$$

here x represents the mRNA concentration for a particular gene, $f(t)$ the concentration of active TF (assumed in this case to be an activator), A and κ_1 are kinetic constants, b a baseline expression rate and λ the mRNA decay rate. In many biological processes, TF transit from inactive to active state as a consequence of fast post-translational modifications; as a consequence, a logical approximation to Equation (1), whereby the TF activity is modelled as a binary variable $\mu(t)\in\{0,1\}\quad\forall t$, is often reasonable, leading to

$$\frac{\mathrm{d}x}{\mathrm{d}t}=A\mu(t)+b-\lambda x. \tag{2}$$

The *inverse problem* that we wish to address is: *given measurements of x, can we infer the switching TF activity* $\mu(t)$ *and model parameters*? The inverse problem for model (2) was solved in the SIM hypothesis in Sanguinetti *et al*. (2009). Let us now assume that we have a set of genes which are controlled jointly by a number of TFs; for simplicity, we will discuss the case when the number of TFs is two, the generalization to more TFs is theoretically straightforward (although more data will be needed given the increased number of parameters). The transcription model takes the form

$$\frac{\mathrm{d}x^i}{\mathrm{d}t}=A_1^i\mu_1(t)+A_2^i\mu_2(t)+A_{12}^i\mu_1(t)\mu_2(t)+b^i-\lambda^i x^i. \tag{3}$$

This model encompasses both activation and repression, and contains as special cases, the OR gates ($A_{12}=0$, $A_1=A_2$) and the AND gate ($A_1=A_2=0$). Solving the inverse problem in this case is far more challenging. In the rest of this section, we demonstrate our solution on both simulated data and real data; the details of the derivation are given in the Supplementary Material.

The conceptual steps in the development of our approach are as follows. First, we define the inference problem, and describe the approximations we make to render the system tractable. We then propose a dual inference approach, where the minimization required in variational inference is shown to be equivalent to a (simpler) saddle point problem, with strong guarantees of convergence. Finally, we highlight a mean-field approximation that allows to reduce the complexity in the case of multiple interacting TFs.

2.1 Inference

The first step in setting up a Bayesian inference framework is to define prior distributions for the variables involved. Following (Sanguinetti *et al*., 2009), we model the latent TF activity $\mu(t)$ as a *telegraph process*, a two states Markov Jump Process. A quantity of particular importance in stochastic models is the *single time marginal*, i.e. the probability of the switch μ being in the on state at time t. As a consequence of the Markov assumption, it can be shown that the single time marginal is given by the following Master equation:

$$\frac{dp_1(t)}{dt}=-(f_++f_-)p_1(t)+f_+, \tag{4}$$

here $p_1(t)=p(\mu(t)=1)$ and f_+ and f_- represent the *transition rates* of the process, defined as

$$f_+(t)=\lim_{\delta t\to 0}\frac{p\big(\mu(t+\delta t)=1|\mu(t)=0\big)}{\delta t}$$

and analogously for f_-. In the case of multiple TFs, the respective processes are assumed to be independent *a priori*. Notice that the stochastic process prior on the TFs implies that the mRNA concentrations are also a stochastic process, even though the relationship between x and μ is entirely deterministic. The prior distribution then is combined with an observation model (likelihood) that relates the observed variables to the latent variables. In this case, we model the observations $y_i(t)$ of mRNA species i at time t as normally distributed around the value of the random variable $x(t)$, i.e

$$y_i(t)|\mu_{1,2}(t)\sim\mathcal{N}\left(x_i(t),\sigma_i^2\right). \tag{5}$$

Given a prior model and a likelihood, we can then combine these into Bayes' theorem to obtain the posterior over the process as

$$p(\mu|\mathbf{y},\Xi)=\frac{1}{Z}p(\mathbf{y}|\mu,\Xi)p(\mu), \tag{6}$$

where $\mathbf{y}$ denotes collectively the observations, Ξ are all the parameters involved in Equation (3) and Z a normalization constant independent of μ. It should be noticed that the probability measures involved in Equation (6) are to be understood as probabilities over the space of *trajectories* of the system, not as single time marginals. We will use a variational formulation of the inference problem (Sanguinetti *et al*., 2009).Variational inference is a powerful inference method based on tools from optimization. Free form (i.e. *unconstrained*) variational inference is entirely equivalent to the general inference problem. Often, however, variational inference is used as an approximation technique: given an intractable probability distribution p, the variational approach finds an optimal approximation q within a certain family of distributions. This is usually done by minimizing the *Kullback–Leibler (KL) divergence* between the two distribution

$$\mathrm{KL}[q\|p]=E_q\left[\log\frac{q}{p}\right]=\int \mathrm{d}x q(x)\log\frac{q(x)}{p(x)}.$$

By selecting a suitable family of approximating distributions, the inference problem is then turned into an optimization problem. It can be shown that the KL divergence is a convex functional of q and is equal to zero iff $q=p$ (e.g. Cover and Thomas, 2006). In this case, we will choose the approximating process q to be again a Markov Jump Process, so that the required KL is given by

$$\begin{aligned}\mathrm{KL}\big[q\|p_{\mathrm{post}}\big]&=\int \mathrm{d}\mu q\log\frac{q}{p_{\mathrm{post}}}=\\ &\log Z-E_q\big[\log p(\mathbf{y}|\mu,\Xi)\big]+\mathrm{KL}\big[q\|p_{\mathrm{prior}}\big],\end{aligned} \tag{7}$$

here Z is a normalization constant, $E_q\big[\log p(\mathbf{y}|\mu,\Xi)\big]$ the expectation of the likelihood of the observations under the approximating process and $\mathrm{KL}\big[q\|p_{\mathrm{prior}}\big]$ the KL divergence between the prior process and the approximating process. The integration element $\mathrm{d}\mu$ denotes integration over all possible paths of the random process μ. The KL divergence between the two general Markov Jump Processes was derived in Opper and Sanguinetti (2007); a derivation is added in the Supplementary Material. Therefore, the inference problem can be turned into an optimization problem (in an infinite dimensional space). Notice that Equation (7) involves the computation of $E_q\big[x_i^2\big]=\bar{x_i^2}$ in the expectation of the likelihood (5). This introduces long-range correlations that make the posterior process non-Markovian (and significantly increase the computational burden; Sanguinetti *et al*., 2009). However, assuming that the latent TFs perform rapid switches between their

on/off states, the variance of the x process can be neglected, so that $\bar{x_i^2} \sim \bar{x}_i^2$. This is equivalent to replacing the correct likelihood (5) with an approximate version

$$y_i(t)|\mu_{1,2}(t) \sim \mathcal{N}\left(\bar{x}_i(t), \sigma_i^2\right). \tag{8}$$

As we neglect the long-range correlations introduced by $E[x^2]$, it can be showed, by considering the structure of the relevant graphical model, that the posterior computed with this approximate likelihood will be Markovian, yielding significant computational advantages (see Supplementary Material for details).

2.2 Dual inference problem

By direct computation, minimization of the KL functional (7) can be represented as the saddle point problem

$$\hat{F} = \max_{\theta} \min_{q} \left\{ \sum_{i=1}^{n} \left[\theta_i \left(y_i - \bar{x}(t_i) \right) - \frac{\sigma^2}{2} \theta_i^2 \right] + \mathrm{KL}\left[q \| p_{\mathrm{prior}} \right] \right\}$$

where $\bar{x} = E_q[x]$ and we have introduced auxiliary variables θ_i (one for each observation). By inspection and using the properties of the KL divergence, we see that this functional is concave in θ and convex in q. Hence we can exchange min and max. Performing the max first yields the result. This also shows that there is only a unique saddle point solution.

Setting $x(0) = 0$ for simplicity and using

$$x(t) = \int_0^t \exp\left(\lambda(s-t)\right) [A\mu(s) + b] ds$$

we get

$$\begin{aligned} \hat{F} = \max_{\theta} \min_{q} \Big\{ & \mathrm{KL}\left[q \| p_{\mathrm{prior}} \right] \\ & - \sum_i \left[\theta_i \int_0^{t_i} \exp\left(\lambda(s-t_i)\right) [A q_1(s) + b] ds - \left(\theta_i y_i - \frac{\sigma^2}{2} \theta_i^2 \right) \right] \Big\} \end{aligned} \tag{9}$$

where we have used the definition of the single time marginals of the q process $E_q[\mu(t)] = q_1(s)$. Thus, we have to perform

$$\min_q \left\{ \mathrm{KL}(q,p) - \int_0^T [q_1(t) + b] W(t) \mathrm{d}t \right\}$$

with

$$W(t) = \sum_i \theta_i e^{\lambda(t-t_i)} \Theta(t_i - t)$$

and with $\Theta(x)$ the unit step function. Minimization of this functional is done by computing its functional derivatives w.r.t. the marginals of the approximating process q *and* the transition rates for the approximating process g, which are allowed to depend on time. Naturally, these quantities are not independent, but are related by the Master Equation (4), so that a constraint must be added to the functional using Lagrange multipliers. Setting to zero the functional derivatives leads to an ODE for the Lagrange multiplier, which is solved backwards in time. This allows us to compute the process rates, and finally the single time marginals for the approximating process are computed by numerically solving the Master Equation (4) forward in time. This procedure is closely related to the familiar forward–backward procedure used in Hidden Markov Models (e.g. Bishop, 2006). Details of the minimization procedure are given below in the general multi-TF case.

Having performed the minimization w.r.t. q, we can update θ by performing gradient ascent. Computation of the explicit gradient w.r.t. θ of the functional (9) is trivial, and implicit dependencies on θ (through the optimal q) need not be taken into account since we are operating at the minimum w.r.t. q. Hence, choosing a sufficiently small learning step, we are guaranteed that this procedure will converge to the saddle point solution. Solving the saddle point problem is interleaved with estimations of the parameters in a expectation-maximization scheme (see Supplementary Material).

2.3 Multiple TFs

We will restrict the discussion to the case of two TFs, extension to $n > 2$ is conceptually straightforward. The input TFs are assumed to be a priori independent (i.e. the prior process p is factorized); this is a reasonable assumption as different TFs normally sense different stimuli. For simplicity, we will also assume that the prior transition rates for the two TFs are the same. More importantly, we will impose an *a posteriori* factorization, in that the process q will be assumed to factorize as $(q^1(\mu_1), q^2(\mu_2))$. Thus, the variational approach will no longer perform exact inference, rather obtain the mean-field solution.The KL divergence between the true posterior and the approximating process now becomes

$$\begin{aligned} \mathrm{KL}[q\|p] = & \frac{1}{Z} + \mathrm{KL}\left[q^1 \| p\right] + \mathrm{KL}\left[q^2 \| p\right] \\ & + \sum_i \sum_j \frac{1}{2(\sigma^i)^2} \left[y_j^i - \bar{x}^i(t_j) \right]^2 \end{aligned} \tag{10}$$

where y are the noisy observations and $\bar{x}$ are the expectations under the joint posteriors for both TFs. Due to the independence assumption, we have that

$$\begin{aligned} \bar{x}^i(t) = & \exp[-\lambda^i t] \int_0^t \exp[\lambda^i s] \Big[A_1^i q_1^1(s) + A_2^i q_1^2(s) \\ & + A_{12}^i q_1^1(s) q_1^2(s) + b^i \Big] \mathrm{d}s \end{aligned}$$

where q_1^j is the (posterior) probability that the j-th TF is on.

The saddle point representation of the optimization problem is analogous to the single TF process. The min–max objective function is given by

$$\begin{aligned} F\left[q^1, q^2, g\right] = & \mathrm{KL}\left[q^1 \| p\right] + \mathrm{KL}\left[q^2 \| p\right] \\ & - \sum_i \int_0^T \left[A_1^i q_1^1(t) + A_2^i q_1^2(t) + A_{12}^i q_1^1(t) q_1^2(t) + b^i \right] W^i(t) \mathrm{d}t. \end{aligned} \tag{11}$$

As before we have introduced

$$W^i(t) = \sum_j \theta_j^i \exp\left[\lambda^i (t - t_j)\right] \Theta(t_j - t).$$

Once again, optimization w.r.t. the approximating process q can be interleaved with optimization w.r.t. the parameters of the model in a variational expectation-maximization scheme.

The main difference in solving the saddle point problem is the fact that now the minimization step is no longer exact, since we are approximating the posterior with a factorized process. Therefore, we need to iterate minimization of the two species by forward–backward solving of ODEs (see Supplementary Material). It is also important to notice that, without any information on the parameters A, the model admits multiple solutions, i.e. permutations between TFs can render the model non-identifiable. This identifiability problem can be resolved if prior information about the connectivity of the regulatory network is available, in particular, if it is known that some genes are not bound by one of the TFs (implying some A parameters are zero). This type of information is partly available for model organisms such as yeast and *Escherichia coli*, and its availability for selected regulons is becoming increasingly available due to the advent of functional genomics techniques such as ChIP-on-chip. A more formal discussion of identifiability issues, including potential problems that can arise in specific situations, is given in the Supplementary Material.

3 RESULTS

3.1 Synthetic data

Computational experiments on simulated data are often precious to test model identifiability and to assess limitations due to the existence of a gold standard. We tested our approach on a simple yet non-trivial synthetically generated example, consisting of three

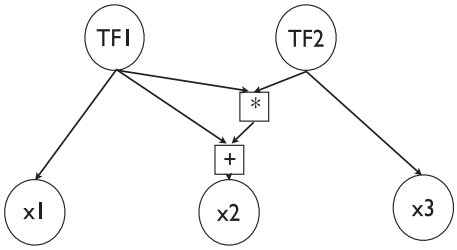

Fig. 1. Schematic of the regulatory structure of the simple network used to generate synthetic data: arrows denote activation and boxes denote interactions. $X1$, $X2$ and $X3$ are gene promoters; plus sign denotes additive interaction of TFs, asterisk multiplicative interaction. The two TFs interact non-trivially at the promoter of the second gene, where a weak response to TF1 is integrated with a strong response to the presence of both TFs.

genes jointly controlled by two TFs[1]. Gene 1 is activated by TF1 and not bound by TF2, Gene 3 conversely is solely activated by TF2, and this information was used in the inference by clamping

$$A_2^1 = A_{12}^1 = A_1^3 = A_{12}^3 = 0.$$

Gene 2 has a rather complex control structure: it needs TF1 to be expressed at all, but TF1's activation can be greatly enhanced if TF2 is also present. TF2 on its own, on the other hand, has no effect on the expression of Gene 2. The model, however, was not presented with any of this information, so that all three parameters A_1^2, A_2^2 and A_{12}^2 remain to be learned. A schematic of the system is given in Figure 1.

Starting from an input signal $(\mu_1(t), \mu_2(t))$ for the two TFs, we generated expression profiles for the three genes using the model (3) and sub-sampled it at regular intervals to obtain three short time series of 20 observations. To mimic a typical experimental situation e.g. a microarray experiment, we then corrupted the observations by adding Gaussian noise whose SD is ∼10% of the total variance of each time course. To assess empirically the statistical reliability of our analysis, we repeated the experiment for three different sets of parameter values and five independently generated time courses for each set of the parameters, resulting in 15 runs of the algorithm.

Figure 2A–C shows an example of a noisy time course used for this analysis. The bottom row shows some of the results of the analysis. The first two figures show the posterior mean for the TF activities (μ_1, μ_2), showing a good agreement with the true profiles. This is perhaps not surprising, since we have constrained the model by providing information on the connectivity. Notice, however, that the true profiles are in the same state for the majority of time, making this quite a tough inferential problem. What is more remarkable is Figure 2F. This shows the posterior distributions over the A parameters for the second gene, which are responsible for its highly non-trivial transcriptional control. The results clearly indicate that the model is able to correctly identify the control mechanism responsible for the expression profile of the second gene, even with such a short and noisy time course. The computational time for a run of the algorithm was ∼5 min on a standard laptop machine; further results on the scaling of the run times on larger datasets are given in the Supplementary Material.

[1]All software to recreate the experimental results can be freely downloaded from `http://www.dcs.shef.ac.uk/~guido/software.html`.

To analyse the statistical reliability of the parameter estimation, we repeated the analysis on 15 datasets corresponding to three different sets of parameters (each with five independently generated time series). For each parameter estimated, we computed an empirical deviance as

$$z = \frac{\hat{\xi} - \xi}{\mathrm{std}(\xi)} \tag{12}$$

where ξ denotes the true value of the parameter, $\hat{\xi}$ the estimate computed from the data and std(ξ) the square root of the posterior variance computed using a variational Bayes approach (using a flat prior). By expanding the terms in Equation (6), it can be shown (see Supplementary Material) that the variational posterior for the parameters A and b is a truncated Gaussian. If the mean of the Gaussian is much higher than its SD, this can be well approximated by a Gaussian, so that the deviance defined in (12) should theoretically be distributed as a unitary Gaussian, $z \sim \mathcal{N}(0,1)$. Figure 3 shows a comparison of the histogram obtained from the empirical deviances and the theoretical unitary Gaussian histogram (notice that, since the centre of the bins in the empirical histogram are not exactly symmetric about zero, the Gaussian histogram also does not appear to be symmetric). As we can see, there is a reasonable agreement, although the higher tail tends to be slightly higher than expected, indicating a slight bias towards overestimating the parameters. A graphical summary of the parameter inference for the sensitivity parameter A is given in the Supplementary Material.

3.2 Control of ribosomal protein production in yeast metabolic cycle

As a real example to test our approach, we considered the dynamics of control of ribosomal protein production in yeast respiration. The respiratory cycle of yeast (also known as metabolic cycle) was assayed using microarrays by Tu *et al*. (2005). This dataset consists of 36 time points sampled at 25 min intervals through three cycles of yeast respiration induced by starvation followed by a period of constant supply of glucose. The study determined that over half of the yeast genes had oscillatory expression patterns (at 95% significance level). The qualitative behaviour of these genes generally resembled more a square wave than a sinusoidal one, indicating that regulatory transitions happen over a fast time scale (compared to the sampling rate).

Two separate genome-wide ChIP-on-chip studies have been published describing the architecture of the yeast transcriptional regulatory network (Harbison *et al*., 2004; Lee *et al*., 2002). Both these studies performed ChIP-on-chip on over 100 TFs, resulting in several thousands TF–DNA interactions. The overlap between these datasets is substantial but not total; to include as much information as possible, we merged the two datasets, including a link if at least one experiment reported the TF binding the promoter of the gene at a 0.001 P-value.

To test our model on this data, we selected two important transcriptional regulators controlling the production of ribosomal proteins, FHL1 and RAP1 (Schawalder *et al*., 2004). In total, we selected 10 genes to perform posterior inference of TF activities. To ensure identifiability, we included three genes that are regulated solely by FHL1 according to the ChIP-on-chip data (TOS4, YLR030W and TKL2), and two genes that are regulated solely by RAP1 (VTS1 and PFK27). The remaining five genes code for

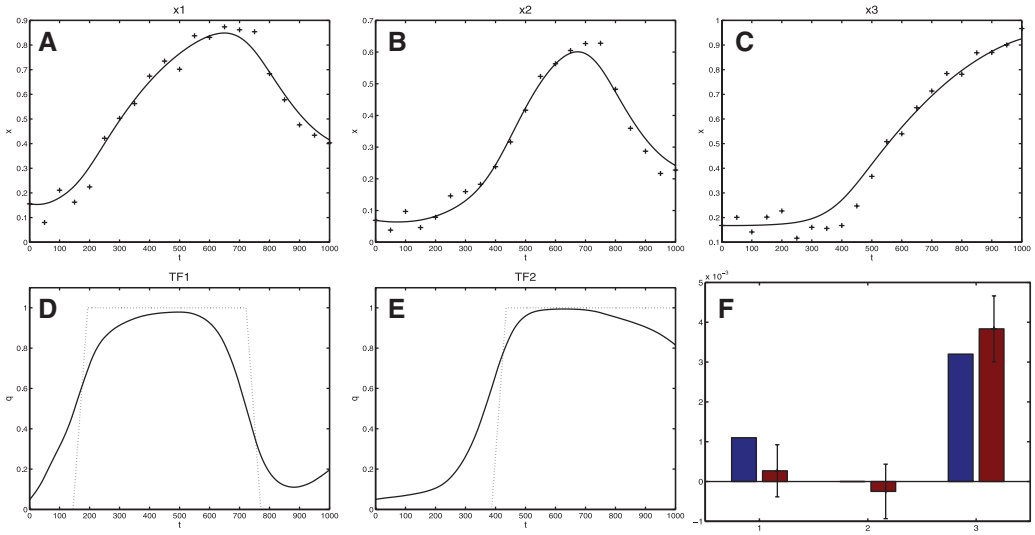

Fig. 2. Results on synthetic data: (**A**–**C**), noise-corrupted synthetic observations (crosses) and posterior mean reconstructed profiles for the three genes. (**D** and **E**) inferred posterior mean activity (solid) versus true input (dotted) for the two TFs, and true values of A_1^2, A_2^2 and A_{12}^2 (blue columns) versus inferred posterior values of the same parameters (red columns), obtained from the expression profiles shown on top. Notice that, as TF2 has no effect on Gene 2 in isolation, there is no blue bar in the middle position in (**F**).

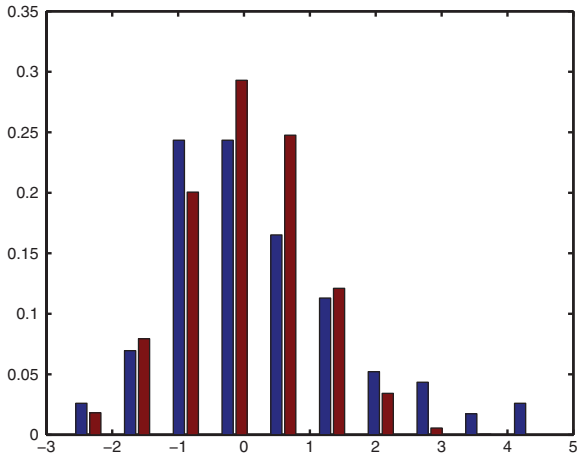

Fig. 3. Empirical analysis of the parameter estimates: the blue bars are the normalized histogram of the empirical deviances from the true values; red bars are the corresponding theoretical distribution (unitary Gaussian).

ribosomal proteins and are jointly regulated by FHL1 and RAP1, although the precise nature of the control is not known. These genes were RPL9A, RPL13A, RPL17B, RPL30 and RPS16B. Three of these genes (RPL13A, RPL17B and RPS16B) were chosen since they exhibit the largest variance in the microarray time course, and hence are likely to provide a cleaner representation of the output of the system. The other two genes were included, despite having noisier profiles, since their expression level was measured in a FHL1 mutant strain (Schawalder *et al*., 2004), allowing a quantitative test of our predictions.[2]

[2]Although RAP1 binding at the RPL9A promoter was not detected in either of the ChIP-on-chip data we used, it was confirmed using a different experimental methodology in Schawalder *et al*. (2004).

The posterior TF profiles are shown in Figure 4A and B; additional figures showing the fit of the model to the data are given in the Supplementary Material. Both inferred TF profiles sow a noisy periodic behaviour; however, the two TF profiles are sufficiently different to allow identifiability of the response parameters A_i^j in Equation (3). Figure 4C gives the values of the parameters A_i^j for the five ribosomal protein genes considered (RPL17B, RPL13A, RPS16B, RPL9A and RPL30). The blue column represents the response A_1^j to FHL1 alone, the green column is the response A_2^j to RAP1 alone and the red column represents the joint response A_{12}^j. In all cases, the model predicts a significant activation by RAP1 alone, while the effect of FHL1 alone is much smaller and sometimes indistinguishable from zero. FHL1, however, is predicted to play a significant role in enhancing the response to RAP1 when both TFs are present; in other words, in all cases we predict a significant combinatorial activation. These predictions can be verified by mutagenetic techniques, i.e. by knocking out one of the TFs. A recent study on the regulation of ribosomal protein genes quantified the transcriptional effect of a FHL1 deletion on mRNA expression of RPL9A and RPL30 (Schawalder *et al*., 2004), reporting a decrease in expression levels (relative to actin) of $\sim 60 \pm 10\%$ and $40 \pm 10\%$, respectively. Comparing our predictions to these results is not trivial, since the experimental setup is different; in particular, in the data we consider, cells have been synchronized according to their respiration state, which is not the case in the mutagenetic experiment in Schawalder *et al*. (2004). To obviate this problem, we consider the time average of the (posterior average) dynamic response to the TFs over the whole experiment of Tu *et al*. (2005), computed as

$$\bar{R}_{wt} = \int \mathrm{d}t A_1^i q_1(t) + A_2^i q_2(t) + A_{12}^i q_1(t) q_2(t).$$

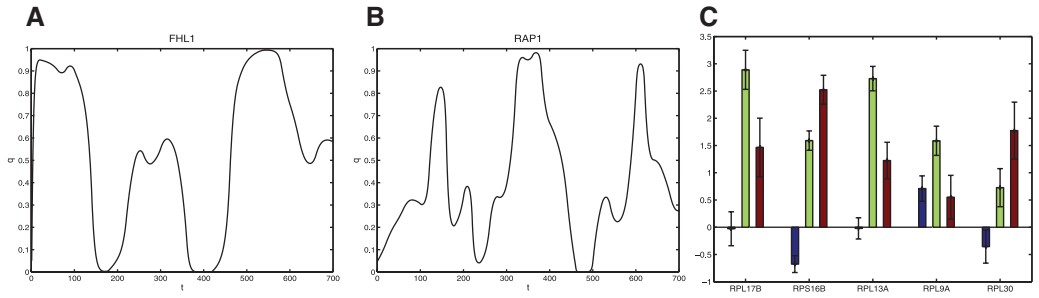

Fig. 4. Results on yeast metabolic cycle data: (**A**) posterior mean profile for FHL1, (**B**) posterior mean profile for RAP1 and (**C**) bar-chart representation of the parameters A_i^j for the three ribosomal protein genes considered (RPL17B, RPL13A, RPS16B, RPL9A and RPL30), giving the logical structure of the interaction between the two TFs.

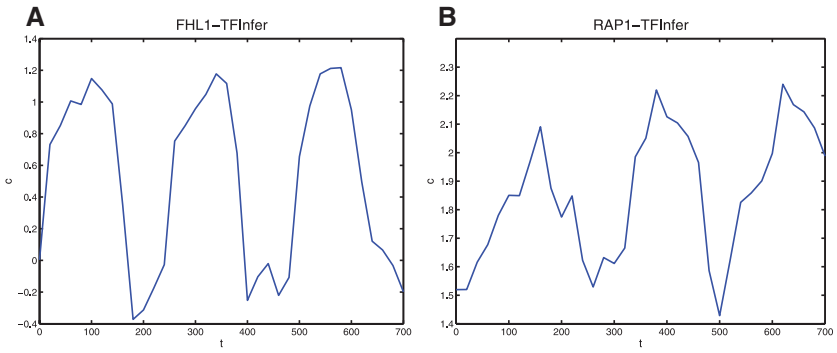

Fig. 5. Results on yeast metabolic cycle data using the discrete-time SSM of Sanguinetti *et al*. (2006): (**A**) posterior mean profile for FHL1 and (**B**) posterior mean profile for RAP1.

By setting $q_1(t) = 0 \forall t$, we can then predict what the effect of a FHL1 deletion would be on the observed mRNA dynamic range. In this way, we obtain that our model predicts an average reduction in mRNA expression for RPL9A of 44% (min 27% and max 56%), and of 32% for RPL30 (min 72% and max unchanged). While this is encouraging, it should be borne in mind that our model of regulation of ribosomal protein genes is still a simplified model, as several cofactor proteins known to be involved in the process are not explicitly modelled (Hu and Li, 2007).

It is interesting to compare the results of our approach with existing approaches to reverse engineering TF activities. As far as we know, no other ODE-based reverse engineering approach can handle multiple TF inputs. We, therefore, compared with a discrete-time approach (Sanguinetti *et al*., 2006). While the inferred TF profiles, shown in Figure 5, are broadly similar to the ones obtained with our method, showing a clear periodic behaviour, the discrete-time approach fails signally in decoding the non-trivial promoter structures, attributing almost entirely the expression of the ribosomal protein genes to the TF RAP1, in contrast with the existing experimental evidence (Schawalder *et al*., 2004). This is not surprising, as the discrete time model is not capable of handling non-linear effects.

4 CONCLUSION

In this article, we have presented a novel computational model to reverse engineer simultaneously both the activity of TFs and the logical structure of promoters from time-series gene expression data. The approach relies on a detailed model of transcription, which is an approximation to the Michaelis–Menten model from classical enzyme kinetics, and therefore should be able to capture accurately the effects that changes in TF activity have on gene expression dynamics.

There has been a considerable level of interest in parameter inference and reverse engineering for biological systems, whose dynamic is governed by systems of ODEs (for a recent survey, e.g. Lawrence *et al*., 2010). In most cases, the problem is tackled in the autonomous case, i.e. when the parameters are time-independent and no external input needs to be inferred. This reduces the problem to a finite dimensional (albeit non-linear) inference problem, for which a number of techniques can be used. In our case, instead, we focus on systems where a time-dependent external input is present, which is mediated as a time-varying TF activity that needs to be inferred at all time points. This is an *infinite dimensional* problem that we tackle by placing a stochastic process prior over the TFs' activities. As far as we know, previous work on this type of systems was restricted to the SIM scenario, where a single TF controls a number of genes (Barenco *et al*., 2006; Lawrence *et al*., 2006; Sanguinetti *et al*., 2009).

The main contribution of our approach is the ability to capture the logical structure of gene promoters, i.e. to predict whether the simultaneous binding of more than one species of TF will result in non-linear, combinatorial effects. This predictive power was demonstrated on a simulated dataset, and model results on a real

yeast dataset are shown to yield predictions that can be tested using standard molecular biology techniques such as mutagenesis. As far as we are aware, previous computational models of combinatorial regulation (Wang *et al*., 2005) adopt strong simplifying assumptions on the dynamics of transcription, or simply base their inference on the position of TF-binding sites in the genome, without modelling transcriptional dynamics (Hu *et al*., 2007). The key strength of our method is to render the system identifiable through an explicit model of TF dynamics (the telegraph process). This is both biologically plausible and constrains the system enough to allow both a good data fit and a strong identifiability of the parameters and latent functions. A further important feature of our method is its scalability: the complexity of the model is linear in the number of TFs, time points and genes involved, allowing in principle for generalizations to much larger systems.

There are several interesting possible extensions of the current model. Fluorescence data is becoming increasingly available, offering long time series of gene expression at a more refined resolution in terms of cellular populations. Intrinsic stochastic fluctuations in this type of data mean that stochastic differential equations (SDEs) provide more appropriate models, and it would be interesting to extend our approach to handle multi-stable systems of SDEs as in Archambeau *et al*. (2007). Another interesting extension would be to consider hierarchical structures in transcriptional regulatory networks, such as feed-forward loops or auto-regulatory motifs (Alon, 2006).

ACKNOWLEDGEMENTS

We thank Prof. Jeff Green (Molecular Biology and Biotechnology, University of Sheffield) for useful comments on a draft manuscript.

Funding: Engineering and Physical Sciences Research Council (grant GR/S84347/01 to G.S.).

Conflict of Interest: none declared.

REFERENCES

Alon,U. (2006) *An Introduction to Systems Biology*. Chapman and Hall, London.

Archambeau,C. *et al*. (2007) Gaussian process approximations of stochastic differential equations. *J. Mach. Learn. Res. Workshop Conf. Proc.*, **1**, 1–16.

Barenco,M. *et al*. (2006) Ranked prediction of p53 targets using hidden variable dynamical modelling. *Genome Biol.*, **7**, R25.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer, Singapore.

Cover,T. and Thomas,J. (2006) *Elements of information theory*. Wiley, New York.

Demin,O. and Goryanin,I. (2008) *Kinetic Modelling in Systems Biology*. CRC press, New York.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hu,H. and Li,X. (2007) Transcriptional regulation in eukaryotic ribosomal protein genes. *Genomics*, **90**, 421–423.

Hu,Z. *et al*. (2007) Prediction of synergistic transcription factors by function conservation. *Genome Biol.*, **8**. R257.

Khanin,R. *et al.* (2007) Statistical reconstruction of transcription factor activity using Michaelis–Menten kinetics. *Biometrics*, **63**, 816–823.

Lawrence,N.D. *et al*. (2006) Modelling transcriptional regulation using Gaussian processes. In Scholkopf,B. *et al*. (eds) *Advances in Neural Information Processing Systems 19*, Vancouver.

Lawrence,N.D. *et al*. (eds) (2010) *Learning and Inference in Computational Systems Biology*. MIT Press, Cambridge, MA.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Liao,J.C. *et al*. (2003) Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA*, **100**, 15522–15527.

Opper,M. and Sanguinetti,G. (2007) Variational inference for Markov jump processes. In Platt,J. *et al*. (eds) *Advances in Neural Information Processing Systems 20*, Vancouver.

Partridge,J.D. *et al*. (2007) Transition of *Escherichia coli* from aerobic to micro-aerobic conditions involves fast and slow reacting regulatory components. *J. Biol. Chem.*, **282**, 11230–11237.

Ptashne,M. and Gann,A. (2002) *Genes and Signals*. CSH press, Cold Spring Harbor.

Rogers,S. *et al*. (2007) Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics*, **8** (Suppl. 2).

Sabatti,C. and James,G.M. (2006) Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, **22**, 739–746.

Sanguinetti,G. *et al*. (2006) Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, **22**, 2775–2781.

Sanguinetti,G. *et al*. (2009) Switching regulatory models of cellular stress response. *Bioinformatics*, **25**, 1280–1286.

Schawalder,S.B. *et al*. (2004) Growth-regulated recruitment of the essential yeast ribosomal protein gene activator Ifh1. *Nature*, **432**, 1058–1061.

Tu,B.P. *et al*. (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, **310**, 1152–1158.

Wang,W. *et al*. (2005) Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl Acad. Sci. USA*, **102**, 1998–2003.