

# Reconstructing transcription factor activities in hierarchical transcription network motifs

Andrea Ocone and Guido Sanguinetti\*

School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK

Associate Editor: Alex Bateman

## ABSTRACT

**Motivation:** A knowledge of the dynamics of transcription factors is fundamental to understand the transcriptional regulation mechanism. Nowadays, an experimental measure of transcription factor activities *in vivo* represents a challenge. Several methods have been developed to infer these activities from easily measurable quantities such as mRNA expression of target genes. A limitation of these methods is represented by the fact that they rely on very simple single-layer structures, typically consisting of one or more transcription factors regulating a number of target genes.

**Results:** We present a novel statistical inference methodology to reverse engineer the dynamics of transcription factors in hierarchical network motifs such as feed-forward loops. The approach we present is based on a continuous time representation of the system where the high-level master transcription factor is represented as a two state Markov jump process driving a system of differential equations. We solve the inference problem using an efficient variational approach and demonstrate our method on simulated data and two real datasets. The results on real data show that the predictions of our approach can capture biological behaviours in a more effective way than single-layer models of transcription, and can lead to novel biological insights.

**Availability:** <http://homepages.inf.ed.ac.uk/ganguin/software.html>

**Contact:** [g.sanguinetti@ed.ac.uk](mailto:g.sanguinetti@ed.ac.uk)

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

Received on March 23, 2011; revised on June 24, 2011; accepted on August 18, 2011

## 1 INTRODUCTION

Unraveling the regulation mechanisms of gene expression is a fundamental problem in systems biology. Recent years have seen tremendous advances towards this goal, driven by technological advances in experimental techniques as well as a systematic application of mathematical modelling. High-throughput techniques, such as microarrays and chromatin immunoprecipitation (ChIP), have uncovered much important information on gene expression profiles and the architecture of biological transcriptional networks. Combining this information in predictive mathematical models can yield precious insights in the way biological systems work. Nonetheless, in order to understand and predict mechanistically the behaviour of a transcriptional control system,

an understanding of the dynamics of transcription factors' (TFs) response to environmental signals is essential. Nowadays, an experimental measure of TFs activities represents a challenge, since TFs are often present at very low concentrations and they are frequently post-transcriptionally regulated through allosteric changes. This motivated many groups to develop quantitative statistical models in order to infer activities of TFs by combining mRNA expression measurements of their target genes with data about the architecture of the regulatory network, usually obtained from ChIP-on-chip experiments.

Broadly speaking, two different classes of approach have been proposed: simplified models of large transcriptional networks and detailed kinetic models of small subnetworks. The first type of models are usually discrete time models that use log-linear approximations to infer the activities of hundreds of TFs from thousands of target genes (Liao *et al.*, 2003; Sabatti and James, 2006; Sanguinetti *et al.*, 2006). While recent advances (Asif and Sanguinetti, 2011) can model non-linear interactions between transcription factors on a genome-wide scale, these models are still unable to capture the dynamics of gene expression, and rely on steady-state assumptions. While these models do provide useful insights in biological processes (McLean *et al.*, 2010; Partridge *et al.*, 2007), their simplified nature means that many of the subtleties of regulation will inevitably be lost. The other class of approaches, to which this work belongs, adopts a more realistic model of the dynamics of transcriptional regulation based on ordinary differential equations (ODEs) and then infers the profile of the TFs directly from a continuous time representation of the system. This more faithful representation of the regulatory mechanism, however, comes at a higher computational cost so that inference in this class of models has so far been possible only on restricted transcriptional networks with simple single-layer architectures. In particular, all methods we are aware of consider simple networks with a single layer of unobserved TFs. In most cases, inference is restricted to the single-input module (SIM) network motif, which is composed of a number of target genes regulated by a single TF (Gao *et al.*, 2008; Khanin *et al.*, 2006; Lawrence *et al.*, 2006; Rogers *et al.*, 2007; Sanguinetti *et al.*, 2009) or at most a few TFs that jointly regulate a number of target genes (Oppen and Sanguinetti, 2010). While these simple models constitute a strong proof of principle of the methodology, and may indeed be useful in specific situations (Honkela *et al.*, 2010), many important information processing functions in cells are carried out through hierarchical motifs which entail multiple stages of transcriptional regulation.

In this article, we present an ODE-based inference methodology for the most fundamental hierarchical transcriptional network

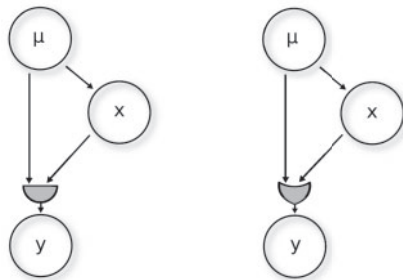
\*To whom correspondence should be addressed.

structure, the feed-forward loop (FFL) network motifs. These network motifs consist of a master TF that directly regulates (transcriptionally) a slave TF; both master and slave TF then control the expression of (a number of) target genes, possibly with non-linear interactions at the target promoters. They are frequently encountered in transcriptional regulatory networks due to their robustness and important functions in biological signal processing, such as filtering biological noise fluctuations (Mangan and Alon, 2003). Predicting the dynamics of these fundamental circuits is trivial if we know the parameters of the models and the activity profile of the master TF. Solving the reverse problem of inferring the master TF activity and model parameters from observations is instead difficult, due to the inevitable non-linearities in these circuits. We use the variational Bayesian approach for approximate inference in continuous-time stochastic processes (Oppen and Sanguinetti, 2007), using a two state Markov jump process as prior distribution over the master TF activity (Sanguinetti *et al.*, 2009). The slave TF is assumed to be transcriptionally regulated and we use a logical approximation so that it becomes active when its concentration crosses a critical threshold (inferred from the data).

We test the model extensively on simulated data to assess its identifiability, reporting accurate continuous time inference of TFs activities, good fitting to data and parameters estimation. We then apply the model on two real datasets: a study of the tumour suppressor protein p53 (Barenco *et al.*, 2006), which was previously used as a benchmark for ODE-based inference models (Barenco *et al.*, 2006; Lawrence *et al.*, 2006; Wang and Tian, 2010), and a time-course experiment of *Escherichia coli* undergoing a transition from aerobic to anaerobic environment (Partridge *et al.*, 2007). We show that our approach can be more effective at predicting independent validity experiments than existing single-layer approaches, as well as being a useful tool for producing novel testable biological hypothesis.

## 2 MODEL AND METHODS

We consider a FFL consisting of a master TF (whose binary activity state<sup>1</sup> is denoted as  $\mu$ ), a slave TF (whose protein expression we denote as  $x$ ) and a target gene whose mRNA expression we denote as  $y$ . The regulation of the target gene is given by a combination of master and slave TFs activities through a logic OR or AND gate. A graphical representation of the network is given in Figure 1. Usually, the master TF functions as a



**Fig. 1.** FFLs network motifs: OR gate FFL (right) and AND gate FFL (left).

<sup>1</sup>A binary representation for both master and slave TF activities is motivated by the fact that transitions between protein states are very rapid (Alon, 2006).

sensor for environmental signals; to enable rapid reactions, many TFs have evolved to function as binary switches (active/inactive) triggered by fast post-translational modification (e.g. a phosphorylation) (Alon, 2006). We can capture the switch-like behaviour of the master TF by modelling it as a two state Markov jump process (*telegraph process*) (Sanguinetti *et al.*, 2009). In contrast, we assume that the slave TF is transcriptionally regulated, so that its activity is a (non-linear) function of its protein abundance. To allow for saturation effects, we model the regulation of the target gene by the slave TF by a logical function of the slave TF's protein (Alon, 2006). Mathematically, the model with OR gate is described by the following equations<sup>2</sup>:

$$\frac{dx(t)}{dt} = A_1\mu(t) + b_1 - \lambda_1 x(t), \quad (1)$$

$$\frac{dy(t)}{dt} = A\mu(t) + b - \lambda y(t) + A_2\Theta[x(t) - c]. \quad (2)$$

Here  $\Theta$  represents the Heaviside step function and  $c$  represents a critical threshold of slave TF protein concentration, such that its regulatory effect becomes important only when its mRNA concentration exceeds  $c$ . Equation (1) refers to the regulation of the slave TF protein  $x$  by the master TF. Here,  $\mu(t)$  is the activity of the master TF, given a telegraph process prior;  $A_1$  is the sensitivity of the gene encoding  $x$  for the master TF;  $b_1$  and  $\lambda_1$  represent the basal transcription rate and decay rate of the slave TF  $x$ , respectively. In practice, we seldom have access to protein measurements of the slave TF, so we will use its mRNA concentration as a proxy for its protein concentration. This is a potentially fraught assumption, as it effectively assumes that the slave TF is transcriptionally regulated *and* that the protein equilibrates fast compared with the sampling interval in our experimental design. These assumptions are not always met but can be sometimes justified, if only as a rough approximation in the absence of better data. Equation (2) is similar to the Equation (1) but it contains an additional term  $A_2\Theta[x(t) - c]$ , which takes into account the regulation of the target gene by the slave TF. Transcription of target gene  $y$  takes into account of the effect of  $x$  only if  $x(t) > c$ ; therefore, the Heaviside function  $\Theta[x(t) - c]$  represents the binary activation state of slave TF. The model can perform both activation or repression over the three edges of the FFL motif, depending on the value of sensitivity parameters  $A_i$ . In addition, it can collapse into a SIM network motif when the parameter  $A_2$  is null.

In the AND gate FFL, the TF inputs on the target gene are multiplied. In this case, Equation (2) of the model is described by

$$\frac{dy(t)}{dt} = A\mu(t)\Theta[x(t) - c] + b - \lambda y(t). \quad (3)$$

Also the AND gate FFL can collapse into a SIM if the threshold  $c$  is null, with  $y$  becoming a target gene only of the master TF.

Mathematical simulation of the above ODEs is trivial *given* the profile of the master TF and the model parameters. The problem we wish to address is the reverse problem: given mRNA observations of both the target gene and the slave TF, can we reconstruct both master and slave TF activity profiles (state inference) and the parameters of the model (parameters estimation)? This is an analytically intractable problem due to the strong non-linearity introduced by the Heaviside function in Equations (2) and (3); we next describe an approximate procedure that can effectively and efficiently handle the problem. Briefly, we compute an approximate variational free energy by introducing two main approximations: the posterior process for the master TF is assumed to be Markovian, and the induced process on the slave TF is approximated as Gaussian using a Laplace-type approximation. This enables us to derive a gradient for the free energy, which can then be used in a gradient descent algorithm.

### 2.1 Inference

The telegraph process representing the activity of the master TF is a continuous time stochastic process that switches with transition rates  $f_{\pm}(t)$

<sup>2</sup>Our definition of OR gate is slightly different from the usual Boolean one; we use OR as a synonym of linearly additive response.

between an ON state and an OFF state. The single time marginal, which represents the probability to be at a given time point in a certain state, is given by the following master equation:

$$\frac{dp_1(t)}{dt} = -f_-(t)p_1(t) + f_+(t)p_0(t) \quad (4)$$

$$\frac{dp_0(t)}{dt} = -f_+(t)p_0(t) + f_-(t)p_1(t), \quad (5)$$

where  $p_1(t)$  and  $p_0(t)$  represent the marginal probability that at time  $t$  the process is in ON and OFF state, respectively. Notice that by means of the telegraph process we have turned our model, described by deterministic ODEs, into a stochastic model that can be studied in terms of probability distributions.

We assume we have mRNA observations corrupted by i.i.d. zero mean Gaussian noise. In particular, the probability to observe the value  $\hat{y}_i$  if the true value is  $y_i$  is given by

$$p(\hat{y}_i|y_i) = \mathcal{N}(\hat{y}_i|y_i, \sigma_i). \quad (6)$$

The same noise model (possibly with different noise variance) is assumed also for the slave TF mRNA observations,  $\hat{x}$ . By means of Bayes' theorem, we combine the prior distribution over the master TF with the likelihood, in order to compute the posterior distribution of the process  $\mu_{0:T}$ , given the observations  $\hat{x}$  and  $\hat{y}$ :

$$p_{\text{post}}(\mu_{0:T}|\hat{x}, \hat{y}) = \frac{1}{Z} p(\hat{x}|\mu_{0:T}) p(\hat{y}|\mu_{0:T}, \hat{x}) p_{\text{prior}}(\mu_{0:T}|f_{\pm}). \quad (7)$$

This defines the posterior distribution of the master TF activity for all time points ( $\mu_{0:T}$ ) as a measure over the space of continuous time paths of the  $\mu$  process. However, as remarked above, direct computation from Equation (7) is not possible; we adopt a variational approach to solve the problem. Variational inference is widely used in machine learning (Bishop, 2006): it relies on a lower bound on the log marginal likelihood to find the optimal approximate posterior within a given family of processes. In our case, the family of approximating processes will be the family of inhomogeneous telegraph processes; we will denote the transition rates of the approximating telegraph process as  $g_{\pm}(t)$ , emphasizing their dependence on time. It can be shown that the variational approximation is equivalent to minimizing the *Kullback–Leibler (KL) divergence* of the approximating process  $q(\mu_{0:T})$  with the true posterior  $p(\mu_{0:T})$ :

$$KL[q||p_{\text{post}}] = \int dq \log \frac{q}{p_{\text{post}}}. \quad (8)$$

The KL divergence is a functional of the transition rates  $g_{\pm}$  of the approximating process, which in turn determine its single time marginals by Equations (4) and (5). The state inference problem, therefore, turns into an optimization problem (in an infinite dimensional space).

In addition, since the KL divergence is also a function of the model parameters, parameter learning can be easily done through optimization. Therefore, state inference is carried out in parallel with parameters estimation. It is possible to include prior distributions for the parameters as well to obtain a fully Bayesian (approximate) inference framework; however in the following, we will place a prior distribution only on the critical threshold  $c$ , for which we define a Gaussian prior distribution centred at half of  $x$  gene expression.

## 2.2 Optimization

By plugging Equation (7) into the KL divergence (8), we obtain the following expression:

$$KL[q||p_{\text{post}}] = -E_q[\ln \prod_{i=1}^N p(\hat{x}_i|x_i) + \ln \prod_{i=1}^N p(\hat{y}_i|y_i)] + KL[q||p_{\text{prior}}] + \ln Z. \quad (9)$$

The second term on the right-hand side of Equation (9) represents the KL divergence between two telegraph processes: the KL divergence between two

general Markov jump processes has been derived by Oppen and Sanguinetti (2007), and it can be easily computed in terms of process rates and single time marginals of the approximating process. The difficulty lies in computing the likelihood terms in (9): these contain the first and second moment of both the variables  $x(t)$  and  $y(t)$  under the marginals of the approximating distribution  $q(\mu(t))$ .

It is important to remark that an exact minimization of the KL functional is impossible as it would involve calculating intractable expectations of the Heaviside function  $\Theta[x(t) - c]$  (which arise in the likelihood terms involving the target mRNA  $y$ ). These, in principle, involve computing all the moments of the slave TF  $x$ . In practice, we resort to a Laplace-type approximation assuming that  $x(t)$  is normally distributed, so that

$$\langle \Theta[x(t) - c] \rangle \simeq \int_c^\infty \mathcal{N}(x(t)|\langle x(t) \rangle, \langle x(t)^2 \rangle - \langle x(t) \rangle^2) dx(t). \quad (10)$$

In practice, in order to minimize Equation (9) with respect to  $g_{\pm}$ , it is more convenient to work in an extended space where the KL divergence is considered to be a functional of the single time marginals  $q(t)$  also. Naturally, process rates and single time marginals are not independent, but they are linked by the master equation. Furthermore, the KL divergence is a functional of the first and second moments of the observed slave gene ( $\langle x \rangle$  and  $\langle x^2 \rangle$ ) and target gene ( $\langle y \rangle$  and  $\langle y^2 \rangle$ ) as well. These moments are also related to transition rates  $g_{\pm}$  and single time marginals by a number of ODEs. So the problem turns into a constrained optimization problem. This can be solved by adding Lagrange multipliers to the KL divergence and performing an efficient gradient descent method based on solving ODEs forward and backward in time.

The derivation of the optimization method is given in the Supplementary Material. Parameter estimation is obtained in parallel through the same strategy.

## 2.3 Heaviside step moments

For the inference in FFL models, we have to consider  $y(t)$  and most of the problems arise from the presence of the Heaviside step function  $\Theta[x(t) - c]$ . In order to compute the expectation of  $y(t)$  under the approximating process  $q(t)$ , we have to compute the expectation of the Heaviside step function. This is a non-trivial problem, since the argument of the Heaviside step function contains  $x(t)$ ,

$$x(t) = e^{-\lambda_1 t} \left[ x(0) + \int_0^t e^{\lambda_1 s} (A_1 \mu(s) + b_1) ds \right], \quad (11)$$

which in turn contains the whole history of the stochastic process  $\mu(t)$ . We address the problem by considering that the Heaviside step function is a deterministic function whose values belong to  $\{0, 1\}$  depending on the sign of its argument being negative or positive, respectively. Therefore, the expectation of the Heaviside step function is the cumulative distribution function:  $\langle \Theta[x(t) - c] \rangle = P(x(t) > c)$ . We can use a Laplace-type approximation for the distribution of  $x(t)$  giving Equation (10). In other words, we have approximated the probability density function of  $x(t)$ ,  $p(x(t))$ , as a Gaussian distribution with first moment and variance of the process  $x(t)$ , as its mean and variance, respectively. The rationale behind this approximation is that the process  $x(t)$  is proportional to a sum over many steps of the stochastic process  $\mu(s)$  [see Equation (11)]. As a consequence of the central limit theorem, the probability distribution over  $x(t)$  tends to be Gaussian with the increase of the switching rate.

By means of this approximation, the expectation of the Heaviside step function can be computed through the *error function*.

For the first moment of  $y(t)$  in the AND gate FFL, the situation is even more complex since we have to compute the expectation of the quantity  $\mu(s)\Theta[x(s) - c]$ . In this case, it makes sense to consider the two processes independent, since the Heaviside step function follows the behaviour of  $x(s)$ , whose time scale is described by slower dynamics compared with the switching process  $\mu(s)$ . This results in the following approximation  $\langle \mu(s)\Theta[x(s) - c] \rangle \simeq \langle \mu(s) \rangle \langle \Theta[x(s) - c] \rangle$ , which can be solved taking advantage of the previous Laplace approximation.

Another expectation that is needed is the autocorrelation function of the process  $\Theta[x(t)-c]$ ,  $\langle\Theta[x(s)-c]\Theta[x(t)-c]\rangle$ . In this case, it is possible to show that the value of autocorrelation function for the process  $x(t)$  decreases exponentially with time interval  $t-s$  and decay constant  $\lambda_1$ . As a consequence, since the Heaviside step function codomain is defined in the range  $\{0, 1\}$ , we can assume that

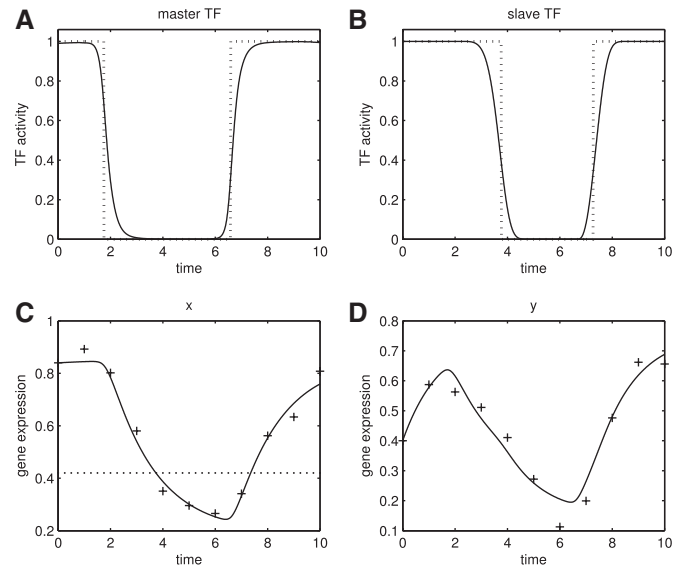
$$\langle\Theta_s\Theta_t\rangle = \langle\Theta_s\rangle + [\langle\Theta_s\rangle\langle\Theta_t\rangle - \langle\Theta_s\rangle] \cdot (1 - e^{-\lambda_1(t-s)}), \quad (12)$$

where we have used the short notation  $\Theta_t$  and  $\Theta_s$  for  $\Theta[x(t)-c]$  and  $\Theta[x(s)-c]$ , respectively. In simple words, it means that autocorrelation is equal to  $\langle\Theta_s\rangle$  when  $s=t$ , and decreases to the product of the expectations  $\langle\Theta_s\rangle\langle\Theta_t\rangle$  as the two Heaviside step functions become uncorrelated.

### 3 RESULTS

To benchmark our model and assess the quality of the approximations, we tested the FFL models on simulated data. We use the same experimental set up as in Sanguinetti *et al.* (2009). Observations  $\hat{x}$  and  $\hat{y}$  are given by adding Gaussian noise with SD of 0.03 (additive noise of relative magnitude  $\sim 5\%$ ) to 10 discrete time points drawn from the model equations with a given TF activity (input) and known parameters. The master TF is initially on, then switches off and eventually switches on towards the end of the period of interest. The parameters regulating the slave TF,  $x(t)$ , are  $A_1 = 3.7 \times 10^{-3}$ ,  $b_1 = 8 \times 10^{-4}$ ,  $\lambda_1 = 5 \times 10^{-3}$ . The parameters regulating the target gene are instead the following:  $A = 2.7 \times 10^{-3}$ ,  $b = 8 \times 10^{-4}$ ,  $\lambda = 8 \times 10^{-3}$ . The additional parameter for the OR gate FFL is  $A_2 = 2.5 \times 10^{-3}$  and finally we set the critical threshold to a value which allows different regulation areas to contain a sufficient number of observation time points (in this case,  $c = 0.42$ ). Notice that when we set also the threshold  $c$ , then the process  $\Theta[x(t)-c]$  is known; therefore, it is possible to evaluate the accuracy of the inference not only for the master TF activity  $\mu(t)$ , but also for the slave TF activity  $\Theta[x(t)-c]$ . Parameter estimation is performed in parallel with state inference during gradient descend optimization. However, the critical threshold  $c$  is harder to evaluate, as it appears within the non-differentiable function  $\Theta[x(t)-c]$ . Therefore, we learned the critical threshold off-line, by selecting the value that minimizes the *variational free energy* among a set  $N_c$  of discrete values used to run different simulations. Figure 2 shows the inferred posterior master (Fig. 2A) and slave (Fig. 2B) TF activities, compared with the true TF activities. The posterior slave TF activity is defined as  $\langle\Theta[x(t)-c]\rangle_q$ . As can be observed, the method gives an excellent reconstruction of both TF activities, with a precise prediction of the correct transition times and an appropriate switch-like behaviour. Figure 2C and D show the posterior first moment of  $x$  and  $y$ , with observations. Again, the model gives a credible fit to the data, giving confidence that a good optimum of the free energy has been reached. Further results on simulated data, including parameter estimates across several repetitions, are shown in the Supplementary Material.

To assess the robustness of the method, we performed 150 simulations on different synthetic datasets obtained by changing the values of the parameters over an order of magnitude. The results show a consistent estimation, giving correlation coefficients between the vectors of true and inferred parameters of  $\sim 0.95$  for both OR gate FFLs and AND gate FFLs. Full details are given in the Supplementary Material. We also investigated the robustness of the model against mismatch in the noise model. To do so, we generated synthetic data from the model and added Gamma distributed noise,



**Fig. 2.** Results on simulated data. (A) Inferred posterior master TF activity (solid) versus true input impulse (dashed). (B) Inferred posterior slave TF activity (solid) versus true input impulse (dashed). (C) Posterior first moment of  $x$  (solid) and observations of  $x$  (crosses). True critical threshold (dashed). (D) Posterior first moment of  $y$  (solid) and observations of  $y$  (crosses).

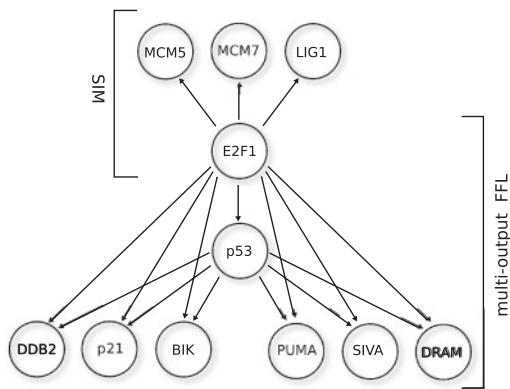
and run our algorithm (which assumes Gaussian noise). In our experiments, we found that this did not make a significant difference, and the model was still capable of giving a good reconstruction of TF activity (see Supplementary Material).

#### 3.1 Inference of p53 activity in human leukaemia cell line

As a first real dataset, we tested our model on the p53 prediction task, which has been already tackled by a number of authors with ODE-based models using a SIM architecture (Barenco *et al.*, 2006; Lawrence *et al.*, 2006). This dataset focuses on the tumour suppressor protein p53. Data were obtained from three independent replicates and represent gene expression at seven discrete time steps from an irradiated human leukaemia cell line. After the irradiation, tumour suppressor protein p53 increases its activity in order to trigger genes whose function is to protect the cell and eventually induce its apoptosis. An attractive feature of this dataset is the existence of a (semi-quantitative) experimental measure of the p53 activity, through a western blot analysis; this allows us to validate, at least qualitatively, the predictions of the model. We compare our predictions with results obtained by Barenco *et al.* (2006) using their HVDM model, which pre-supposes a SIM architecture; results using different inference methodologies (but still a SIM architecture) do not differ qualitatively (Lawrence *et al.*, 2006).

Previous models based on a SIM structure could not reproduce the p53 experimental measurements (Barenco *et al.*, 2006; Lawrence *et al.*, 2006), predicting a decrease of the p53 activity after about 5 h time, rather than the experimentally measured 10 h time. To explain the discrepancy, an unknown mechanism of migration of a part of the p53 protein from the nucleus to the cytoplasm was postulated, but no experimental evidence is available to support this mechanism.





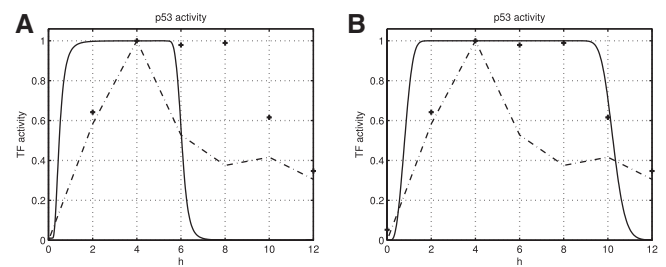
**Fig. 3.** p53 network architecture. E2F1 is the master TF, p53 is the target TF and both regulate target genes *DDB2*, *p21*, *BIK*, *PUMA*, *SIVA*, *DRAM*. Target genes of the only E2F1 (*MCM5*, *MCM7*, *LIG1*) have been included.

While a careful selection of the p53 target genes may moderate the problem (Wang and Tian, 2010), we are interested in testing whether the discrepancy may arise from model inadequacy. In particular, it is known that p53 belongs to a modified FFL where another important transcriptional regulator, E2F1, controls the expression of p53 and directly common p53 targets (Polager and Ginsberg, 2009). This represents a FFL where E2F1 is the master TF and p53 is the slave TF (Fig. 3).

The target genes are *DDB2*, *p21*, *BIK*, a subgroup of the genes used in Barenco *et al.* (2006) which are regulated by p53 and E2F1 (Chinnadurai *et al.*, 2009; Hiyama *et al.*, 1998; Prost *et al.*, 2006); in addition, we use the apoptotic genes *PUMA*, *SIVA* and the gene *DRAM*, involved in the process of autophagy, regulated again by both p53 and E2F1 (Polager and Ginsberg, 2009).<sup>3</sup> We infer the activity of both E2F1 and p53 from the observation of the expression of these common target genes and p53 itself. We run two different simulations, first using only the SIM part of the model to infer the activity of p53 from its target genes and then using the whole FFL model with also another SIM part regarding other genes (*MCM5*, *MCM7*, *LIG1*) regulated only by E2F1 (Bracken *et al.*, 2004). In the former case, the p53 activity is qualitatively similar to the one predicted by Barenco *et al.* (Fig. 4A). On the contrary, inclusion of the FFL structure leads to a completely different prediction of p53 activity, which fits the experimental measure much better (Fig. 4B).

Inference of E2F1 is also obtained but it is not as interesting as the posterior p53 activity, since we do not have an experimental validation for it. Instead, it is interesting that, on this particular real biological dataset, both an AND and an OR FFL seem capable of fitting the data reasonably (no strong evidence was found using standard model selection heuristics such as BIC or AIC). Experimental evidence derived from a knock-out of p53 shows that the common target genes of p53 and E2F1 are still activated, indicating that master TF E2F1 and slave TF p53 cannot be combined into a pure AND gate (Polager and Ginsberg, 2009). It is important to remark that the assumption that the activity of the slave TF is transcriptionally regulated is violated in this case: in fact, it is

<sup>3</sup>Other apoptotic and non-apoptotic genes mentioned in Polager and Ginsberg (2009) have not been included due to their highly noisy expression profiles.

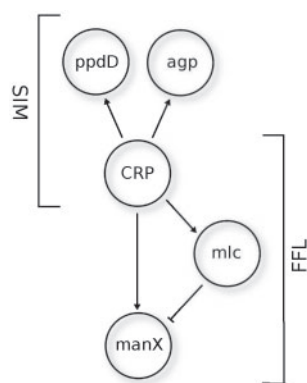


**Fig. 4.** Results on p53 dataset. (A) Posterior inferred p53 activity using our SIM model (solid), compared with Barenco *et al.*'s prediction (dashed) and experimental measurements (crosses). (B) Posterior inferred p53 activity using our FFL model (solid), compared with Barenco *et al.*'s prediction (dashed) and experimental measurements (crosses).

known that p53 activity is post-translationally regulated by multi-phosphorylation (Lee *et al.*, 2010) and E2F1 does not directly affect the transcription of p53. Nonetheless, E2F1 directly regulates the transcription of other genes that either phosphorylate or interact with p53, resulting in a stimulation of its apoptotic activity. Therefore, we may assume that that E2F1's activity results in a more efficient activation of p53 protein, thus justifying the assumption that slave TF expression is a good proxy for its activity. However, the lack of a direct regulation of p53 expression by E2F1 means that the physical meaning of the parameters  $A_1$ ,  $b_1$  and  $\lambda_1$  is unclear in this particular case. What our results in fact show is that the inclusion of a hierarchical structure does lead to a considerable improvement in the prediction. Intuitively, this could be explained by the model using the flexibility given by the threshold parameter  $c$  to adjust the mRNA profile (in this case a suboptimal proxy) to be as close as possible to the true slave TF activity profile.

### 3.2 Sugar foraging in *E.coli* during aerobic-anaerobic transition

As a further example of the applicability of our methodology, we considered a study of the transcriptional response of *E.coli* during oxygen withdrawal. FFLs are particularly over-represented in bacterial transcriptional networks (Alon, 2006). Furthermore, there is experimental evidence that during transitions *E.coli* adopts a nutrient foraging strategy based on hierarchical transcriptional regulation (Liu *et al.*, 2005). We therefore focus on a subnetwork composed of the cAMP receptor protein (CRP) as master transcriptional regulator, the major regulator of carbon catabolism in *E.coli*. Among the CRP target genes, we consider three genes (*manX*, *manY* and *manZ*) belonging to the PEP-dependent sugar transporting PTS system. The essential function of the *manXYZ* operon is to make sugar substrates available for metabolism, principally through glycolysis. In addition to the positive regulation by CRP, *manXYZ* is subject to a negative regulation by Mlc, another global regulator of bacterial metabolism. *mlc* is in turn positively regulated by CRP, leading to the FFL structure shown in Figure 5 (Görke and Stülke, 2008). The dataset contains *E.coli* global gene expression measurements taken during the transition from aerobic to anaerobic condition, at initial time and four successive discrete time points (5, 10, 15 and 60 mins) (Partridge *et al.*, 2007). The expression profiles of these genes, as well as the model fits, are shown in the Supplementary Material.



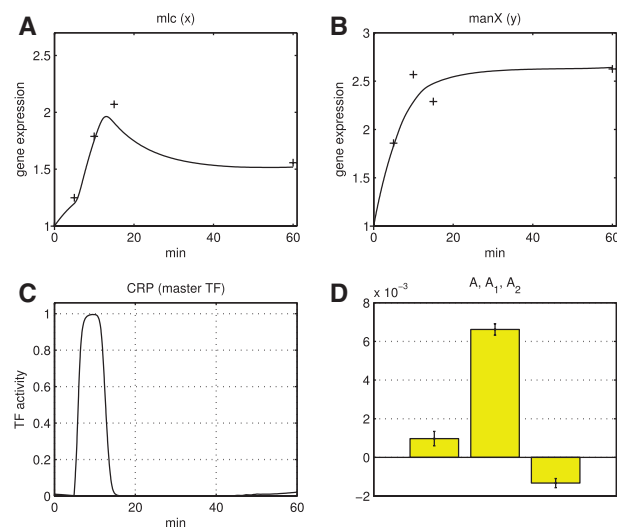
**Fig. 5.** CRP network architecture. CRP represents the master TF, Mlc represents the slave TF and both regulate target genes *manXYZ* (here only *manX* is showed). Target genes of the only CRP (*ppDD* and *agp*) have been included too (Keseler *et al.*, 2009).

Using microarray time courses from target genes *manX*, *manY* and *manZ*, we infer the activity of Mlc and more interestingly the activity of CRP using an OR gate FFL. Posterior expectations give a good fit of the (noisy) target genes expressions (Fig. 6A and B); in addition, we note that the inferred parameters  $A$  and  $A_1$  are positive (Fig. 6D), whereas parameter  $A_2$  is negative, representing inhibition, which is in accordance with existing knowledge about the sign of regulation in this network. In Figure 6C, we show posterior activity of CRP: it is interesting to note that this activity rapidly transits to the active state and after a short interval it goes back to the inactive state. This behaviour can be rationalized using existing knowledge of the regulation of PEP by the two pyruvate kinases of *E.coli*, whose different temporal behaviour may induce a transient response of CRP. In any case, it directly leads to a testable prediction on the dynamics of cAMP during aerobic–anaerobic transitions, which may lead to insights in the survival strategy of *E.coli* under stress. While this hypothesis clearly needs experimental validation, it gives a useful example of the kind of insights this type of modelling can offer.

## 4 CONCLUSIONS

Differential equation models of gene regulation have been enormously successful in the last decade in providing a flexible predictive tool for systems biology. However, their effectiveness depends crucially on the availability of model parameters and external inputs (in the case of non-autonomous systems). Often, these are simply not available, motivating the need for integrating statistical inference tools in ODE-based methodologies. Methodologies for statistical inference in these systems are difficult to develop, but can often offer insights in the underlying biological system which are complementary to those gained by traditional mechanistic modelling.

In this article, we introduced an approximate inference methodology for ODE-based models of FFLs driven by an unobserved binary TF. While FFLs are very simple structures, they can be viewed as the fundamental building block for hierarchical regulatory networks; from the methodological point of view, there are no significant conceptual difficulties to extending the proposed



**Fig. 6.** Results on *E.coli* dataset. (A and B) Shows posterior inferred mean of *manX* and *mlc* (solid) and observations (crosses). (C) Shows posterior inferred CRP activity. (D) Shows estimated parameters of the FFL architecture.

approach to more complex multi-layer networks. Inference of TF activity in ODE-based models has recently attracted considerable attention, and several tools now exist for network motifs involving a single layer of hidden TFs (Barenco *et al.*, 2006; Lawrence *et al.*, 2006; Oppen and Sanguinetti, 2010; Rogers *et al.*, 2007; Sanguinetti *et al.*, 2009). To the best of our knowledge, similar methodologies for hierarchical networks of transcriptional regulators have not been previously reported; given the importance and frequent occurrence of these motifs, the lack of methodology for inference represented a significant obstacle to the adoption of statistical methodologies in ODE-based models of biological networks. Our contribution means that inference methods are now available for all the fundamental building blocks of transcriptional networks. While we view our contribution as mainly methodological, application to two real datasets shows the potential for this type of methodology to yield consistent predictions and to lead to testable hypotheses in biology.

From the methodological point of view, our method belongs to the class of deterministic variational approximate inference techniques, which originated in the statistical physics literature and have been widely used in the machine learning community in recent years (Bishop, 2006). While these provide often accurate and efficient estimates of the posterior statistics of interest, they are inherently biased by the choice of family of approximating processes (in our case, Markovian processes). An unbiased alternative would be to resort to stochastic approximations such as Markov chain Monte Carlo. While these have been used successfully in stochastic processes with continuous state space (Wilkinson, 2006), it is not obvious how to derive an efficient sampling strategy for a highly non-linear hybrid model as the one we consider here.

There are several further avenues of research in which this work can be extended. Inference in stochastic models of gene regulation, which are capable of capturing the intrinsic fluctuations of fluorescence data, are at the moment only available for SIM architectures (Oppen *et al.*, 2010), and it would be important to

extend them to the FFL case. Extrinsic noise sources contributions have also been considered in a Bayesian inference model of gene regulation (Komorowski *et al.*, 2010). A linear noise approximation is adopted to deal with the mesoscopic nature of the observed fluorescent data. For the different levels of gene expression, they perform a noise analysis, which would be even more relevant but not straightforward for a hierarchical model such as the one presented here.

Another important direction would be to remove the use of the slave TF mRNA as a proxy for its active protein concentration. This is in principle straightforward in the case when the regulation is purely transcriptional (so that the difference between mRNA and protein profile is entirely due to the difference in half-life between protein and mRNA), and can be done by introducing a simple model of translation (Gao *et al.*, 2008).

## ACKNOWLEDGEMENT

We gratefully acknowledge useful discussions with Jeff Green about *E.coli* metabolism, and with Manfred Oppen and Nicolas Oury about stochastic processes.

**Funding:** Scottish Government through the Scottish Informatics and Computer Science Alliance (SICSA) to G.S.

**Conflict of Interest:** none declared.

## REFERENCES

- Alon, U. (2006) *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Champan and Hall, London.
- Asif, H. and Sanguinetti, G. (2011) Large scale learning of combinatorial transcriptional dynamics from gene expression. *Bioinformatics*, **27**, 1277–1283.
- Barenco, M. *et al.* (2006) Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biol.*, **7**, R25.
- Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*. Springer, New York.
- Bracken, A. *et al.* (2004) E2F target genes: unraveling the biology. *Trends Biochem. Sci.*, **29**, 409–417.
- Chinnadurai, G. *et al.* (2009) BIK, the founding member of the BH3-only family proteins: mechanisms of cell death and role in cancer and pathogenic processes. *Oncogene*, **27**, S20–S29.
- Gao, P. *et al.* (2008) Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, **24**, i70–i75.
- Görke, B. and Stülke, J. (2008) Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nat. Rev. Microbiol.*, **6**, 613–624.
- Hiyama, H. *et al.* (1998) Regulation of the cdk inhibitor p21 gene during cell cycle progression is under the control of the transcription factor E2F. *Oncogene*, **16**, 1513–1523.
- Honkela, A. *et al.* (2010) Model-based method for transcription factor target identification with limited data. *Proc. Natl Acad. Sci. USA*, **107**, 7793–7798.
- Keseler, I. *et al.* (2009) EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res.*, **37**, D464–D470.
- Khanin, R. *et al.* (2006) Reconstructing repressor protein levels from expression of gene targets in E.coli. *Proc. Natl Acad. Sci. USA*, **103**, 18592–18596.
- Komorowski, M. *et al.* (2010) Using a single fluorescent reporter gene to infer half-life of extrinsic noise and other parameters of gene expression. *Biophys. J.*, **98**, 2759–2769.
- Lawrence, N. *et al.* (2006) Modelling transcriptional regulation using Gaussian processes. In *Advances in Neural Information Processing Systems 19*. MIT Press.
- Lee, C. *et al.* (2010) Graded enhancement of p53 binding to CREB-binding protein (CBP) by multisite phosphorylation. *Proc. Natl Acad. Sci. USA*, **107**, 19290–19295.
- Liao, J. *et al.* (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA*, **100**, 15522–15527.
- Liu, M. *et al.* (2005) Global transcriptional programs reveal a carbon source foraging strategy by Escherichia coli. *J. Biol. Chem.*, **280**, 15921–15927.
- Mangan, S. and Alon, U. (2003) Structure and function of the feed-forward loop network motif. *Proc. Natl Acad. Sci. USA*, **100**, 11980–11985.
- McLean, S. *et al.* (2010) Peroxynitrite toxicity in *Escherichia coli* K-12 elicits expression of oxidative stress responses, and protein nitration and nitrosylation. *J. Biol. Chem.*, **285**, 20724–20731.
- Oppen, M. and Sanguinetti, G. (2007) Variational inference for Markov jump processes. In *Advances in Neural Information Processing Systems 20*. MIT Press.
- Oppen, M. and Sanguinetti, G. (2010) Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics*, **26**, 1623–1629.
- Oppen, M. *et al.* (2010) Approximate inference for continuous-time Gaussian-jump processes. In *Advances in Neural Information Processing Systems 23*. MIT Press.
- Partridge, J. *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.
- Polager, S. and Ginsberg, D. (2009) p53 and E2f: partners in life and death. *Nat. Rev. Cancer*, **9**, 738–748.
- Prost, S. *et al.* (2006) E2F regulates DDB2: consequences for DNA repair in Rb-deficient cells. *Oncogene*, **26**, 3572–3581.
- Rogers, S. *et al.* (2007) Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics*, **8**(Suppl. 2), S2.
- Sabatti, C. and James, G. (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.
- Wang, J. and Tian, T. (2010) Quantitative model for inferring dynamic regulation of the tumour suppressor gene p53. *BMC Bioinformatics*, **11**.
- Wilkinson, D. (2006) *Stochastic Modelling for Systems Biology*. Champan and Hall, London.