BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 7 2013, pages 910–916
doi:10.1093/bioinformatics/btt069

Systems biology

Advance Access publication February 13, 2013

Hybrid regulatory models: a statistically tractable approach to model regulatory network dynamics

Andrea Ocone[1], Andrew J. Millar[2,3] and Guido Sanguinetti[1,2,*]

[1]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, [2]SynthSys—Systems and Synthetic Biology, University of Edinburgh, Edinburgh EH9 3JD and [3]School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JR, UK

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Computational modelling of the dynamics of gene regulatory networks is a central task of systems biology. For networks of small/medium scale, the dominant paradigm is represented by systems of coupled non-linear ordinary differential equations (ODEs). ODEs afford great mechanistic detail and flexibility, but calibrating these models to data is often an extremely difficult statistical problem.

Results: Here, we develop a general statistical inference framework for stochastic transcription–translation networks. We use a coarse-grained approach, which represents the system as a network of stochastic (binary) promoter and (continuous) protein variables. We derive an exact inference algorithm and an efficient variational approximation that allows scalable inference and learning of the model parameters. We demonstrate the power of the approach on two biological case studies, showing that the method allows a high degree of flexibility and is capable of testable novel biological predictions.

Availability and implementation: http://homepages.inf.ed.ac.uk/gsanguin/software.html.

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

Contact: G.Sanguinetti@ed.ac.uk

Received on September 28, 2012; revised on December 18, 2012; accepted on February 5, 2013

1 INTRODUCTION

Understanding the dynamics of gene regulatory networks (GRNs) is a fundamental area of research in systems biology. *In silico* predictions of the network's response to altered conditions can often give deep insights in the functionality of the biological system under consideration, as well as being crucial in biomedical and biotechnological applications.

Bioinformatics data analysis methods are invaluable in extracting information in large datasets, and can be useful to predict the main changes in regulatory behaviours (Asif and Sanguinetti, 2011; Sanguinetti *et al.*, 2006). However, detailed predictions of the dynamics of small/medium scale complex regulatory networks cannot avoid dealing with the non-linear and continuous time nature of such systems, calling for more sophisticated mathematical modelling techniques. By some distance, the dominant paradigm to model GRNs' dynamics is given by systems of coupled non-linear ordinary differential equations (ODEs). ODEs provide an ideal framework for the detailed modelling of mechanistic systems, and of course can rely on refined analysis tools developed over hundreds of years of mathematical research. Nevertheless, mechanistic detail often comes at the cost of including many unknown parameters, as well as novel variables that are not observed (e.g. post-translational modifications of proteins). Although there are many parameter estimation tools available (Georgoulas *et al.*, 2012; Hoops *et al.*, 2006; Liepe *et al.*, 2010; Vyshemirsky and Girolami, 2008), parameter estimation in systems of non-linear ODEs is often an intrinsically difficult statistical problem owing to the severe multimodality of the likelihood landscape. This is further compounded by the limited amount of data usually available in most biological scenarios.

Here, we propose a novel statistical modelling framework to model regulatory interactions in GRNs, which maintains some key features of non-linear ODE models while being amenable to a principled statistical treatment. Statistical modelling has become increasingly central in systems biology (Lawrence *et al.*, 2010). Many different statistical models have been proposed in the context of mechanistic systems biology models, ranging from ODEs with uncertain parameters to fully stochastic models (Vyshemirsky and Girolami, 2008; Wilkinson, 2011). Naturally, the key question is to select a representation that is complex enough to capture the behaviour of the system, but simple enough to allow tractable inference. Here, we build on recently proposed statistical models for transcriptional regulation (Ocone and Sanguinetti, 2011; Opper and Sanguinetti, 2010; Sanguinetti *et al.*, 2009) and represent GRNs using a hybrid continuous/discrete stochastic process, consisting of binary promoter states (occupied/vacant) that drive a stochastic differential equation describing protein dynamics. In this way, we bypass much of the statistical difficulties introduced by detailed modelling of transcription/translation and subsequent post-translational modifications. On the other hand, the introduction of a latent stochastic promoter state can capture much of this complexity, giving a flexible framework. Our key advance is the introduction of a model of how promoters can depend on upstream protein states, and of a modular approach to approximate inference in this model class that scales linearly with the number of genes in the network. In this way, we can handle medium-sized networks of arbitrary topology. We complement our theoretical analysis with an empirical analysis of our method on simulated data, as well as on two real biological systems: the

*To whom correspondence should be addressed.

benchmark yeast synthetic network IRMA (Cantone *et al.*, 2009), and the circadian clock of the picoalga *Ostreococcus tauri* (Troein *et al.*, 2011). We compare with existing ODE models, and show that our approach achieves excellent fits and robust predictions. By comparing predictions on different data types, our model also provides a new testable hypothesis about the structure of the *O.tauri* clock network.

2 MODEL AND METHODS

Our aim is to obtain plausible, yet statistically tractable, models of the dynamics of transcription–translation networks. A central requirement is therefore to include a plausible model of gene expression at the heart of the framework. In this approach, we use the *on/off model* of gene expression (Ptashne and Gann, 2002), a simple, yet powerful, model where the rate of transcription of a gene can vary between two levels depending on the occupancy of the promoter of the gene. Assuming for simplicity a tight coupling of transcription and translation, we will use the stronger assumption that protein production can also happen at two distinct rates depending on the occupancy of the promoter. Our network models are therefore composed of a number of connected blocks of two separate types, each of them representing a protein node and a promoter state.

It is convenient to adopt a graphical notation for the statistical models. We denote protein states as circles, and promoter states as squares. Measured protein values are denoted by shaded circles, and we will always assume measurements to occur at discrete times with i.i.d. Gaussian noise; promoter states are assumed not to be observed. Figure 5A shows an example of our graphical representation of a two-gene network.

2.1 Promoter model

We model promoters as Markovian continuous time random variables with two possible states, occupied or unoccupied; we denote promoter states as μ and represent them as *telegraph processes*. The time marginal probability $p_{\mu=1}(t) = p_1(t)$ obeys the chemical master equation

$$\begin{aligned}\frac{dp_1(t)}{dt} &= -f_-(t)p_1(t) + f_+(t)p_0(t),\\ \frac{dp_0(t)}{dt} &= -f_+(t)p_0(t) + f_-(t)p_1(t)\end{aligned} \tag{1}$$

where f_+ and f_- are called switching rates. They represent the transition probabilities per unit time for the switching of the promoter state from 0 to 1 and the other way round, respectively. The time marginal probability $p_\mu(t)$ represents the probability of the promoter state to have a certain value (either 0 or 1) at a given time t. For example, the marginal probability $p_1(t)$ is the probability for the promoter state to be 1 at time t.

Naturally, the rate at which a promoter becomes occupied depends on the state (concentration) of upstream proteins which can bind the promoter. Mathematically, we encode this property by enforcing that the switching rates of the telegraph process μ_i are functions of the transcription factor (TF) concentration x_j. As $f^i_\pm$ represent probabilities per unit time, these functions must be always positive. We use a log-linear model for f^i_+, primarily owing to its mathematical convenience for approximate inference (see Supplementary Material). On the other hand, the switching rate f^i_- is set to a positive constant value, reflecting the fact that unbinding of the TF (i.e. switch from state 1 to state 0) does not depend on x_i (Schultz *et al.*, 2007). In formulae we have:

$$f^i_+ = k_p \exp\left(k_e x_j\right) \tag{2}$$

$$f^i_- = k_m \tag{3}$$

where $k_{p,m,e}$ are hyperparameters. Notice that this model implies that the steady-state probability of being bound has a saturating Hill-type dependence on the concentration of protein x (see Supplementary Material).

2.2 Protein model

Protein production is modelled as a stochastic on/off model. We use a continuous approximation to the underlying discrete system and model the transcriptional–translational dynamics through the following stochastic differential equation (SDE):

$$\mathrm{d}x_i = (A_i\mu_i(t) + b_i - \lambda_i x_i)\mathrm{d}t + \sigma \mathrm{d}w(t) \tag{4}$$

where subscript i refers to the target gene and its promoter. Here, $\Theta_i = [A_i, b_i, \lambda_i]$ is the set of kinetic parameters: A_i represents the efficiency of the promoter in recruiting polymerase when occupied. Its sign defines the type of regulation: either activation or repression. Parameter b_i represents a basal transcriptional–translational rate and λ_i is the exponential decay constant for x_i, which is inversely proportional to x_i half-life. Note that Equation (4) is a linear SDE *conditioned on the history of the promoter state*, which entails significant computational efficiency. However, the time-varying nature of the promoter state allows plenty of flexibility to capture non-stationary behaviours. The term $\sigma \mathrm{d}w(t)$, where $w(t)$ is a Wiener process, represents a white noise-driving process with non-zero variance σ^2. This accounts for the presence of intrinsic noise in the protein concentration x_i, whereas the stochastic process μ_i takes into account the extrinsic noise in gene expression (Elowitz *et al.*, 2002; Swain *et al.*, 2002).

2.3 Approximate inference

As the model of promoter and proteins is jointly Markovian, exact inference can be carried out by numerically solving the Chapman–Kolmogorov forward and backward equations along the lines of Sanguinetti *et al.* (2009) (see Supplementary Material). This, however, requires the numerical solution of a system of high dimensional partial differential equations, leading to severe computational problems when parameters need to be estimated or when more than two genes are present in the network.

We therefore adopt an approximate Bayesian approach for the reconstruction of promoter states μ, protein states x and the estimation of model parameters. The quantity we are interested in is the conditional probability distribution $p(\mu, x|\mathbf{y})$, the joint (posterior) probability of the state of the promoter μ and the promoter concentration x at all time points $t = t_0, \ldots, T$, conditioned on the observations $\mathbf{y}$. We compute an approximation to this quantity by minimizing the *Kullback–Leibler* (KL) divergence functional under a restrictive ansatz for the approximating process. We give details for a two-gene network: extension to more genes is straightforward.

The choice of the family of approximating distributions $q(\mu, x)$ is often crucial for computational reasons; here, we use a mean-field approximation (Opper and Saad, 2001). The approximating posterior then factorizes as follows

$$\begin{aligned}&q(x_{20:T}, \mu_{20:T}, x_{10:T}, \mu_{10:T}) =\\ &q_{x_2}(x_{20:T})q_{\mu_2}(\mu_{20:T})q_{x_1}(x_{10:T})q_{\mu_1}(\mu_{10:T})\end{aligned} \tag{5}$$

where we are considering a network with two genes, x_1 and x_2. q_{x_i} and q_{μ_i} represent pure diffusion processes and pure telegraph processes, respectively. By using this factorized distribution, the KL divergence becomes a sum of terms that are analytically computable

$$\begin{aligned}KL[q||p] = \log Z &- \sum_{j=1,2}\sum_{i=1}^{N} \langle \log p(y_{ji}|x_j(t_i))\rangle_{q_{x_j}}\\ &+ \sum_{j=1,2} \langle KL\left[q_{x_j}||p(x_{j0:T}|\mu_{j0:T})\right]\rangle_{q_{\mu_j}}\\ &+ \sum_{\substack{i=1,2\\ j=2,1}} \langle KL\left[q_{\mu_i}||p(\mu_{i0:T}|x_{j0:T})\right]\rangle_{q_{x_j}}.\end{aligned} \tag{6}$$

Here, the last two terms include the KL divergence between two diffusion processes and between two telegraph processes, respectively. The second KL term can be computed using the variational approximation for telegraph processes (Opper and Sanguinetti, 2007); in addition, it requires an expectation with respect to the diffusion process, which involves the prior switching rates Equations (2) and (3). This expectation can be computed exactly, as the diffusion process is approximated with a Gaussian process; the resulting term is linear in the marginals of the μ process, so that the computation of the approximating telegraph processes can be done efficiently using a forward–backward algorithm (Opper *et al.*, 2010). The computation of the approximating diffusion process is more involved, as the expectation of the telegraph switching rates introduce non-linear terms of the form $\langle \exp[x] \rangle$; these terms can, however, still be computed analytically under the Gaussian process ansatz, so that an efficient gradient descent algorithm can be used (see Supplementary Material).

The inference problem in this way becomes a constrained optimization problem: we iterate between minimizing with respect to each factor, and each of these minimizations is carried out exactly using a forward–backward procedure. The posterior inference problem is solved together with the parameter estimation problem, in a variational expectation–maximization style. Our mean-field variational approximation does not require large computational resources and at the same time provides a solution whose quality is comparable with a computationally expensive exact inference method (see Supplementary Material). Full algorithmic and implementation details are reported in Supplementary Material.

2.4 Approximate variational Bayesian scheme

Here, we report an algorithmic description of the iterative procedure for KL minimization; mathematical details are found in the Supplementary Material. The iterative algorithm consists of the following three steps:

(i) computation of the approximating diffusion process;
(ii) computation of the approximating jump process;
(iii) update of the kinetic parameters $\Theta = [A, b, \lambda]$.

In the first step, we need to compute the approximating diffusion process; under the restrictive assumption of Gaussianity, this is equivalent to computing marginal mean $m(t)$ and variance $c^2(t)$ of the process. Again as a consequence of the Gaussian assumption, we know that the diffusion process is governed by a linear SDE with drift $dx = \alpha(t)x + \beta(t)$ (α and β are variational parameters to be optimized). As mentioned above, some of the terms involving $m(t)$ and $c^2(t)$ are non-linear; therefore, we cannot use forward–backward Kalman recursions. Instead, we adopt a gradient descent algorithm and minimize the KL divergence with respect to α and β, subject to constraints involving the approximating moments. This is done by incorporating the constraints, through Lagrange multipliers, into the KL functional. Then solving forward for the moments and backward for Lagrange multipliers, we finally compute the gradients with respect to the variational parameters.

In the second step, we compute the approximating jump process marginals and rates. Inspection of the KL divergence reveals that the marginals are only involved linearly, so that fast forward–backward recursions can be used for these computations (see Supplementary Material).

The KL functional is also a quadratic function of the kinetic parameters; therefore, they can be easily updated using quadratic programming. This provides an estimation for the mean of each parameter. In addition, assuming the parameters are Gaussian distributed, we also get an estimation of the variance of each parameters, which is simply given by the diagonal elements of the inverse Hessian matrix. This information provides a confidence interval for the parameter estimation and can be used to evaluate statistically the goodness of the estimation.

3 RESULTS

In this section, we assess the performance of the hybrid regulatory model on two real datasets. The main features we are interested in are the quality of the fits to the training data (i.e. whether the model has sufficient flexibility to capture the complex behaviour of biological circuits) and the ability to predict unseen data in perturbed conditions (i.e. whether it is able to generalize). Statistically, identifiability is also an important issue: we address this in the Supplementary Material through a study on simulated data, where comparisons with the ground truth and with the results of exact inference show empirically an excellent identifiability. Throughout the results section, the hybrid regulatory model has three free parameters per gene, corresponding to the kinetic parameters in Equation (4). Hyperparameters in the transition rates, as well as the system noise, are fixed to reasonably vague values (their precise identifiability would require longer time series than usually available).

3.1 Modelling the IRMA synthetic yeast network

As a first application, we considered the IRMA network (Cantone *et al.*, 2009), a synthetic network embedded in the yeast *Saccharomyces cerevisiae*. IRMA is composed of five genes: *ASH*1, *CBF*1, *GAL*4, *SWI*5 and *GAL*80. Figure 1 shows a representation of our hybrid regulatory model for the IRMA network, where the interactions between the five genes can be easily detected by looking at the thick black lines. The network was engineered to respond to changes in the sugar supplied (galactose versus glucose). Gene expression from all the five genes was measured during the transitions from glucose to galactose and from galactose to glucose, giving two sets of data that are referred to as switch-on and switch-off time series.

To analyse the dynamics of the IRMA network, we compared two different models: our hybrid regulatory model and the non-linear delay differential equation (DDE) model of Cantone *et al.* (2009). Our model consists of five SDEs with 3×5 free parameters, while the model of Cantone *et al.* consists of five DDEs,

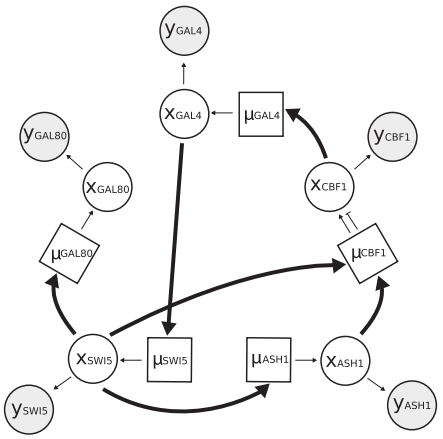

Fig. 1. IRMA yeast synthetic network. Thick black lines, which model promoter activations, show the IRMA network topology. Activation and repression arrows in the transcriptional–translational dynamics between μ_{CBF1} and x_{CBF1} model the fact that *ASH*1 and *SWI*5 are a repressor and an activator of *CBF*1, respectively

modelling the mRNA levels of the five genes, with a total of 31 parameters. Both models were trained by using only the switch-on time series: for the hybrid regulatory model, we adopt the variational Bayesian scheme described above, whereas the Cantone *et al*. model is trained using stochastic optimization (see Supplementary Material). Figure 2 shows the results of this analysis. The left hand column shows the fit to the training switch-on data: the dark grey lines represent the fits of the Cantone *et al*. model, while the light grey lines are the hybrid regulatory model posterior predictions (with confidence intervals). Both models give a qualitatively good fit, with a slightly better fit for the hybrid regulatory model. The right hand column in Figure 2 shows the simulated switch-off behaviour obtained using the parameters estimated from the switch-on transition data. Both models capture the general de-activation trend, but the hybrid regulatory model seems to give a slightly better prediction of the initial transient behaviour of *ASH*1. The results for the hybrid regulatory model were obtained in less than a minute on a standard dual-core desktop machine.

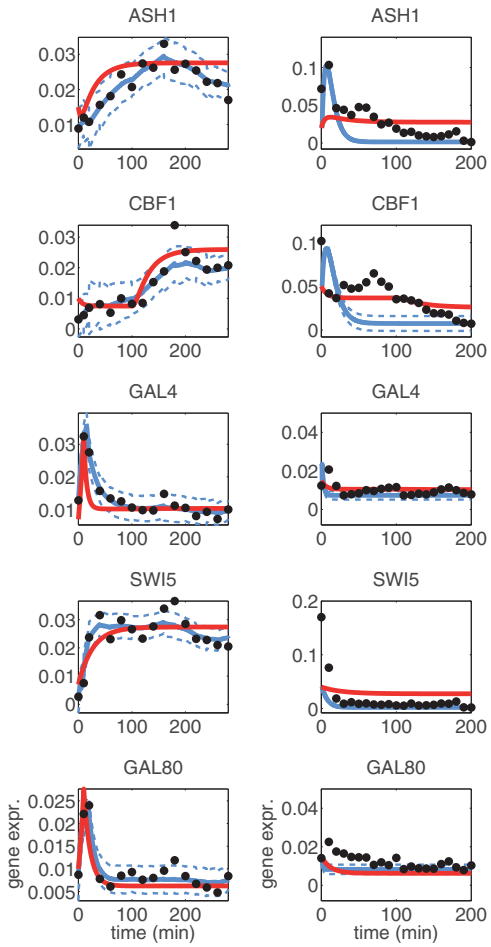

Fig. 2. Model fitting and predictions on IRMA dataset. Left column represents the fit of the models to the training, switch-on data. Right column are predictions on the switch-off data. The dark grey lines (red lines in the online version) are the results of the Cantone *et al*. model, light grey lines (blue lines in the online version) the results of the hybrid regulatory model (with dashed confidence intervals)

3.2 Modelling circadian clock in *O.tauri*

Despite its complex topology and the relatively large number of genes involved, the IRMA network does not exhibit particularly complex dynamics during the two transitions. As a second example, therefore, we consider a circadian clock, i.e. a network that can sustain oscillatory dynamics autonomously. By standard results in dynamical systems theory, this implies the presence of feedback loops in the network architecture. Typically, transcription–translation models are used to explain the sustained oscillations of gene expression. In a minimal model, the translational product of a clock gene becomes the transcriptional activator/inhibitor of another clock gene. Given the importance of circadian clocks for biomedical applications, and the availability of many tools to study oscillatory time series in mathematics and engineering, circadian clocks have become a major focus of systems biology research

The picoalga *O.tauri* has recently emerged as a powerful, yet simple, model of plant circadian clocks owing to its compact genome and extremely simple physiology. Notably, only the clock genes TOC1 and CCA1, represented by multiple members in the higher plant *Arabidopsis thaliana*, are encoded in *O.tauri*, along with a cryptochrome-like gene with possible clock involvement (Heijde *et al*., 2009). This has led to the hypothesis that its clock network consists of a minimal oscillator of a single loop between these two genes. This hypothesis has been explored mathematically in a number of articles (Morant *et al*., 2010; Thommen *et al*., 2010; Troein *et al*., 2011); most recently, Troein *et al*. provided comprehensive datasets consisting of several luciferase time series measurements of TOC1 and CCA1 protein abundance in a synchronized population of *O.tauri* cells. Troein *et al*. then proposed a detailed ODE model of the system, and used 144 luciferase time series to parametrize the model.

We compare the results of our hybrid approach with the ODE approach of Troein *et al*. on the *O.tauri* circadian clock data. The structure of the model is a simple negative feedback loop (NFL) network, including the evening gene TOC1 and the morning gene CCA1 (Fig. 5A). We consider the presence of a single light input affecting the CCA1 promoter state. This is to mimic light-induced phosphorylation of the TOC1 transcription factor (Troein *et al*., 2011), which affects its ability to bind the CCA1 promoter (see Supplementary Material for how this input contribution is modelled).

To evaluate the performance of our approach, we compare our stochastic hybrid approach and the complex clock model of Troein *et al*. (2011). Our hybrid regulatory model has only two SDEs and six free parameters, with mild non-linearities coming from the exp(x) terms in the master equation; Troein *et al*.'s model is a system of seven ODEs with 19 parameters.

All models were given two time series of TOC1 and CCA1 protein concentrations sampled hourly across three cycles. The data were obtained by measuring luciferase (LUC) luminescence sampled at regular intervals during 12 h:12 h light–dark cycles (L:D 12:12) in transgenic lines where LUC was fused to the protein of interest. We indicate these data as translational reporter data (y_{TOC1} and y_{CCA1}). The parameters of the ODE model were determined again by stochastic optimization (see Supplementary Material), while in our model they are

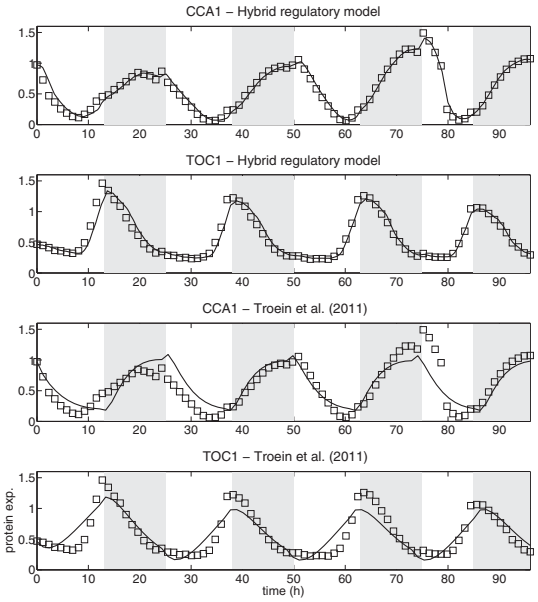

Fig. 3. Fit of the two models to training data: *upper*, posterior mean of hybrid regulatory model; *bottom*, model of Troein *et al.* (2011)

learned during the variational Bayesian scheme as described in section 2.4.

Figure 3 shows the results of applying these procedures on the training data: the top two panels show the posterior mean of the hybrid regulatory model, while the bottom two panels show the optimized fit of the ODE model. The Troein *et al.*'s model is sufficiently complex to afford a credible fit to the data. Nevertheless, even with its complexity, it cannot explain the higher level of CCA1 expression in the third cycle, while the stochastic model of course can accommodate that by a slightly higher activation of the CCA1 (latent) promoter variable.

Next, we compare the predictions of the two models on independent data where the light input (length of the light periods, *photoperiod*) was altered. We focus here on predicting the expression of TOC1, as it is not directly affected by the light input. To do this, we use the two models parameterized using the L:D 12:12 translational reporter data. We then simulate an entraining phase where the oscillator is driven for a long time by an L:D 12:12 cycle and then suddenly alter the photoperiod of the cycle to L:D 6:18, followed by a period of constant light. This mimics the experimental setting in which the data were collected (Troein *et al.*, 2011). We stress that these predictions are truly *out of sample* predictions in a statistical sense: the data we compare with have not been used in any form to parameterize or tune the models (except for a global scaling factor due to the arbitrariness of the units on the LUC signal). Figure 4 shows the results of the simulation of TOC1 expression from the two models. Both models accurately predict a reduction of amplitude of the oscillations during the altered L:D cycles. However, Troein *et al.*'s model completely misfits the final constant light period, both in terms of frequency and amplitude. On the other hand, the stochastic hybrid approach provides robust predictions that continue to oscillate with the same period after the change to constant light. Furthermore, it predicts an increase in the average

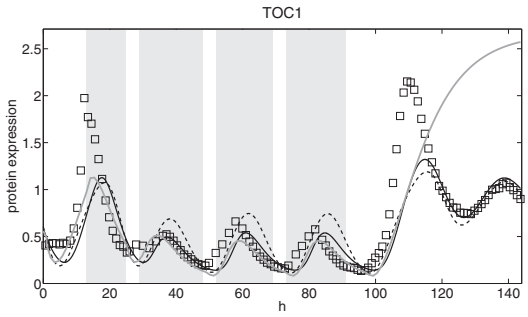

Fig. 4. Prediction of TOC1 protein level after exposure to light dark cycles with altered photoperiod: dashed, prediction using the hybrid regulatory model (NFL); grey solid line, prediction using the Troein *et al.*'s model (Troein *et al.*, 2011); black solid line, prediction using the hybrid regulatory model (repressilator TOC1-X-CCA1); empty squares, observed data

value of the TOC1 protein, and a dampening of the amplitudes of the oscillations in constant light. As a further control, in the Supplementary Material, we also consider whether a simple deterministic approach (obtained by training by optimization of the mean behaviour of the hybrid regulatory model) could yield good fits and predictions. The results in the Supplementary Material show that the simple model can indeed yield robust predictions but is unable to fit the training data satisfactorily: the stochastic hybrid regulatory models appears to strike a good compromise between the flexibility given by the latent promoter process, and the robustness given by the simplicity of the underlying model.

3.3 Predicting the clock's structure

Troein *et al.* (2011) also provide an indirect measurement of promoter states in the form of luciferase time series. This is obtained by inserting in the *Ostreococcus* genome another copy of the TOC1 or CCA1 promoters directly fused to luciferase. We call these additional data sources transcriptional reporters and denote them as y_{pTOC1} and y_{pCCA1}.

Statistically, these two data types could be represented with the simpler models of Figure 5B; nevertheless, obviously the promoter state profiles inferred from transcriptional reporters using the model in Figure 5B should match reasonably well the profiles inferred from translational reporters using the model in Figure 5A.

Figure 6 shows the results of this approach for CCA1 (left) and TOC1 (right) arranged in a negative feedback loop. Surprisingly, while the predicted promoter states of TOC1 match well, CCA1 promoters present different dynamics when inferred from transcriptional and translational reporter data, exhibiting an average phase shift of $\sim 40^o$. More worryingly, the phase shift is highly asymmetrical, with accurately matched off-time estimates and widely divergent on-time estimates.

We then decided to explore the possibility that this mismatch may be due to an incorrect network topology. Recent results have shown that, in *A.thaliana*, TOC1 acts as a repressor (Huang *et al.*, 2012), and that the core structure of the *Arabidopsis* clock is better represented as a three-node network known as a repressilator (Elowitz and Leibler, 2000; Pokhilko

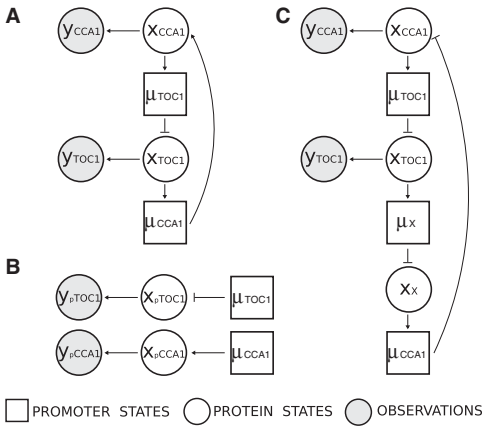

Fig. 5. Statistical models for *O.tauri*: negative feedback loop (**A**), transcriptional models (**B**), repressilator TOC1-X-CCA1 (**C**). Note that in order to compare inference results obtained with transcriptional and repressilator models, we need to consider a repressive regulation between μ_{CCA1} and x_{pCCA1}

et al., 2012). Therefore, we introduced a hypothetical third gene X in the NFL network, leading to a more complex clock network. There are only two possible repressilator configurations after introduction of a putative clock gene X into the previous network, depending on whether X is repressed by TOC1 or by CCA1. We will refer to them as TOC1-X-CCA1 and CCA1-X-TOC1, where X has the role of repressor for CCA1 and TOC1, respectively (Fig. 5C).

We then repeated the inference of the promoter states using the repressilator model. Naturally, in this case, we do not have translational and transcriptional reporter data for the hypothetical gene X; however, marginalization of this additional latent variable is straightforward in the Bayesian setting. Therefore, we can use the translational reporter data y_{TOC1} and y_{CCA1} to infer the promoter states μ_{TOC1}, μ_{CCA1} and μ_X (as well as the protein states for all three genes).

Figure 7 left panel shows the predicted promoter state of CCA1 using the TOC1-X-CCA1 architecture (the TOC1 promoter gives a good agreement also with this architecture). As can be seen, the average phase shift is greatly reduced, and the inferred promoter states overlap symmetrically.

Interestingly, using the CCA1-X-TOC1 repressilator structure, the model fails to predict the CCA1 promoter state. Therefore, our approach predicts that the *O.tauri* clock should have a repressilator structure, and that the third gene should be repressed by TOC1. As TOC1 is expressed mainly in the evening, it follows that the third gene X should be an afternoon gene, as predicted by our model also (Fig. 7 right panel). This is consistent with the existing knowledge of the *A.thaliana* clock.

Next, we checked whether the repressilator model trained on L:D 12:12 translational reporter data is able to predict TOC1 profiles in altered photoperiods. The results are shown in Figure 4, which compares the repressilator predictions (black solid line) with the NFL predictions (dashed). It is apparent that the repressilator provides a more accurate prediction, particularly during the final constant light period. Further predictions on different altered photoperiods are shown in the Supplementary Material.

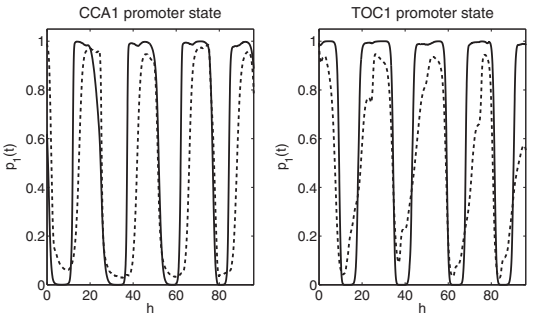

Fig. 6. Inferred promoter states for CCA1 (left) and TOC1 (right), obtained with the NFL model using translational (solid lines) and transcriptional (dashed lines) reporters

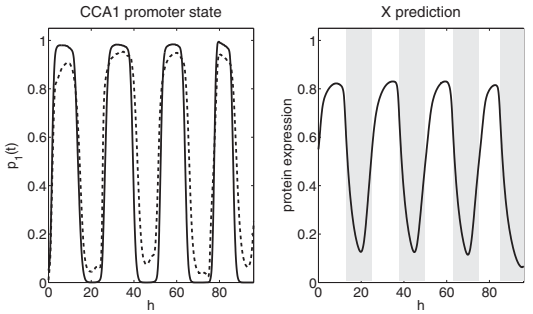

Fig. 7. Inferred promoter states for CCA1 (left), obtained with the repressilator model TOC1-X-CCA1 using translational (solid lines) and transcriptional (dashed lines) reporters. The right panel shows the mean prediction of the hypothetical gene X

4 DISCUSSION

Mathematical modelling of GRNs is fundamental to our attempts to understand the structure and dynamics of gene networks. Non-linear ODE models provide an excellent framework to elucidate and predict complex regulatory mechanisms in small-to-medium scale GRNs. However, they can be vulnerable to incomplete knowledge of the system, and calibrating complex models to limited data may pose an unsurmountable statistical challenge.

Here, we have presented a statistical approach to modelling transcription–translation networks, which aims at retaining the flexibility allowed by non-linear ODE models while making possible a statistical exploration of the model's parameterization. The approach relies on a stochastic hybrid representation of the system where the transcription–translation mechanism is modelled using only two variables: promoter (latent) states and protein states. By replacing complex non-linearities and additional unknown parameters of ODE models with latent variables, the model becomes simpler, more robust and more identifiable.

Our empirical study demonstrates the identifiability of our approach and shows how on two real biological problems, it can yield comparable or better predictions than competing methods, as well as leading to novel testable biological hypotheses. Our prediction that a repressilator structure underpins the *O.tauri* clock is in line with recent discoveries on the structure

of the *A.thaliana* clock (Pokhilko *et al*., 2012), and indeed with the structure of known animal circadian clocks (Ukai-Tadenuma *et al*., 2011). If validated, this finding would suggest that the repressilator structure is an evolutionarily conserved feature of eukaryotic circadian clocks. Furthermore, our model predicts that TOC1 acts as a repressor (while remaining an indirect activator of CCA1 through a double repression); again, the repressor role of TOC1 was recently demonstrated in *A.thaliana* (Gendron *et al*., 2012; Huang *et al*., 2012), leading further weight to our hypotheses. Although the repressilator model substantially ameliorates the model misfit of the NFL model, there remains some residual unexplained discrepancy between inferences from transcriptional and translational reporters. Although this may be due to noise in the data, it cannot be excluded that the complexity of the *O.tauri* oscillator may be even greater, as is the case of other plant oscillators (Pokhilko *et al*., 2012).

We believe that these results show the promise of this approach as an effective tool in addressing systems biology problems. Nevertheless, this work opens further avenues for development. From the biological point of view, validation of the novel structure of the *O.tauri* clock would be an important step, which is likely to require substantial bioinformatics research. However, perhaps even more interesting would be to computationally explore the links between the transcription–translation oscillator we study and the recently described non-transcriptional oscillator of *O.tauri* (O'Neill *et al*., 2011). From the computational point of view, this study does not address the important problem of *de novo* reconstruction of the structure of the regulatory network, but relies on pre-existing network structures. A systematic method to combine structure learning with dynamical modelling remains a desirable goal (Oates *et al*., 2012); we hope that this work will represent an advance in that promising direction.

ACKNOWLEDGEMENT

We thank Dr Gerben Van Oijen for useful comments on a draft manuscript, and Dr Carl Troein for useful discussions on *O.tauri* circadian clocks.

Funding: SynthSys is a Centre for Integrative Systems Biology supported by BBSRC and EPSRC award D190621. G.S. acknowledges support from the European Research Council under grant MLCS306999.

Conflict of Interest: none declared.

REFERENCES

Asif,H. and Sanguinetti,G. (2011) Large-scale learning of combinatorial transcriptional dynamics from gene expression. *Bioinformatics*, **27**, 1277–12183.

Cantone,I. *et al.* (2009) A yeast synthetic network for *in vivo* assessment of reverse-engineering and modeling approaches. *Cell*, **137**, 172–181.

Elowitz,M. and Leibler,S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335–338.

Elowitz,M. *et al*. (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.

Gendron,J. *et al*. (2012) Arabidopsis circadian clock protein, TOC1, is a DNA-binding transcription factor. *Proc. Natl Acad. Sci. USA*, **109**, 3167–3172.

Georgoulas,A. *et al.* (2012) A subsystems approach for parameter estimation of ode models of hybrid systems. In: Bartocci,E. and Bortolussi,L. (eds) *Hybrid Systems and Biology, EPTCS*, Vol. 92. Open Publishing Association, Newcastle upon Tyne, UK.

Heijde,M. *et al.* (2009) Characterization of two members of the cryptochrome/photolyase family from *Ostreococcus tauri* provides insights into the origin and evolution of cryptochromes. *Plant Cell Environ*., **33**, 1624–1626.

Hoops,S. *et al.* (2006) COPASI—a complex pathway simulator. *Bioinformatics*, **22**, 3067–3074.

Huang,W. *et al.* (2012) Mapping the core of the Arabidopsis circadian clock defines the network structure of the oscillator. *Science*, **336**, 75–79.

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

Liepe,J. *et al*. (2010) ABC-SysBio—approximate Bayesian computation in Python with GPU support. *Bioinformatics*, **26**, 1797–1799.

Morant,P. *et al*. (2010) A robust two-gene oscillator at the core of Ostreococcus tauri circadian clock. *Chaos*, **20**, 045108.

Oates,C.J. *et al.* (2012) Network inference using steady-state data and Goldbeter-Koshland kinetics. *Bioinformatics*, **28**, 2342–2348.

Ocone,A. and Sanguinetti,G. (2011) Reconstructing transcription factor activities in hierarchical transcription network motifs. *Bioinformatics*, **27**, 2873–2879.

O'Neill,J. *et al.* (2011) Circadian rhythms persist without transcription in a eukaryote. *Nature*, **469**, 554–558.

Opper,M. and Saad,D. (2001) *Advanced Mean Field Methods: Theory and Practice*. The MIT Press, Cambridge, MA.

Opper,M. and Sanguinetti,G. (2007) Variational inference for Markov jump processes. In: Platt,J., *et al.* (eds) *Advances in Neural Information Processing Systems*, Vol. 20. Curran Associates, Inc., Vancouver, British Columbia, Canada.

Opper,M. and Sanguinetti,G. (2010) Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics*, **26**, 1623–1629.

Opper,M. *et al*. (2010) Approximate inference for Gaussian-jump processes. In: *Advances in Neural Information Processing Systems*, Vol. 24. Curran Associates, Inc., Vancouver, British Columbia, Canada.

Pokhilko,A. *et al*. (2012) The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. *Mol. Syst. Biol*., **8**, 574.

Ptashne,M. and Gann,A. (2002) *Genes and Signals*. Cold Harbor Spring Laboratory Press, New York.

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.

Schultz,D. *et al*. (2007) Molecular level stochastic model for competence cycles in bacillus subtilis. *Proc. Natl Acad. Sci. USA*, **104**, 17582–17587.

Swain,P. *et al*. (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 12795–12800.

Thommen,Q. *et al*. (2010) Robustness of circadian clocks to daylight fluctuations: hints from the picoeucaryote Ostreococcus tauri. *PLoS. Comput. Biol*., **6**, e1000990.

Troein,C. *et al*. (2011) Multiple light inputs to a simple clock circuit allow complex biological rhythms. *Plant J*., **66**, 375–385.

Ukai-Tadenuma,M. *et al*. (2011) Delay in feedback repression by cryptochrome 1 is required for circadian clock function. *Cell*, **144**, 268–281.

Vyshemirsky,V. and Girolami,M.A. (2008) Bayesian ranking of biochemical systems models. *Bioinformatics*, **24**, 833–839.

Wilkinson,D. (2011) *Stochastic Modelling for Systems Biology*. Chapman and Hall/CRC Press, London.