BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–7

Gaussian Process Modelling of Latent Chemical Species: Applications to Inferring Transcription Factor Activities

Pei Gao [1], Antti Honkela [2], Magnus Rattray [1] and Neil D. Lawrence [1]

[1]School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester, M13 9PL
[2]Adaptive Informatics Research Centre, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Finland

Received on Feb 2008; accepted on Apr 2008

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Inference of *latent chemical species* in biochemical interaction networks is a key problem in estimation of the structure and parameters of the genetic, metabolic and protein interaction networks that underpin all biological processes. We present a framework for Bayesian marginalisation of these latent chemical species through Gaussian process priors.

Results: We demonstrate our general approach on three different biological examples of single input motifs, including both activation and repression of transcription. We focus in particular on the problem of inferring transcription factor activity when the concentration of active protein cannot easily be measured. We show how the uncertainty in the inferred transcription factor activity can be integrated out in order to derive a likelihood function that can be used for the estimation of regulatory model parameters. An advantage of our approach is that we avoid the use of a coarse-grained discretization of continuous-time functions, which would lead to a large number of additional parameters to be estimated. We develop exact (for linear regulation) and approximate (for non-linear regulation) inference schemes, which are much more efficient than competing sampling-based schemes and therefore provide us with a practical toolkit for model-based inference.

Availability: The software and data for recreating all the experiments in this paper is available in MATLAB from `http://www.cs.man.ac.uk/~neill/gpsim`.

Contact: `neill@cs.man.ac.uk`

1 INTRODUCTION

Ordinary differential equations (ODEs) are the most common framework in use for modelling biological sub-systems (Alon, 2006). Well established methodologies have been developed for estimating the parameters of these equations in the context of a particular experiment or set of experiments, using *e.g.* least squares and maximum likelihood combined with an appropriate optimisation algorithm (Mendes and Kell, 1998). More recently, significant progress has been made on Bayesian parameter estimation in the context of ODEs (*e.g.* Coleman and Block, 2006). Through the use of advanced Monte Carlo techniques it is even possible to, given a specific data set, rank model structures through the use of Bayes factors (Vyshemirsky and Girolami, 2008). This shows the potential for ODE models to be closely integrated with biological investigations, informing the process of biological experimental design.

A challenging problem for parameter estimation in ODE models occurs where one or more chemical species influencing the dynamics are controlled outside of the sub-system being modelled. For example, a signalling pathway can be triggered by a signal external to the pathway itself. In a regulatory sub-system, one or more transcription factors (TFs) may influence the expression of a set of target genes, but these TFs may not be regulated at the transcriptional level, instead being activated by another sub-system such as a signalling pathway. Similarly, in a metabolic pathway external metabolites and enzymes will influence the dynamics of the pathway. If these external chemical species have a constant influence, *e.g.* as in the case of steady-state behaviour of a metabolic pathway, then they can simply be treated as additional parameters of the model and their effect can be estimated along with the other model parameters. However, more often these external factors are time-varying quantities. In this case they are functional parameters and cannot be estimated by the standard methods discussed above. One approach for dealing with this is to discretize in time, treating the time varying function as a sequence of discrete parameters. However, this leaves the problem of choosing the correct granularity for the discretization and either ignoring temporal continuity, or assuming a simple Markovian relationship and thereby introducing further parameters and assumptions. Here we propose an alternative approach. We deal with these parameters as continuous functions of time, avoiding the need for arbitrary discretisation.

To further compound the problem of dealing with the time-varying effects of these chemical species, their concentration is often not directly observable and their dynamics must therefore be inferred indirectly according to their influence on measured elements of the system. This is a common problem and it is a natural consequence of the fact that some quantities are relatively easy to measure in a high throughput manner (*e.g.* mRNA concentrations with a microarray), whereas others are much more difficult to measure (*e.g.* the concentration of transcription factors located in the nucleus).

In this paper we advocate the use of Gaussian processes to define prior distributions over these *latent chemical species*. This allows us to marginalise their contributions in the interaction network of interest. We present a basic toolkit of algorithms based on Gaussian processes which allow us to consider different response models (Michaelis Menten kinetics, repression responses) and cascades of interactions in which chemical species of interest are missing. The application domain we consider is inference of transcription factor (TF) activity in both developmental and signalling networks.

Inference of TF activity in a given network is a well studied problem with both genome wide approaches (*e.g.* Liao et al., 2003, Sanguinetti et al., 2006b,a) and algorithms designed for a subset of genes (*e.g.* Nachman et al., 2004, Rogers et al., 2006, Khanin et al., 2006, Barenco et al., 2006). Our approach is most directly inspired by Barenco et al. [2006] who infer transcription factor activity in a single input module network motif through a differential equation with a linear response. We build on their work to consider simple cascade networks and non-linear response models such as Michaelis Menten kinetics and repression responses (Alon, 2006, Khanin et al., 2006).

Gaussian processes (Rasmussen and Williams, 2006) are probabilistic models of functions that encode particular assumptions about the function such as smoothness and timescale. They are commonly applied in the context of regression and interpolation. Gaussian process modelling provides not only the inference of continuous quantities without discretization but also the natural capability of handling uncertainty. Another attractive characteristic of the Gaussian process (GP) is that the result of any linear operator on the function leads to another GP and we will exploit this when applying GPs in the context of differential equations.

Our focus in this paper will be the inference of transcription factor activities given mRNA concentrations. We start by considering the model given in Barenco et al. [2006] and reviewing the work done by Lawrence et al. [2007] who provided the first treatment of this model with GPs. Then, in Section 2.3, we extend our model to the case of repression. We follow Khanin et al. [2006] and apply the model in the context of the SOS system in *E. coli* where genes are controlled by the transcriptional repressor protein LexA. In both these cases, no model of transcription factor translation is included as the proteins are post-translationally regulated. In our final example (Section 2.4) we show how, in the context of *Drosophila* mesoderm development, a model of translation can also be incorporated to improve the quality of the transcription factor inference in cases where TFs are primarily regulated at the transcriptional level.

2 METHODS AND RESULTS

The following linear model of gene activation was considered by Barenco et al. [2006],

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = B_j + S_j f(t) - D_j x_j(t), \tag{1}$$

where the changing level of a gene j's expression, $x_j(t)$, is given by a combination of a basal transctiption rate, B_j, a sensitivity, S_j, to its governing TF's activity, $f(t)$, and the decay rate of the mRNA, D_j. The differential equation can be solved for $x_j(t)$ giving

$$x_j(t) = \frac{B_j}{D_j} + S_j \int_0^t e^{-D_j(t-u)} f(u)\,\mathrm{d}u, \tag{2}$$

where we have ignored any transient terms (we can assume these are zero if the TF has zero concentration at times before $t = 0$).

2.1 Gaussian process inference

Our general approach is to assume that $f(t)$ was drawn from a Gaussian process. A Gaussian Process (GP) is a prior distribution over functions that leads to highly flexible non-linear functions in which characteristics such as the stationarity, the roughness and the timescale of the signal can be controlled. GPs are the functional analogue of the Gaussian distribution, they are fully specified by a mean function, $\mu(t)$, and a covariance function, $k(t,t')$. The mean function is an unconstrained function of time, whilst the covariance function is constrained to be a positive definite function. Various covariance functions can be used but we will predominantly make use of the squared exponential covariance (sometimes known as the Gaussian or RBF covariance),

$$k(t,t') = \exp\left(-\frac{(t-t')^2}{l^2}\right). \tag{3}$$

An important characteristic of a GP is that any linear operation applied to the function drawn from a GP leads to a function that is drawn from a related GP. This is the functional analogue of linear operations applied to samples from a Gaussian distribution. We note that (2) is a linear operation on $f(t)$ since the integral is the functional analogue of a weighted sum. The properties of the GP tell us that if $f(t)$ was drawn from a GP with covariance function $k_{f,f}(t,t')$ then $x_j(t)$ will also be drawn from a GP with a covariance function given by

$$k_{x_j,x_j}(t,t') = S_j^2 \int_0^t \int_0^{t'} \mathrm{e}^{-D_j(t-u+t'-u')} k_{f,f}(t,t')\,\mathrm{d}u\mathrm{d}u'. \tag{4}$$

Furthermore, the cross covariances between the $x_j(t)$ and $f(t)$ will be given by

$$k_{x_j,f}(t,t') = S_j \int_0^t e^{-D_i(t-u)} k_{f,f}(t,t')\,\mathrm{d}u. \tag{5}$$

Finally, if our data consist of several genes, $\{x_j(t)\}_{j=1}^N$, cross covariances between the genes can be computed,

$$k_{x_i,x_j}(t,t') = S_i S_j \int_0^t \int_0^{t'} e^{-D_i(t-u)-D_j(t'-u')} k_{f,f}(t,t')\,\mathrm{d}u\mathrm{d}u'. \tag{6}$$

The basal transcription rate then appears in the mean function to define the mean of the Gaussian process for $x_j(t)$ as a constant B_j/D_j.

As was shown in Lawrence et al. [2007], if $f(t)$ is drawn from a GP with a squared exponential covariance function,

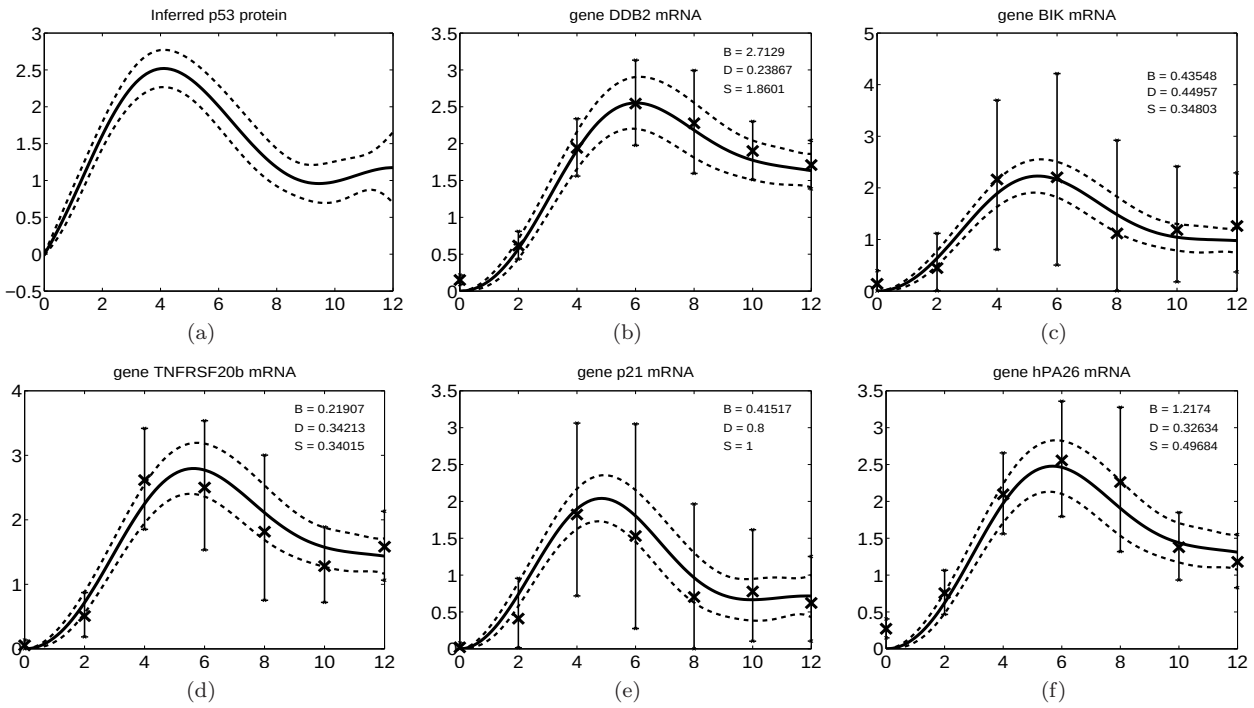

Figure 1. Results for p53 using the linear model and a GP with RBF kernel. Results are shown for one of the replicates: (a) predicted protein concentration; (b) predicted expression level for DDB2; (c) predicted expression level for BIK; (d) predicted expression level for TNFRSP20b; (e) predicted expression level for p21; (f) predicted expression level for hPA26. Solid lines represent the mean inference, dashed lines show the 95% credible intervals, and the crosses are the observed gene expression data with error bars showing the technical error from each individual Affymetrix microarray processed using the puma package (Liu et al., 2005). Maximum likelihood model parameters, estimated using all replicates, are shown for each target gene. Data and reconstructed profiles are shown on an unlogged normalised scale. Time is measured in hours.

we can analytically solve the integrals in equations (4–6) and thereby define a probabilistic model of the expression data for which the parameters of the differential equation, B_j, S_j, D_j and the scale l can all be determined, either through maximum likelihood or Bayesian sampling.

A likelihood function for the model parameters $\theta = \{B_j, S_j, D_j\}_{j=1}^N$ and GP scale l is obtained by *integrating out* the latent function $f(t)$

$$L(\theta, l) = \int \left(\prod_j p(\boldsymbol{x}_j | \theta, f(t)) \right) p(f(t)|l) \, \mathrm{d}f(t) \ , \qquad (7)$$

where the data $\boldsymbol{x}_j = [x_j(t_i)]$ are collected at discrete times t_i and modelled using Gaussian observation noise with either known or estimated variance (this is for a single replicate). Integrating out $f(t)$ avoids having to carry out maximum likelihood for this infinite dimensional parameter which could be problematic as estimates based on limited data would be associated with large uncertainty. Approaches based on carrying out maximum likelihood over a discretized $f(t)$ suffer from the same problem, *i.e.* the discretisation introduces many new parameters which have to be estimated and therefore only a very coarse discretization will be tractable (Khanin et al., 2006). The associated error in these parameters confounds the estimation of other parameters by maximum liklihood and while Bayesian approaches avoid this problem there is a large associated computational bottleneck in using Markov chain Monte Carlo (MCMC) on these extra parameters (Rogers et al., 2006, Barenco et al., 2006).

Once the model parameters have been estimated then the model can be used to estimate the predictive GP distribution for $f(t)$ and for each target gene $x_j(t)$. The mean and covariance of this predictive distribution for $f(t)$ are given by,

$$\langle f \rangle_{\mathrm{post}} = K_{f\boldsymbol{x}} K_{\boldsymbol{x}\boldsymbol{x}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \qquad (8)$$

$$K_{ff}^{\mathrm{post}} = K_{ff} - K_{f\boldsymbol{x}} K_{\boldsymbol{x}\boldsymbol{x}}^{-1} K_{\boldsymbol{x}f} \qquad (9)$$

where $\boldsymbol{x}$ and $\boldsymbol{\mu}$ have the dimension of $N \times T$ with the elements of $x_{j,i} = x_j(t_i)$ (for $j = 1, \ldots N$ and $i = 1, \ldots T$) and $\mu_{j,i} = B_j/D_j$ (for all i) respectively, $(K_{ff})_{p,q} = k_{ff}(t_p, t_q)$, $(K_{f\boldsymbol{x}})_{p,rj} = k_{f,x_j}(t_p, t_r)$ and $(K_{\boldsymbol{x}\boldsymbol{x}})_{ri,mj} = k_{x_i,x_j}(t_r, t_m) + \delta_{ij}\delta_{mr}\sigma_{ir}^2$ where σ_{ir}^2 is the measurement noise associated with $x_i(t_r)$. Here t_p and t_q denote any choice of points over which one wishes to evaluate the GP, whereas t_r and t_m are specifically the times at which the data are measured. This equation was used to plot the credible intervals for f in the figures presented in this paper. It is similarly possible to compute the predictive distribution for the data by replacing f with x_i in the above expression. The predictive distribution can be

used for ranking genes according to their fit to the model (as in the application considered by Barenco et al. [2006]).

We illustrate our results from applying the method below, first on the previously studied linear activation model and then showing how we can extend the approach to non-linear response functions for activation and repression, followed by a two-layer activation model.

2.2 Activation

In this section we consider a system in which a single TF activates a number of targets. The example we consider is the TF p53 which is a tumour repressor activated during DNA damage. According to Barenco et al. [2006], irradiation is performed to disrupt the equilibrium of the p53 network stimulating transcription of p53 target genes. Seven samples of the expression levels of the target genes in three replicas are collected as the raw time course data.

2.2.1 Linear Activation In this section we recreate the results presented by Lawrence et al. [2007] for the linear model with several key differences. Firstly, in their original paper Barenco et al. [2006] constrained $f(0)$ to be zero, forcing the basal transcription rate to account for all transcription at time $t = 0$. This constraint was not included in Lawrence et al. but is included here. This allows us to incorporate the prior information of the latent TF profiles as much as possible. Secondly, Lawrence et al. used an unnormalised version of the Affymetrix array data. We found that simple median based normalisation removed the effect of a couple of repeats that were anomalously high. Inspection of the processed data used by Barenco et al. showed that they had also dealt with these anomalies.

In Figure 1 we show results of the TF inference as well as the model parameter estimations. The latent TF activity profiles are reconstructed with 95% credible intervals. The result resembles the activity profile of p53 measured by Western blot (Barenco et al., 2006) and the kinetic parameters in the model are also closely matched with the results there (we followed Barenco et al. in fixing kinetic parameters for p21 to improve identifiability).

2.2.2 MAP-Laplace Approximation The differential equation with a linear response is an attractive model to use in the context of GPs as it allows the joint distribution over the gene expression and TF activity to be determined analytically, given the model parameters. However, as a model, it has some shortcomings. Firstly, it treats both the gene expression and the TF activity as GPs. Since a GP cannot encode the information that a function is constrained positive, this means that the concentrations are *a priori* allowed to be negative. Whilst the posteriors, in the region where there is data, tend to stay positive (see Figure 1), when the predictions move away from the data they allow the TF activity to become negative. In Figure 2(a) we show the results using the linear model for a different replicate to the one used in Figure 1. In this case it can be observed that, although the mean inferred profile remains positive or very close to zero,

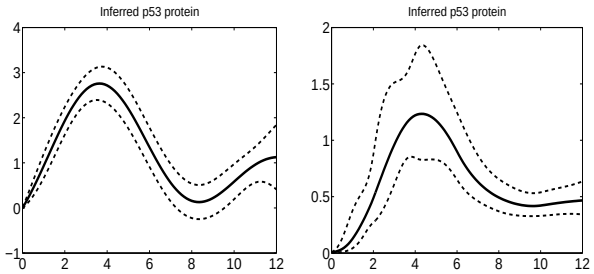

Figure 2. Inferring p53 activity for a different replicate than Figure 1 (a) using the linear model; (b) using the nonlinear model with Michalis Menten kinetics and a positively constrained TF concentration.

the inferred distribution of profiles summarized by the credible intervals includes profiles with negative concentrations. This contradicts our prior knowledge that concentrations are positive quantities.

A potential solution to this problem is to place a GP prior over, for example, the log of the TF activity. However, this, in effect, is a non-linear response in the differential equation. The non-linear response means that it is no longer possible to construct the joint distribution over gene expression and TF activity in a closed form. We must, necessarily, turn to approximations to make progress. In Lawrence et al. [2007] the use of a MAP-Laplace approximation is suggested. They demonstrate how the concentration of the TF activity can be constrained positive by placing the GP in log space.

Consider the following modification to the model,

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = B_j + S_j g(f(t)) - D_j x_j(t), \tag{10}$$

where $g(\cdot)$ is a non-linear function. The differential equation can still be solved,

$$x_j(t) = \frac{B_j}{D_j} + S_j \int_0^t e^{-D_j(t-u)} g_j(f(u))\,\mathrm{d}u \tag{11}$$

but there is now a non-linear operation on $f(t)$ within the integral. The gene expression level is therefore no longer a GP. The MAP-Laplace approach involves finding a maximum *a posteriori* (MAP) estimate for the function $f(t)$ and making a second order Taylor approximation at that point to the log likelihood. This approximation is itself a Gaussian process and leads to an approximation to the marginal likelihood (Rasmussen and Williams, 2006). Derivatives can then be taken with respect to the model parameters and the approximation maximised. The MAP-Laplace's approximation becomes exact in the case where $g(\cdot)$ is *linear*. This is also useful in practice, naive implementation of the linear response model for a general covariance function would require, in general, numerical evaluation of the double integrals in equations (4) and (6). The use of the MAP-Laplace approach involves only a single numerical integral.

We now build on this work in two major ways. We go beyond the simple exponential response model considered

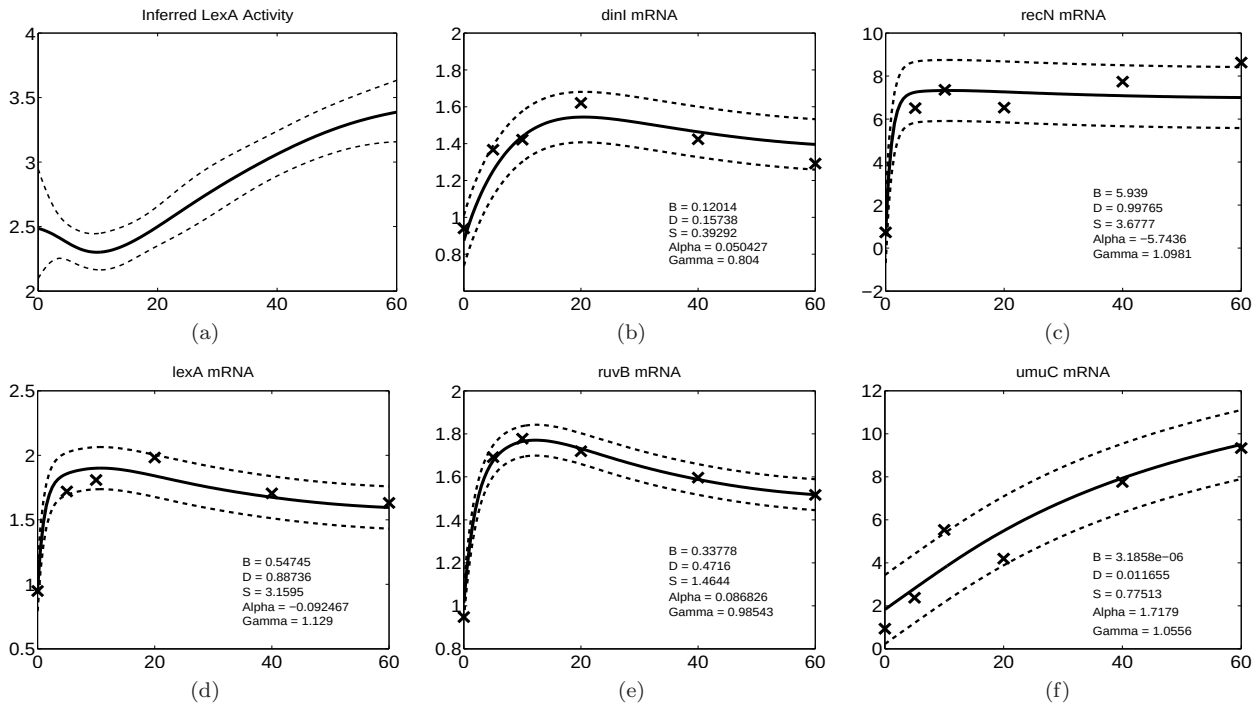

Figure 3. Results for the repressor LexA. (a) predicted LexA concentration; (b) predicted expression level for dinI; (c) predicted expression level for recN; (d) predicted expression level for lexA; (e) predicted expression level for ruvB; (f) predicted expression level for umuC. Parameters shown were obtained by maximum likelihood. Standard errors were not available for this two-dye microarray data set and therefore a gene-specific noise variance parameter was estimated for each target gene. Data and reconstructed profiles are given in the original unlogged scale. Time is measured in minutes.

in (Lawrence et al., 2007) by exploring a Michaelis Menten kinetics inspired response and a response that has been suggested as appropriate for repression (Alon, 2006). We also extend the algorithm to learn the parameters of the differential equation by maximising the approximation to the log likelihood.

2.2.3 Michaelis Menten Kinetics Michaelis Menten reaction kinetics were designed for the case where a chemical reaction is enabled through an enzyme. If the concentration of the enzyme is much lower than the concentration of the substrate the reaction rate becomes limited by the reduced enzyme availability. The justification in the context of modelling transcription is that there are a limited number of binding sites on the genome for the transcription factor. This 'bottle neck' has the same effect as a reduced availability of enzyme and Michaelis Menten kinetics is therefore also appropriate as a model of transcription.

We implemented Michaelis Menten kinetics for the p53 data by taking the non-linearity to have the following form

$$g_j\left(f\left(t\right)\right) = \frac{e^{f(t)}}{\gamma_j + e^{f(t)}}, \tag{12}$$

where the Michaelis constant for the ith gene is given by γ_i and we are using a GP $f\left(t\right)$ to model the log of the TF activity.

Figure 2(b) shows the results of applying the above kinetic model with positively constrained TF concentration to the p53 data set. It can be seen that the inferred distribution of TF profiles no longer includes negative concentrations and the result is now closer in shape to the replicate shown in Figure 1.

2.3 Repression

The same framework can easily be adapted to the case of a repressor by using an analogous Michaelis Menten model of repression,

$$g_j\left(f\left(t\right)\right) = \frac{1}{\gamma_j + e^{f(t)}}, \tag{13}$$

again using $f(t)$ to represent the log TF concentration. Khanin et al. [2006] developed a method to infer the TF profile and model parameters for this model. This involved a coarse discretization of the TF profile, treating $f(t)$ as a piecewise-constant function of time. They acknowledge that more flexible parametric models for $f(t)$ will lead to an increase in the number of parameters, which are therefore impossible to estimate by maximum likelihood. Our approach avoids this explosion of parameters by treating $f(t)$ as a functional parameter which is integrated out before carrying out maximum likelihood for the other model parameters (recall equation (7)).

We applied our method to the same microarray data as Khanin et al. [2006]. They identify 14 target genes which are repressed by the TF LexA in *E. Coli* and mRNA measurements for these genes are collected over six time points. The amount of LexA is reduced after UV irradiation, decreasing for a few minutes and then recovering to its normal level.

In the case of repression we have to include the transient terms in equation (2),

$$x_j(t) = \alpha_j e^{-D_j t} + \frac{B_j}{D_j} + S_j \int_0^t e^{-D_j(t-u)} g_j(f(u))\mathrm{d}u, \quad (14)$$

where $\{\alpha_j\}$ are an additional set of model parameters that have to be inferred.

The result for the inferred TF profile along with predictions for a selection of target genes are shown in Figure 3. Our inferred TF profile and reconstructed target gene profiles are similar to those obtained by Khanin et al. [2006], showing how different model parameters can lead to very different target gene expression given the same stimulus. However, whereas Khanin et al. [2006] had to use interpolation to go from a discretized model to a continuous profile, our GP representation avoids any need for discretization or interpolation. The model parameters shown in the figure are estimated by maximum likelihood.

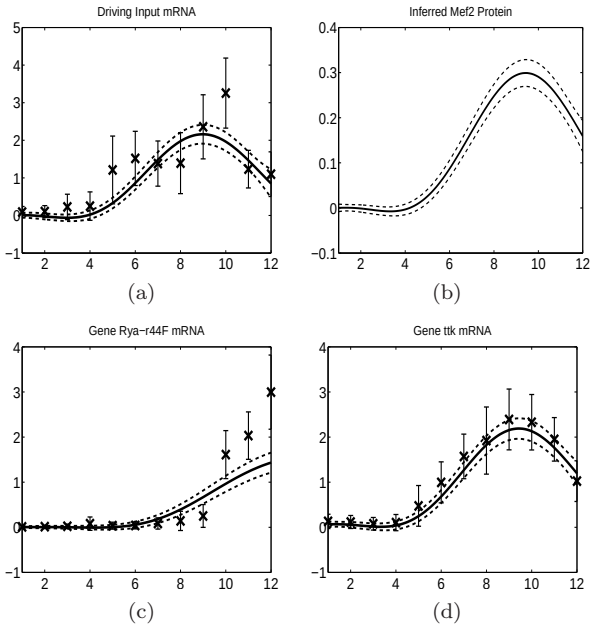

Figure 4. Results for Mef2 from one of the replicates: (a) Driving input mef2 mRNA concentration; (b) inferred Mef2 protein concentration; (c-d) reconstructed mRNA profiles of two example genes from the six used to train the model. Data points and error bars are obtained from probe-level analysis of the Affymetrix data using the puma package (Liu et al., 2005).

2.4 Cascaded Differential Equations

As a final example we consider a simple cascade of differential equations. Returning to the framework of the linear model we consider the case where the TF protein is regulated at the level of transcription. In other words, we model the process of translation from mRNA to protein for the TF, but we are only able to measure the mRNA of the TF and of its targets. The simple model we consider would only be appropriate in the case where the TF does not require activation, *e.g.* by phosphorolation, after translation. The model is, therefore, not appropriate for many signalling pathways, but seems sensible in the context of development where TFs are often functional directly after translation. We therefore considered an example of the development of the mesoderm in *Drosophila* combining target gene predictions from ChIP-chip data [Sandmann et al., 2006] with microarray time-course data from wild-type development [Tomancak et al., 2002].

We take the production rate of active transcription factor to be given by,

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} = \sigma y(t) - \delta f(t) \quad (15)$$

where δ gives the decay rate of the active TF protein, σ gives a rate of translation and $y(t)$ is the concentration of the transcription factor's mRNA. The solution for $f(t)$, setting transient terms to zero, is

$$f(t) = \sigma \int_0^t y(v)\, \mathrm{e}^{\delta(v-t)} dv \ . \quad (16)$$

If we assume that $y(t)$ was drawn from a GP with the squared exponential covariance function, then $f(t)$ is also a GP with the kernel being a special case of the one given by equations (4) and (5). This gives us a new covariance function governing $f(t)$. It turns out that we can use this prior over $f(t)$ in the linear activation model (equation (1)) and derive the associated prior distribution for the target genes of the TF, $\{x_j(t)\}_{j=1}^N$. The derivation and form of the resulting covariance function are rather involved, but remarkably all the integrals required can be calculated analytically for the squared exponential covariance function given by equation (3).

We applied the above driven input model to a simple cascade in *Drosophila* mesoderm development, focusing on the transcription factor Mef2 (Myocyte enhancing factor 2). We selected six targets of Mef2 identified by Chromatin immunoprecipitation (ChIP-chip) assays (Sandmann et al., 2006) and which were observed to be up-regulated after Mef2 is expressed. Affymetrix time course microarray data from wild-type embryos (Tomancak et al., 2002) provides us with observations of Mef2 expression ($y(t)$) and the expression of the target genes $x_j(t)$ at hourly intervals. The microarray time course was replicated three times, and we used all three replicates to fit the model parameters $\delta, \{B_j, D_j, S_j\}$ and the kernel scale l by maximum likelihood (we set $\sigma = 1$ without loss of generality since the scale of the TF protein is arbitrarily determined by the S_j parameters).

Figure 4 shows the results for one of the replicates. Error bars associated with the mRNA data are obtained from the Affymetrix probe-level processing of each replicate (Liu et al.,

2005). In Figure 4(a) we show the inferred mRNA profile for Mef2 along with the microarray data. In Figure 4(b) we show the inferred TF profile for Mef2. Figures 4(c) and (d) show the model predictions for two of the targets, along with the associated expression data, showing again the difference in response possible given different model parameters.

We observe that using the TF mRNA data results in the model obtaining tighter credibility intervals for the inferred TF. The fit to the target gene expression looks reasonable, although the response in Figure 4(c) appears to be sharper than predicted by the model. It also appears that the peak of the inferred Mef2 mRNA profile precedes the peak in the data. The model here highlights the fact that some of our simplifying assumptions are being violated by the biology. The most critical assumption that we have made is that all of the targets used to train the model are unique targets of Mef2. In reality it is likely that other TFs are co-bound with Mef2. In future work we expect to extend the model to account for these other TFs. We are currently using unpublished ChIP-chip data (obtained from E. Furlong, EMBL Heidelberg) for other TFs involved in mesoderm development in order to refine our model and selection of targets.

3 DISCUSSION

Gaussian processes allow for inference over chemical species concentrations in a continuous time manner. In this paper, we have laid out the basic technologies required for this inference and highlighted some of the issues that arise: *e.g.* dealing with non-linear response models and layered systems. Barenco et al. [2006] have already shown that these models can be used for ranking targets of transcription factors. Such rankings can also be given in the context of our models, both for linear and non-linear response models.

Another important direction of future research will be scaling the models used to much larger systems with multiple interacting transcription factors. Larger systems will lead to an increase in computational requirements, particularly if it is necessary to account for correlations between multiple interacting latent chemical species.

The non-linear response models require particular approximations as exact inference in these models is not tractable. One issue with the Laplace approximation we have applied is that it is only responsive to the mode of the posterior distribution. Variational approximations (Jordan et al., 1998) provide an alternative approach which can lead to a rigorous lower bound on the likelihood. Finally, it also makes sense to validate the quality of the approximation through Markov chain Monte Carlo (MCMC) methods. We are working towards efficient MCMC algorithms for validating the quality of our approximate inference.

All the experiments we have shown have been in the context of transcription networks. However, the technologies are equally applicable for other biological systems, such as metabolic networks.

FUNDING

This work is funded by BBSRC Grant. No BBS/B/0076X "Improved processing of microarray data with probabilistic models", EPSRC Grant. No EP/F005687/1 "Gaussian Process Models for Systems Identification with Applications in Systems Biology" and the EU FP6 Network of Excellence PASCAL.

ACKNOWLEDGEMENT

For data provision and assistance with our questions we would like to thank Charles Girardot and Eileen Furlong of EMBL in Heidelberg (mesoderm development in *D. Melanogaster*), Martino Barenco and Mike Hubank at the Institute of Child Health in UCL (p53 pathway) and Raya Khanin and Ernst Wit of the University of Glasgow and the University of Lancaster (*E. coli* repressor system).

REFERENCES

U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, London, 2006.

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. *Genome Biology*, 7(3):R25, 2006.

M.C. Coleman and D.E. Block. *Am. Inst. of Chem. Eng. Journal*, 52:651–667, 2006.

M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. In Michael I. Jordan, editor, *Learning in Graphical Models*, volume 89 of *Series D: Behavioural and Social Sciences*, pages 105–162. Kluwer, Dordrecht, The Netherlands, 1998.

R. Khanin, V. Viciotti, and E. Wit. *Proc. Natl. Acad. Sci. USA*, 103 (49):18592–18596, 2006.

N.D. Lawrence, G. Sanguinetti, and M. Rattray. Modelling transcriptional regulation using Gaussian processes. In B. Schölkopf, J.C. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems*, volume 19, Cambridge, MA, 2007. MIT Press.

J.C. Liao, R. Boscolo, Y.-L. Yang, L.-M. Tran, C. Sabatti, and V.P. Roychowdhury. *Proc. Natl. Acad. Sci. USA*, 100(26):15522–15527, 2003.

X. Liu, M. Milo, N.D. Lawrence, and M. Rattray. *Bioinformatics*, 21: 3637–3644, 2005.

P. Mendes and D. Kell. *Bioinformatics*, 14:869–883, 1998.

I. Nachman, A. Regev, and N. Friedman. *Bioinformatics*, 20(Suppl. 1):248–256, 2004.

C.E. Rasmussen and C.K.I. Williams. MIT Press, Cambridge, MA, 2006.

S. Rogers, R. Khanin, and M. Girolami. In *Probabilistic Modeling and Machine Learning in Structural and Systems Biology*, Tuusula, Finland, 17-18th June 2006.

T. Sandmann, L.J. Jensen, J.S. Jakobsen, M.M. Karzynski, M.P. Eichenlaub, P. Bork, and E.E. Furlong. *Dev. Cell*, 10:797–807, 2006.

G. Sanguinetti, N.D. Lawrence, and M. Rattray. *Bioinformatics*, 22 (22):2275–2281, 2006a.

G. Sanguinetti, M. Rattray, and N.D. Lawrence. *Bioinformatics*, 22 (14):1753–1759, 2006b.

P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S.E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S.E. Celniker, and G.M. Rubin. *Genome Biology*, 3:RESEARCH0088, 2002.

V. Vyshemirsky and M.A. Girolami. *Bioinformatics*, 24(6):833–839, 2008.