Gaussian Process Model for Inferring the Regulatory Activity of Transcription Factor Proteins

Neil D. Lawrence, Guido Sanguinetti
Department of Computer Science
University of Sheffield, U.K.

Magnus Rattray
School of Computer Science
University of Manchester, U.K.

19th May 2006

Abstract

In this paper we show how transcription factor protein concentration can be inferred from gene expression data using a Gaussian process model. We consider the situation where several target genes are regulated by a single transcription factor. We follow recent work by Barenco et al. (2006) in assuming that the relationship between the genes and the regulating protein can be modelled by a simple set of linear differential equations describing the rate of RNA production and decay for each target gene. We compare our results to those of Barenco et al. and we extend the model to deal with a non-linear relationship between the transcription factor concentration and the rate of RNA production.

1 Introduction

Recent advances in molecular biology have brought about a revolution in our understanding of cellular processes. Microarray technology now allows measurement of mRNA abundance on a genome-wide scale, and techniques such as chromatin immunoprecipitation (ChIP) have largely unveiled the wiring of the cellular regulatory network, identifying which genes are bound by which transcription factors. However, to gain a full quantitative description of the regulatory mechanism of transcription one needs to know the concentration levels of active transcription factor proteins, which is still a formidable challenge for the experimentalist. There is, therefore, a great deal of interest in inferring transcription factor protein concentrations indirectly from the expression levels of known targets of the transcription factor. This is often done following one of two complementary approaches. One can formulate a large scale simplified model of regulation (for example assuming a linear response to protein concentrations) and then combine network architecture data and gene expression data to infer transcription factors' protein concentrations on a genome-wide scale (see, for example, Sanguinetti et al., 2006, and references therein). Alternatively, one can attempt to devise a realistic model of the dynamics of a small subnetwork of the regulatory network.

In this paper we follow the second approach, focussing on the simplest subnetwork consisting of only one transcription factor, but using a detailed model of the interaction dynamics with an explicit parametric model of the growth and decay of mRNA concentration. A similar approach was recently taken by Barenco et al. (2006) and by Rogers et al. (2006). In these studies Markov chain Monte Carlo (MCMC) methods were used to carry out Bayesian inference, requiring substantial computational resources and limiting the inference to the discrete time-points where the data was collected. We show here how a Gaussian Process model provides a simple and computationally efficient method for Bayesian inference of continuous transcription factor concentration profiles and associated model parameters. This approach has the advantage of allowing for the inference of continuous quantities (concentration profiles) without discretization.

2 Linear Response Model

We follow Barenco et al. (2006) in first considering a linear differential equation that relates a given gene j's expression level $x_j(t)$ at time t to the concentration of the regulating transcription factor protein $f(t)$,

$$\frac{dx_j}{dt} = B_j + S_j f(t) - D_j x_j(t). \tag{1}$$

Here, B_j is the basal transcription rate of gene j, S_j is the sensitivity of gene j to the transcription factor and D_j is the decay rate of the mRNA. Assuming a linear relationship between rate of transcription and protein concentration is a crude simplification, but it can still lead to interesting results in certain modelling situations. We will consider how to deal with nonlinearities in section 5.

Barenco et al. used the linear differential equation (1) to model the expression levels of the targets of the tumour suppressor p53. They selected five targets of p53 and estimated their rates of transcription from their expression levels using polynomial curve fitting. They then used MCMC to estimate the protein concentration at each time point as well as the basal transcription rate, the sensitivity and the decay for each gene. Their approach has two significant limitations. Firstly, their model includes no concept of temporal continuity that would allow the protein concentration level to be interpolated between time points. Secondly, they needed to make use of numerical solutions to solve their differential equation and had to perform up to 10 million steps of MCMC in the parameter estimation stage.

In this paper we show how the entire model can be resolved using a Gaussian process prior on the protein concentration. We achieve similar results to Barenco et al. (2006) without having to resort to numerical solutions of the equation or MCMC.

3 Gaussian Process Inference

The equation given in (1) can be solved to recover

$$x_j(t) = \frac{B_j}{D_j} + k_j \exp(-D_j t) + S_j \exp(-D_j t) \int_0^t f(u) \exp(D_j u)\, du \tag{2}$$

where k_j arises from the initial conditions, and is zero if we assume an initial baseline expression level $x_j(0) = B_j/D_j$. Importantly this equation involves only linear operations on the function $f(t)$. This means that if we place a Gaussian process prior over $f(t)$ this directly implies a Gaussian process prior over $x_j(t)$. If the covariance function associated with $f(t)$ is given by $k_{ff}(t,t')$ then the covariance function associated with $x_j(t)$ is given by

$$k_{x_j x_j}(t,t') = S_j^2 \exp(-D_j(t+t')) \int_0^t \exp(D_j u) \int_0^{t'} \exp(D_j u')\, k_{ff}(u,u')\, du' du,$$

however to predict $f(t)$ given instantiations of $\{x_j(t)\}_{j=1}^m$ we also need the 'cross-covariance' terms between $x_i(t)$ and $x_j(t')$, $k_{x_i x_j}(t,t')$ and between $x_j(t)$ and $f(t')$, $k_{x_j f}(t,t')$:

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

$$k_{x_j f}(t,t') = S_j \exp(-D_j t) \int_0^t \exp(D_j u)\, k_{ff}(u,t')\, du\ .$$

Furthermore, if the process prior over $f(t)$ is taken to be a squared exponential kernel,

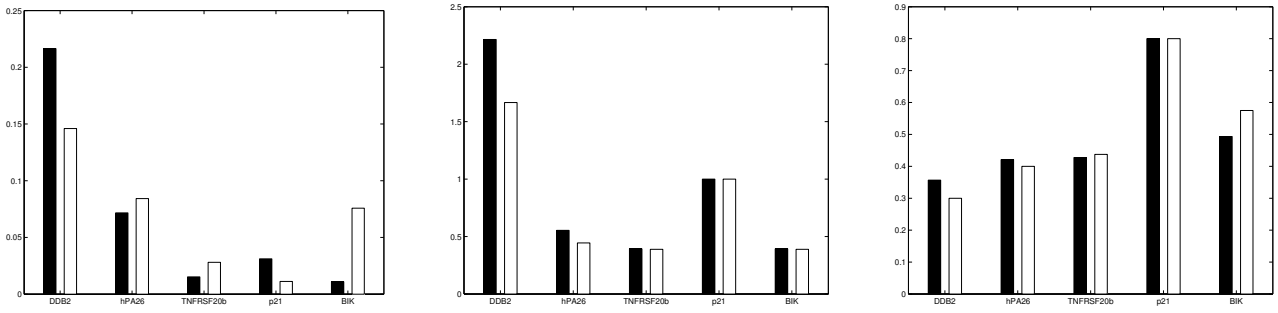
$$k_{ff}(t,t') = \exp\left(-\frac{(t-t')^2}{\sigma^2}\right),$$

where σ controls the width of the basis functions[1], all the integrals can be computed analytically. We can therefore compute a likelihood which relates instantiations from all the observed genes, $\{x_j(t)\}_{j=1}^m$, through dependencies on the parameters $\{B_j, S_j, D_j\}_{j=1}^m$. The effect of $f(t)$ has been marginalised. This allows us to avoid sampling over $f(t)$ in the manner of Barenco et al. (2006) and Rogers et al. (2006), both of which necessarily restrict samples of $f(t)$ to the instantiations of $f(t)$ that are associated with time points where measurements are taken.

4 Experiments

To demonstrate the efficacy of our approach, we followed Barenco et al. (2006) in analysing five targets of p53: *DDB2*, *p21 SESN1/hPA26*, *BIK* and *TNFRSF10b*. Results, including comparisons with results of Barenco et al. are given in Figures 1 and 2.

[1] The scale of the process is ignored to avoid a parameterisation ambiguity with the sensitivities.



(a) (b) (c)

Figure 1: Results of inference on p53 data of Barenco et al.. The bar charts show (a) Basal transcription rates from our model and that of Barenco et al.. Black shows results from our model, white is Barenco et al.. (b) Similar for sensitivities. (c) Similar for decay rates.

As can be seen from the bar plots above, our results in terms of degradations, sensitivities and basal transcription rates are strikingly similar to those of Barenco et al.. However, ours were obtained in approximately 13 minutes in MATLAB on an AMD Athlon 1.79 GHz machine using approximately 600 iterations conjugate gradient optimisation (see `http://www.dcs.shef.ac.uk/~neil/gpsim` for the code used). Timings aren't given by Barenco et al.; they used 10 million iterations of MCMC to obtain their results.

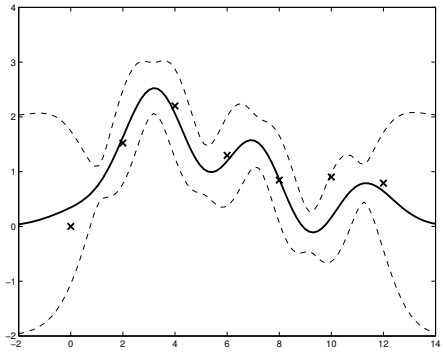


Figure 2: Predicted concentration for p53. Solid line is mean prediction, dashed lines are at two standard deviations. The prediction of Barenco et al. was pointwise and is shown as crosses.

5 Non-linear Response Model

It is more realistic to assume a non-linear relationship between the transcription factor concentration and rate of RNA production. In this case we have a function parameterised by a target-specific vector $\boldsymbol{\theta}_j$, so that,

$$\frac{dx_j}{dt} = B_j + g(f(t), \boldsymbol{\theta}_j) - D_j x_j, \quad x_j(t) = \frac{B_j}{D_j} + \mathrm{e}^{-D_j t} \int_0^t \mathrm{d}u \, g(f(u), \boldsymbol{\theta}_j) \, \mathrm{e}^{D_j u} \ , \tag{3}$$

where we again set $x_j(0) = B_j/D_j$. A reasonable choice of non-linear function is the Michaelis-Menten form considered by Rogers et al. (2006). It is also common to measure expression on a log-scale and to assume multiplicative log-normal noise. In this case it is no longer possible to represent the posterior distribution of $x_j(t)$ and $f(t)$ as a Gaussian process. However, we can derive the functional gradient of the likelihood and prior, and use this to learn the Maximum a Posteriori (MAP) solution for $f(t)$ and other parameters by (functional) gradient descent. Given multiplicative noise-corrupted logged data $\hat{y}_{ij} = \log(x_j(t_i)) + \epsilon_{ij}$ at times t_i we define the functional $E(f) = -\sum_i \log p(\hat{y}_{ij}|f, \theta, B, D, \lambda_{ij}) - \log p(f|k)$ to be minimised. We find the functional gradient,

$$\frac{\delta E}{\delta f(t)} = \sum_{j=1}^{m} \sum_{i=1}^{T} \Theta(t_i - t) \left(\frac{\lambda_{ij} \left(\log x_j(t_i) - \hat{y}_{ij} \right)}{x_j(t_i)} \right) \frac{\partial g(f, \boldsymbol{\theta}_j)}{\partial f} \mathrm{e}^{-D_j (t_i - t)} + \int \mathrm{d}u \, k^{-1}(t, u) f(u) \ , \tag{4}$$

where $\Theta(x)$ is the Heaviside function and the λ_{ij}s are the inverse noise variances. Estimates of the inverse noise variances λ_{ij} can be obtained from the probe-level processing of the microarray data using *e.g.* the recently proposed method of Liu et al. (2005) and can be viewed as additional input data, rather than parameters to be optimised. Second order variations can be used to associate credibility intervals with the estimated profile and we are currently investigating a MAP-Laplace method for model selection in the non-linear case.

Acknowledgements

We thank Martino Barenco for useful discussions and for providing the data. This work was supported by a BBSRC award "Improved procesing of microarray data using probabilistic models".

References

Barenco, M., Tomescu, D., Brewer, D., Callard, R., Stark, J., and Hubank, M. (2006). Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25.

Liu, X., Milo, M., Lawrence, N. D., and Rattray, M. (2005). A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644.

Rogers, S., Khanin, R., and Girolami, M. (2006). Model based identification of transcription factor regulatory activity via Markov Chain Monte Carlo. Presentation at MASAMB '06.

Sanguinetti, G., Rattray, M., and Lawrence, N. D. (2006). A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. *Bioinformatics*. In press.