BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–4

A Gaussian Process Toolkit for Inference of Latent Chemical Species with Appplications in 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]Department of XXXXXXXX, Address XXXX etc.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

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 apply our approach to three different biological data sets. We focus on the problem of inference of transcription activity and consider signallying and developmental networks. Our results demonstrate the promise of the methodology ...

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`

Magnus pointed out that what we do is not just useful for latent functions, but also in the case where the system isn't closed loop. A closed loop system defines the type of function involved at each point, an open system (such as we look at) doesn't. So even if we had observations for all functions involved we would still need our GP priors for continuity. Need to embed this somewhere.

1 INTRODUCTION

The use of ordinary differential equations to model biological subsystems is already a well established approach \cite{}. More recently attention has turned to methodologies for fitting the parameters of these equations in the context of a particular experiment or set of experiments \cite{ maybe copasi?}.

Ordinary differential equations (ODE) are an approximation to the underlying chemical master equation. In an ODE the intrinsic variability in the reactants is ignored. However, this variability becomes less significant as the size of the system increases. Since most biological experiments cannot isolate expression levels in a single cell we are generally viewing data that is collated from a significant number of cells. Therefore, for many experiments of interest we consider the assumptions behind the approximation to be broadly valid.

Significant progress has been made on parameter estimation in the context of differential equations and 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 emirsky and Girolami(2008)s, y. This shows the potential for these techniques to be closely integrated with biological investigations, informing the process of biological experimental design.

A remaining problem for these approaches is dealing with the case where one or more of the chemical species in the system are unobserved. This is a common problem, 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 microarray), whereas others are much more difficult to measure on a large scale (*e.g.* protein concentrations such as transcription factors). 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.* et al.(2003)Liao, Boscolo, Yang, Tran, Sabatti, and Roychowdhurya, iuinetti et al.(2006b)Sanguinetti, Rattray, and Lawrence, n, auinetti et al.(2006a)Sanguinetti,

*to whom correspondence should be addressed

Lawrence, and Rattray,a) and algorithms designed for a sub set of genes (*e.g.* man et al.(2004)Nachman, Regev, and Friedmanc, ars et al.(2006)Rogers, Khanin, and Girolami, g, oin et al.(2006)Khanin, Viciotti, and Wit, a, hnco et al.(2006)Barenco, Tomescu, Brewer, Callard, Stark, and Hubank, r, a). Our approach is most directly inspired by nco et al.(2006)Barenco, Tomescu, Brewer, Callard, Stark, and Hubankr, a 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 (2006)o, l.

Gaussian processes ussen and Williams(2006)s, a 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. An attractive characteristic of the Gaussian process (GP) is that the result of any linear operator on the function leads to another GP, 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 nco et al.(2006)Barenco, Tomescu, Brewer, Callard, Stark, and Hubankr, a and reviewing the work done by ence et al.(2007)Lawrence, Sanguinetti, and Rattrayw, a who provided the first treatment of this model with Gaussian processes. Then, in Section 2.2, we extend our model to the case of repression. We follow in et al.(2006)Khanin, Viciotti, and Wita, h and apply the model in the context of ... ***brief description of Wit data***. In both these cases, no model of transcription factor translation is included. In our final example (Section 2.3) we show how, in the context of *Drosophila mesoderm* development a model of translation can be incorporated to improve the quality of the transcription factor inference.

2 METHODS

Consider the following simple model of gene regulation proposed by nco et al.(2006)Barenco, Tomescu, Brewer, Callard, Stark, and Hubankr, a,

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

here 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

$$x_j(t) = \frac{B_j}{D_j} + S_i \int_0^t \exp\left(-D_i(t-u)\right) f(u)\,\mathrm{d}u, \qquad (1)$$

where we have ignored any transient terms.

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. Gaussian processes 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, we will predominantly make use of the squared exponential covariance (sometimes known as the Gaussian or RBF covariance),

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

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 $f(t)$: it only involves integrals sums and products applied to $f(t)$. The properties of 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

$$k_{x_j,x_j}(t,t') = S_j^2 \int_0^t \exp\left(-D_j(t-u)\right) \int_0^{t'} \exp\left(-D_j\left(t'-u'\right)\right) k_{f,f}(t,t'$$

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

$$k_{x_j,f}(t,t') = \int_0^t \exp\left(-D_i(t-u)\right) k_{f,f}(t,t')\,\mathrm{d}u.$$

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

$$k_{x_i,x_j}(t,t') = S_i S_j \int_0^t \exp\left(-D_i(t-u)\right) \int_0^{t'} \exp\left(-D_j\left(t'-u'\right)\right) k_{f,f}(t,$$

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 $\frac{B_j}{D_j}$.

As was shown in ence et al.(2007)Lawrence, Sanguinetti, and Rattrayw, a, if $f(t)$ is drawn from a GP with a squared exponential covariance function, we can analytically define a probabilistic model of the expression data for which the parameters of the differential equation, B_j, S_j, D_j and the timescale l can all be determined, either through maximum likelihood or Bayesian sampling. We illustrate this by repeating their results here.

2.1 p53 Activity

The transcription factor p53 is a tumour repressor ... **GENERAL PROBLEM DESCRIPTION HERE**

2.1.1 Linear Model In this section we recreate the results presented by Lawrence et al.(2007)Lawrence, Sanguinetti, and Rattray for the linear model with several key differences.

Firstly, in the original paper Barenco et al.(2006)Barenco, Tomescu, Brewer, Callard, Stark, and Hubank 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.(2007)Lawrence, Sanguinetti, and Rattray but is included here. Secondly, Lawrence et al.(2007)Lawrence, Sanguinetti, and Rattray used an unormalised 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.(2006)Barenco, Tomescu, Brewer, Callard, Stark, and Hubank showed that they had also dealt with these anomalies so here we considered the normalised array data.

Below we show results of inference ...

RESULTS WITH LINEAR MODEL HERE.

2.1.2 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), when the predictions move away from the data they allow the TF activity to become negative. A potential solution to this problem is to place a GP over, for example, the log concentration 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 ence et al.(2007)Lawrence, Sanguinetti, and Rattrayw, a 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,

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

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

$$x_j(t) = \frac{B_j}{D_j} + S_i \int_0^t \exp(-D_i(t-u))\, g(f(u))\, \mathrm{d}u, \qquad (2)$$

but there is now a non-linear operation on $f(t)$ before the product and integral. The gene expression level is therefore no longer a GP. The MAP-Laplace approach involves finding a maximum *a posteriori* estimate for the function $f(t)$ and making a second order Taylor approximation at that point to the log likelihood. This approxmation is itself a Gaussian process and leads to an approximation to the marginal likelihood ussen and Williams(2006)s, a. 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, as well as providing a sanity check that the solutions are consistent, the MAP-Laplace approach is unconstrained in the choice of covariance function for $f(t)$. Naive implementation of the linear response model for a general covariance function would require, in general, numerical evaluation of the double integral in () and (). The use of the MAP-Laplace approach involves only a single numerical integral.

RESULTS ON MLP KERNEL WITH p53 DATA HERE

In this paper we build on the work of ence et al.(2007)Lawrence, Sanguinetti, and Rattrayw, a in two major ways. We go beyond the simple exponential response model considered exploring a Michaelis-Menten kinetics inspired response and a response that has been suggested as appropriate for repression (2006)o, l. We also extend the algorithm to learn the parameters of the differential equation by maximising the approximation to the log likelihood.

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

$$g(f(t)) = \ldots.$$

DESCRIPTION OF THE RESULTS HERE

2.2 Repression

DESCRIPTION OF REPRESSION PROBLEM AND RESULTS

2.3 Cascaded Differential Equations

As a final example we consider a simple cascade of differential equations. Returning to the framework of the linear model (although our non-linear approximation can be applied in a straightforward manner here) we consider the case where the TF activity is determined by its own mRNA. In other words we model the process of translation from mRNA to protein for the TF. The simple model we consider would only be appropriate in the case where the TF does not require activation after translation, for example by phosphorolation. The model is, therefore, not appropriate for many signalling pathways, but seems sensible in the context of development where TFs typically can act directly after translation. We therefore considered an example of the development of the mesoderm in *Drosophila Melanogaster*.

CASCADE RESULTS GO HERE

- Mention that we are aware that this is a short cascade ... reference Alon for Cascades in developmental networks? pg 90.

3 DISCUSSION

4 CONCLUSION

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

Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, London, 2006. ISBN 1-58488-642-0.

Martino Barenco, Daniela Tomescu, Daniel Brewer, Robin Callard, Jaroslav Stark, and Michael Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25, 2006.

Raya Khanin, Veronica Viciotti, and Ernst Wit. Reconstructing repressor protein levels from expression of gene targets in *E. Coli*. *Proc. Natl. Acad. Sci. USA*, 103(49):18592–18596, 2006. doi: 10.1073/pnas.0603390103.

Neil D. Lawrence, Guido Sanguinetti, and Magnus Rattray. Modelling transcriptional regulation using Gaussian processes. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems*, volume 19, Cambridge, MA, 2007. MIT Press.

James C. Liao, Riccardo Boscolo, Young-Lyeol Yang, Linh My Tran, Chiara Sabatti, and Vwani P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences USA*, 100 (26):15522–15527, 2003.

Iftach Nachman, Aviv Regev, and Nir Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20(Suppl. 1):248–256, 2004.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 026218253X.

Simon Rogers, Raya Khanin, and Mark Girolami. Model based identification of transcription factor activity from microarray data. In *Probabilistic Modeling and Machine Learning in Structural and Systems Biology*, Tuusula, Finland, 17-18th June 2006.

Guido Sanguinetti, Neil D. Lawrence, and Magnus Rattray. Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22(22):2275–2281, 2006a. doi: 10.1093/bioinformatics/btl473.

Guido Sanguinetti, Magnus Rattray, and Neil D. Lawrence. A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. *Bioinformatics*, 22(14):1753–1759, 2006b. doi: 10.1093/bioinformatics/btl154.

Vladislav Vyshemirsky and Mark A. Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 2008. doi: 10.1093/bioinformatics/btm607. Advance Access.