# Modelling Transcriptional Regulation Using Gaussian Processes

Neil D. Lawrence, Guido Sanguinetti and Magnus Rattray
{neill,magnus}@cs.man.ac.uk, guido@dcs.shef.ac.uk

School of Computer Science
University of Manchester, U.K. and Department of Computer Science
University of Sheffield, U.K.

## Overview

Modelling the dynamics of transcriptional processes in the cell requires the knowledge of a number of key biological quantities. While some of them are relatively easy to measure, such as mRNA decay rates and mRNA abundance levels, it is still very hard to measure the active concentration levels of the transcription factor proteins that drive the process and the sensitivity of target genes to these concentrations. In this poster we show how these quantities for a given transcription factor can be inferred from gene expression levels of a set of known target genes. We treat the protein concentration as a latent function with a Gaussian process prior, and include the sensitivities, mRNA decay rates and baseline expression levels as hyperparameters. We apply this procedure to a human leukemia dataset, focusing on the tumour repressor p53 and obtaining results in good accordance with recent biological studies.

## Introduction

- Understanding of cellular processes is improving through microarrays, chromatin immunoprecipitation *etc.*.
- Quantitative description of regulatory mechanisms requires:
  - transcription factor (TF) concentrations,
  - gene-specific constants such as the baseline expression, mRNA decay rate and sensitivity to TF concentrations (TFCs).
- These quantities are hard to *measure directly.*
- We show how they can be *inferred* using a systems biology model and Gaussian processes (GPs).
- Our work provides a extensible, principled framework for attacking this problem.

## Gaussian Processes

- We use GPs for inference of TFC profiles and model parameters.
  - GPs allow for inference of continuous profiles, accounting naturally for temporal structure.
  - GPs avoid cumbersome interpolation techniques to estimate mRNA production rates from mRNA abundance data.
  - GPs allows us to deal naturally with the noise inherent in the measurements.
  - GPs outstrip pure MCMC techniques for computational efficiency. This will be crucial in future extensions to more complex (and realistic) regulatory networks.
- GPs have previously been proposed for solving differential equations Graepel [2003].

## Linear Response Model

- Data consists of $T$ measurements of mRNA expression level for $N$ different genes.
- We relate gene expression, $x_j(t)$, to TFC, $f(t)$, by

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

$B_j$ basal transcription rate of gene $j$,
$S_j$ is sensitivity of gene $j$
$D_j$ is the decay rate of the mRNA.

- Dependence of mRNA transcription rate on TF is linear.
- Model was proposed by Barenco et al. [2006] for tumour suppressor TF p53 and five targets genes.

## Linear Response Solution

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

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

- If we model $f(t)$ as a GP then as (2) only involves linear operations $x(t)$ is also a GP.

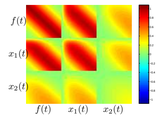
## Covariance for Transcription Model

- RBF Kernel function for $f(t)$

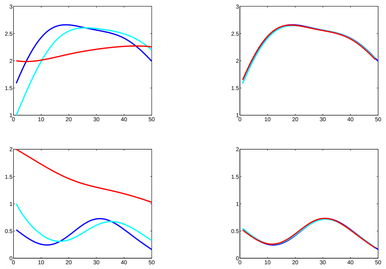
$$x_i(t) = \frac{B_i}{D_i} + S_i \exp(-D_i t) \int_0^t f(u) \exp(D_i u)\, du.$$

- Joint covariance for $x_1(t)$, $x_2(t)$ and $f(t)$.
- Values for this covariance

| $D_1$ | $S_1$ | $D_2$ | $S_2$ |
|---|---|---|---|
| 5 | 5 | 0.5 | 0.5 |




## Joint Sampling of $x(t)$ and $f(t)$



*Left*: joint samples from the transcription covariance, *blue*: $f(t)$, *cyan*: $y_1(t)$ and *red*: $y_2(t)$. *Right*: numerical solution for $f(t)$ of the differential equation from $x_1(t)$ and $x_2(t)$ (blue and cyan). True $f(t)$ included for comparison.

## Covariance Function Computation

- We rewrite equation (2) as

$$x_j(t) = \frac{B_j}{D_j} + L_j[f](t)$$

where

$$L_j[f](t) = S_j \exp(-D_j t) \int_0^t f(u) \exp(D_j u)\, du \tag{3}$$

is a linear operator.

- The new covariance function is then given by

$$\operatorname{cov}\left(L_j[f](t), L_k[f](t')\right) = L_j \otimes L_k\left[k_{tt}\right](t,t').$$

more explicitly

$$k_{x_j x_k}(t,t') = S_j S_k \exp(-D_j t - D_k t') \int_0^t \exp(D_j u) \int_0^{t'} \exp(D_k u')\, k_{tt}(u,u')\, du' du. \tag{4}$$

- For an RBF covariance for $f(t)$

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

these integrals are tractable.

## Noise Corruption

- Allow the mRNA abundance of each gene at each time point to be corrupted by noise, for observations at $t_i$ for $i = 1, \ldots, T$,

$$y_j(t_i) = x_j(t_i) + \epsilon_j(t_i) \tag{5}$$

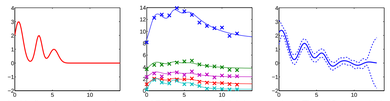
with $\epsilon_j(t) \sim \mathcal{N}\left(0, \sigma_j^2\right)$.

- Estimate noise level using probe-level processing techniques of Affymetrix microarrays (*e.g.* mmgMOS, Liu et al. [2005]).
- The covariance of the noisy process is then $K_{\mathbf{yy}} = \Sigma + K_{\mathbf{xx}}$, with $\Sigma = \operatorname{diag}\left(\sigma_{11}^2, \ldots, \sigma_{1T}^2, \ldots, \sigma_{N1}^2, \ldots, \sigma_{NT}^2\right)$.

## Artificial Data

- Results from an artificial data set.
  - We used a 'known TFC' and derived six 'mRNA profiles'.
  - Fourteen subsamples were taken and corrupted by noise.
  - This 'data' was then used to infer a distribution over plausible TFCs.

## Artificial Data Results



In this figure we illustrate inference of TFC from an artificial data set. *Left:* The TFC, $f(t)$, which drives the system, it is given by the sum of three Gaussian basis functions. *Middle:* Six gene mRNA concentration profiles each obtained by using different parameter sets $\{B_i, S_i, D_i\}_{i=1}^6$ (lines) along with noise corrupted 'data' for the system (crosses). *Right:* The inferred TFC (with error bars). We used our Gaussian process framework to learn the system's parameters and infer the TFC.

## Non-linear Response Model

- Linear model does not account for saturation.
- All the quantities in equation (1) are positive, but direct samples from a GP will not be.
- *Solution*: model response using a positive nonlinear function.

## Formalism

- Introduce a non-linearity $g(\cdot)$ parameterised by $\boldsymbol{\theta}_j$

$$\begin{aligned} \frac{dx_j}{dt} &= B_j + g(f(t), \theta_j) - D_j x_j \\ x_j(t) &= \frac{B_j}{D_j} + \exp(-D_j t) \int_0^t du\, g(f(u), \theta_j) \exp(D_j u) \ldots \end{aligned} \tag{6}$$

- The induced distribution of $x_j(t)$ is no longer a GP.
- Derive the functional gradient and learn a MAP solution for $f(t)$.
- Also compute Hessian so we can approximate the marginal likelihood.

## Log Likelihood

- Given noise-corrupted data $y_j(t_i)$ the log-likelihood is

$$\log p(\mathbf{Y}|f, \{B_j, \theta_j, D_j, \Xi\}) = -\frac{1}{2} \sum_{i=1}^{T} \sum_{j=1}^{N} \left[ \frac{(x_j(t_i) - y_j(t_i))^2}{\sigma_j^2} - \log\left(\sigma_j^2\right) \right] - \frac{NT}{2} \log(2\pi) \tag{7}$$

$\Xi$ — kernel parameters.

- The functional derivative of the log-likelihood wrt $f$ is

$$\frac{\delta \log p(\mathbf{Y}|f)}{\delta f(t)} = -\sum_{i=1}^{T} \Theta(t_i - t) \sum_{j=1}^{N} \frac{(x_j(t_i) - y_j(t_i))}{\sigma_j^2} g'(f(t)) \mathrm{e}^{-D_j(t_i - t)} \tag{8}$$

$\Theta(x)$ — Heaviside step function.

- The negative Hessian of the log-likelihood wrt $f$ is

$$\begin{aligned} w(t,t') = -\frac{\delta^2 \log p(\mathbf{Y}|f)}{\delta f(t)\, \delta f(t')} &= \sum_{i=1}^{T} \Theta(t_i - t) \delta(t - t') \sum_{j=1}^{N} \frac{(x_j(t_i) - y_j(t_i))}{\sigma_j^2} g''(f(t)) \mathrm{e}^{-D_j(t_i - t)} \\ &+ \sum_{i=1}^{T} \Theta(t_i - t) \Theta(t - t') \sum_{j=1}^{N} {\sigma_j}^{-2} g'(f(t)) g'(f(t')) \mathrm{e}^{-D_j(2t_i - t - t')} \end{aligned} \tag{9}$$

$g'(f) = \partial g / \partial f$ and $g''(f) = \partial^2 g / \partial f^2$.

## Implementation

- Implementation requires a discretised time.
- Compute the gradient and Hessian on a grid.
- Integrate them by approximate Riemann quadrature.
- We choose a uniform grid $\{t_p\}_{p=1}^M$ so that $\Delta = t_p - t_{p-1}$ is constant.
- The vector $\mathbf{f} = \{f_p\}_{p=1}^M$ is the function $f$ at the grid points.
- The gradient of the log-likelihood is then given by,

$$\frac{\partial \log p(\mathbf{Y}|\mathbf{f})}{\partial f_p} = -\Delta \sum_{i=1}^{T} \Theta(t_i - t_p) \sum_{j=1}^{N} \frac{(x_j(t_i) - y_j(t_i))}{\sigma_j^2} g'(f_p) \mathrm{e}^{-D_j(t_i - t_p)} \tag{10}$$

- Negative Hessian of the log-likelihood is,

$$\begin{aligned} W_{pq} = -\frac{\partial^2 \log p(\mathbf{Y}|\mathbf{f})}{\partial f_p \partial f_q} &= \delta_{pq} \Delta \sum_{i=1}^{T} \Theta(t_i - t_q) \sum_{j=1}^{N} \frac{(x_j(t_i) - y_j(t_i))}{\sigma_j^2} g''(f_q) \mathrm{e}^{-D_j(t_i - t_q)} \\ &+ \Delta^2 \sum_{i=1}^{T} \Theta(t_i - t_p) \Theta(t_i - t_q) \sum_{j=1}^{N} {\sigma_j}^{-2} g'(f_p) g'(f_q) \mathrm{e}^{-D_j(2t_i - t_p - t_q)} \end{aligned} \tag{11}$$

where $\delta_{pq}$ is the Kronecker delta.

## Implementation II

- Combine these with prior to compute gradient and Hessian of log posterior $\Psi(\mathbf{f}) = \log p(Y|\mathbf{f}) + \log p(\mathbf{f})$ [see Rasmussen and Williams, 2005, chapter 3]

$$\begin{aligned} \frac{\partial \Psi(\mathbf{f})}{\partial \mathbf{f}} &= \frac{\partial \log p(Y|\mathbf{f})}{\partial \mathbf{f}} - K^{-1}\mathbf{f} \\ \frac{\partial^2 \Psi(\mathbf{f})}{\partial \mathbf{f}^2} &= -(W + K^{-1}) \end{aligned} \tag{12}$$

$K$ prior covariance evaluated at the grid points.

- Use to find a MAP solution via, $\hat{\mathbf{f}}$, using Newton's algorithm.
- The Laplace approximation is then

$$\log p(Y) \simeq \log p(Y|\hat{\mathbf{f}}) - \tfrac{1}{2}\hat{\mathbf{f}}^{\mathsf{T}} K^{-1} \hat{\mathbf{f}} - \tfrac{1}{2} \log |I + KW|. \tag{13}$$

## Example: exponential response

- Exponential response constrains protein concentration positive.

$$g(f(t), \theta_j) = S_j \exp(f(t)) \tag{14}$$

## Results

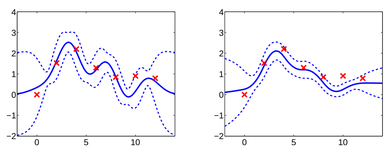
- Recently published biological data set studied using linear response model by Barenco et al. [2006].
- Study focused on the tumour suppressor protein p53.
- mRNA abundance measured at regular intervals using Affymetrix U133A oligonucleotide microarrays.
- Five known target genes of p53: *DDB2*, *p21*, *SESN1/hPA26*, *BIK* and *TNFRSF10b*.
- They used quadratic interpolation for the mRNA production rates, discretised the model, and used MCMC sampling to obtain estimates of the model parameters $B_j$, $S_j$, $D_j$ and $f(t)$.
- Sensitivity and mRNA decay of one target gene, p21, was fixed and $f(0)$ was constrained to be zero.

## Linear response analysis

- We analysed data using the linear response model
- Raw data was processed using the mmgMOS model of Liu et al. [2005] which provides variance as well as expression level.
- Inferred posterior mean function shown below for RBF kernels.
- Note oscillatory behaviour, possible artifact of RBF covariance [see page 123 in Rasmussen and Williams, 2005].
- We repeated the experiment using the "MLP" covariance function Williams [1998].

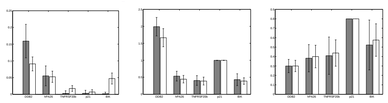
$$k(t,t') = \arcsin\left(\frac{wtt' + b}{\sqrt{(wt^2 + b + 1)(wt'^2 + b + 1)}}\right) \tag{15}$$

## Linear Response Results



Predicted protein concentration for p53 using a linear response model: *Left:* RBF prior on $f$; *Right:* MLP prior on $f$. Solid line is mean prediction, dashed lines are 95% credibility intervals. The prediction of Barenco *et al.* was pointwise and is shown as crosses.

## Linear Response Results II



Results of inference on the hyperparameters for p53 data studied in Barenco et al. [2006]. The bar charts show *Left:* Basal transcription rates from our model and that of Barenco *et al.*. Grey are estimates obtained with our model, white are the estimates obtained by Barenco *et al.* *Middle:* Similar for sensitivities. *Right:* Similar for decay rates.

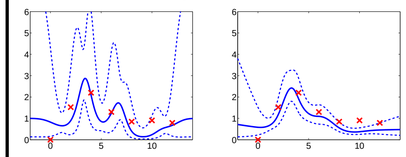
## Linear Response Discussion

- Figure above shows the results of inference on the values of the hyperparameters $B_j$, $S_j$ and $D_j$.
- Samples from the posterior distribution were obtained using Hybrid Monte Carlo (see *e.g.* Neal, 1996).
- Results are in good accordance with the results obtained by Barenco *et al.*
- Differences in estimates of the basal transcription rates probably due to the different methods used for probe-level processing of the microarray data.

## Non-linear response analysis

- Exponential response model (constrains protein concentrations positive).
- Inferred MAP solutions for the latent function $f$ are plotted below.

## Non-linear Response Results



Predicted protein concentration for p53 using an exponential response: *Left:* shows results of using a squared exponential prior covariance on $f$; *Right:* shows results of using an MLP prior covariance on $f$. Solid line is mean prediction, dashed lines show 95% credibility intervals. The results shown are for $\exp(f)$, hence the asymmetry of the credibility intervals. The prediction of Barenco *et al.* was pointwise and is shown as crosses.

## Discussion

- We have described how GPs can be used in modelling dynamics of a simple regulatory network motif.
- Our approach has advantages over standard parametric approaches:
  - there is no need to restrict the inference to the observed time points, the temporal continuity of the inferred functions is accounted for naturally.
  - GPs allow us to handle uncertainty in a natural way.
  - MCMC parameter estimation in a discretised model can be computationally expensive. Parameter estimation can be achieved easily in our framework by type II maximum likelihood or by using efficient hybrid Monte Carlo sampling techniques
- All code on-line `http://www.dcs.shef.ac.uk/~neil/gpsim/`.

## Future Directions

- This is still a very simple modelling situation.
  - We are ignoring transcriptional delays.
  - Should consider more biologically meaningful nonlinearities (*e.g.* Michaelis-Menten model used in Rogers et al. [2006]).
  - Here we have single transcription factor: our ultimate goal is to describe regulatory pathways with more genes.
- All these issues can be dealt with in the general framework we have described.
  - Need to overcome the greater computational difficulties.

## Acknowledgements

We thank Martino Barenco for useful discussions and for providing the data. We gratefully acknowledge support from BBSRC Grant No BBS/B/0076X "Improved processing of microarray data with probabilistic models".

## References

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25, 2006. 1

T. Graepel. Solving noisy linear operator equations by Gaussian processes: Application to ordinary and partial differential equations. In T. Fawcett and N. Mishra, editors, *Proceedings of the International Conference in Machine Learning*, volume 20, pages 234–241. AAAI Press, 2003. ISBN 1-57735-189-4. 1

X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644, 2005. 1

R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996. Lecture Notes in Statistics 118. 1

C. E. Rasmussen and C. K. Williams. *Gaussian Processes for Machine Learning*. MIT press, 2005. 1

S. Rogers, R. Khanin, and M. 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. 1

C. K. I. Williams. Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216, 1998. 1