Gaussian Process Modelling of Latent Chemical Species

Applications to Inferring Transcription Factor Activities

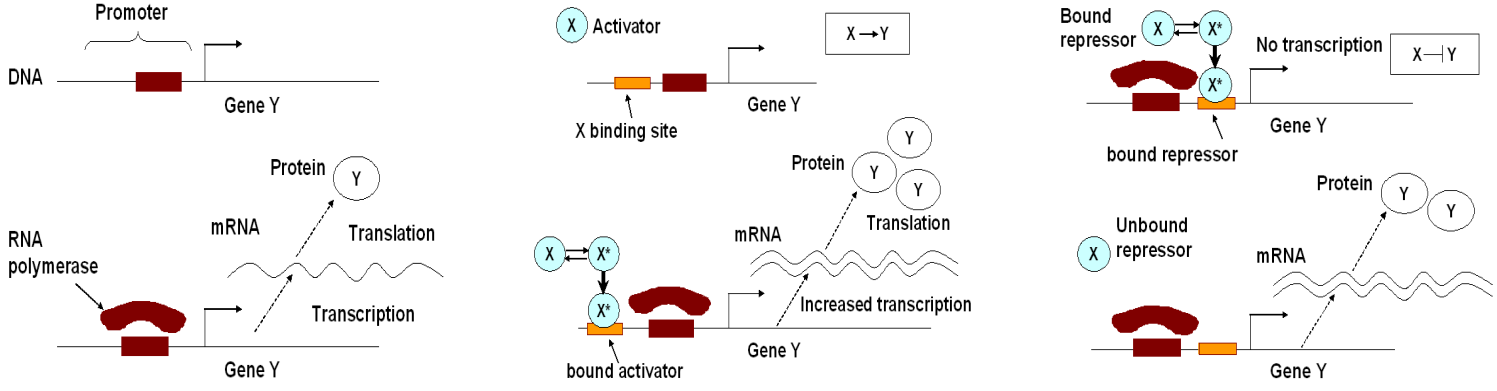
Pei Gao, Antti Honkela, Magnus Rattray and Neil Lawrence

March 25, 2008

School of Computer Science, University of Manchester

Gene transcription regulation

- Gene transcription regulation



Figures are from An Introduction to Systems Biology, by U. Alon, 2006.

- Algorithms for both activation and repression cases.
- Single-input Module(SIM) network motif.

Notations

- $x_j(t)$: gene j's expression
- $f(t)$: Transcription Factor (TF) related activity, drawn from a Gaussian Process.
  - Linear response: TF's activity
  - Nonlinear response: log of the TF activity.
- Important parameters
  - $S_j(t)$: sensitivity of the gene j to its governing TF.
  - $D_j(t)$: decay rate of the mRNA j.
  - $B_j(t)$: basal transcription rate of the gene j.

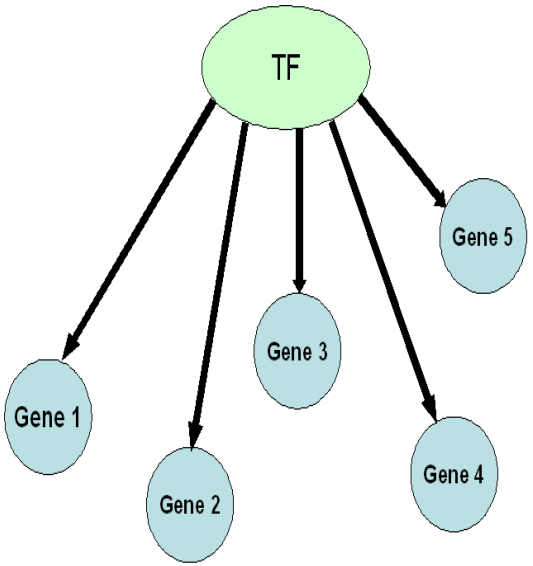
Linear Activation Response

The linear model is considered by Barenco et al. [2006]

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

This differential equation can be solved for $x_j(t)$ as

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



Gaussian Process Inference for the Linear Model

Any linear opearation of a GP $\Longrightarrow$ Related GP

$$f(t) \backsim \mathcal{GP}\left(0, k_{ff}\left(t, t'\right)\right) \Longrightarrow x_j(t) \backsim \mathcal{GP}\left(\frac{B_j}{D_j}, k_{xx}\left(t, t'\right)\right)$$

Hence, under the linear model, the prediction of the TF can be obtained by using

$$\begin{bmatrix} f \\ \mathbf{x} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \frac{\mathbf{B}}{\mathbf{D}} \end{bmatrix}, \begin{bmatrix} K_{ff} & K_{f\mathbf{x}} \\ K_{\mathbf{x}f} & K_{\mathbf{xx}} \end{bmatrix}\right)$$

Standard GP Regression yields

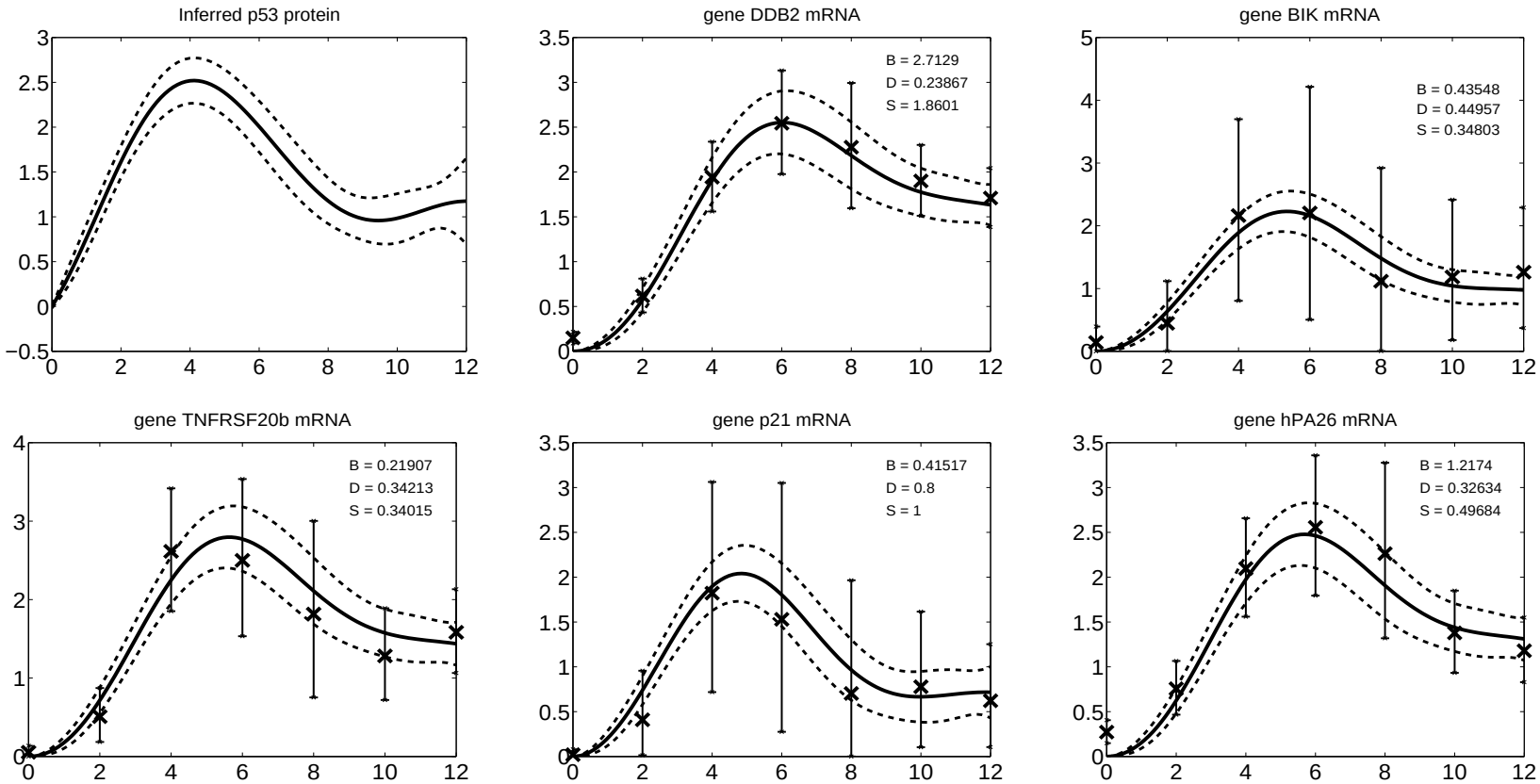
$$\begin{aligned} \langle f \rangle_{post} &= K_{f\mathbf{x}} K_{\mathbf{xx}}^{-1}\left(\mathbf{x} - \frac{\mathbf{B}}{\mathbf{D}}\right) \\ K_{ff}^{post} &= K_{ff} - K_{f\mathbf{x}} K_{\mathbf{xx}}^{-1} K_{\mathbf{x}f} \end{aligned}$$

Example: Inferring p53 activity using the linear model

Experiment Settings

- TF p53 is activated during DNA damege (irradiation).
- Seven samples of the expression levels of the target genes are collected in three replicas.
- Five target genes are selected to train the model.
- $f(t)$ is constraint to be zero at time $t = 0$.
- Data points and error bars are pre-processed by the PUMA package [Liu et al.,2005]. Simple median based normalised version of the Affymetrix array data is used.

Results for p53 using the Linear Model.



Nonlinear Response Model

Advantages of the linear activation model

- Simplicity.
- Allow the joint distribution over the gene expression and the TF activity to be determined analytically.

Disadvantages of the linear activation model

- GP cannot constrain the function to be positive.
- Not capable for the repression cases.

Consider the following modification to the model,

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = B_j + S_j g(f(t)) - D_j x_j(t), \tag{3}$$

Nonlinear Activation Model

The differential equation can still be solved,

$$x_j(t) = \frac{B_j}{D_j} + S_j \int_0^t e^{-D_j(t-u)} g_j(f(u)) \, \mathrm{d}u \tag{4}$$

where $g(\cdot)$ is a non-linear function. For the activation case, Michaelis Menten kinetics Model takes the non-linearity of

$$g_j(f(t)) = \frac{e^{f(t)}}{\gamma_j + e^{f(t)}}, \tag{5}$$

where we are using a GP to model the log of the TF activity, i.e. $f(t)$.

MAP-Laplace Approximation

Based on the Laplace's method,

$$p(f \mid y) = N(\hat{f}, A^{-1}) \propto \exp\left(-\frac{1}{2}\left(f - \hat{f}\right)^T A \left(f - \hat{f}\right)\right) \qquad (6)$$

- Inference of the TF: Optimize the Laplace approximation to the logged un-normalized posterior using Newton's method.
- Estimation of parameters: Maximise the marginal likelihood using the scaled conjugate gradient algorithm.

Results for p53 using Nonlinear Activation Model

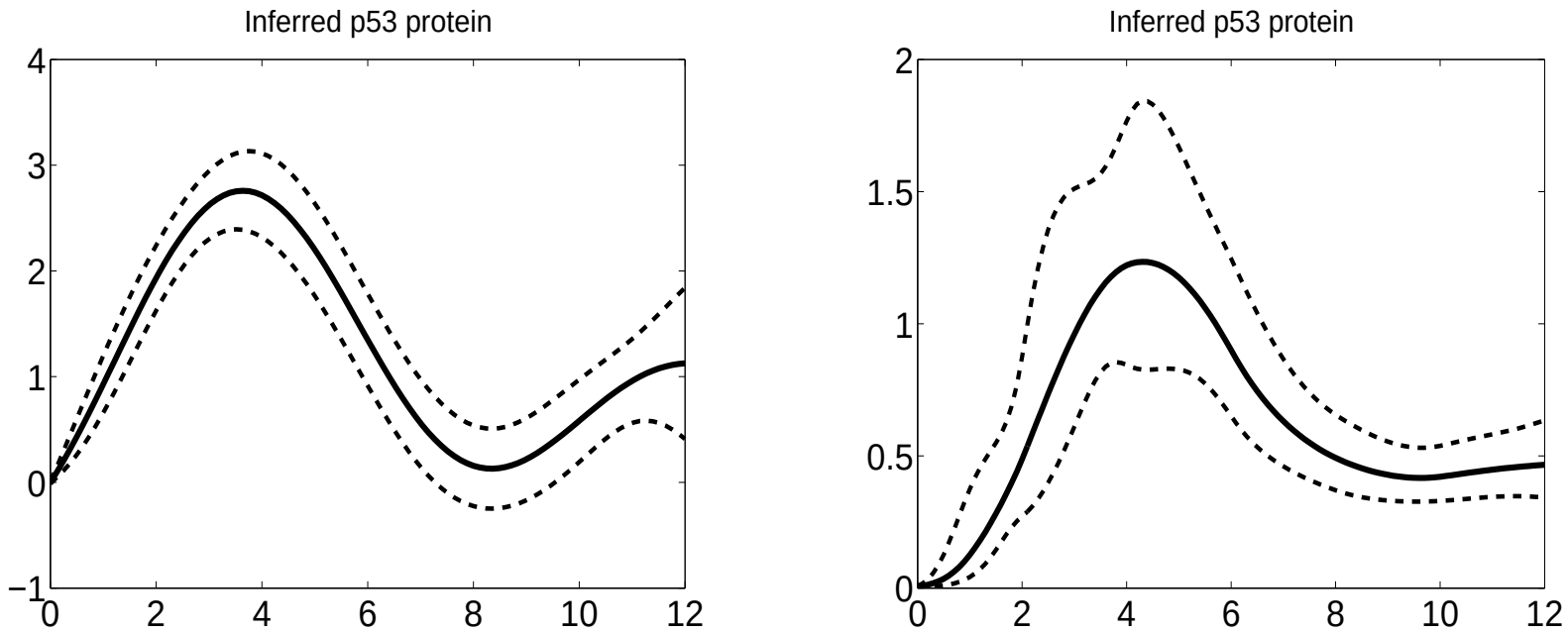


Figure: Inferring p53 activity for a different replicate.

Nonlinear Repression Model

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

$$g_j\left(f\left(t\right)\right) = \frac{1}{\gamma_j + e^{f(t)}} \tag{7}$$

In the case of repression we have to include the transient terms as

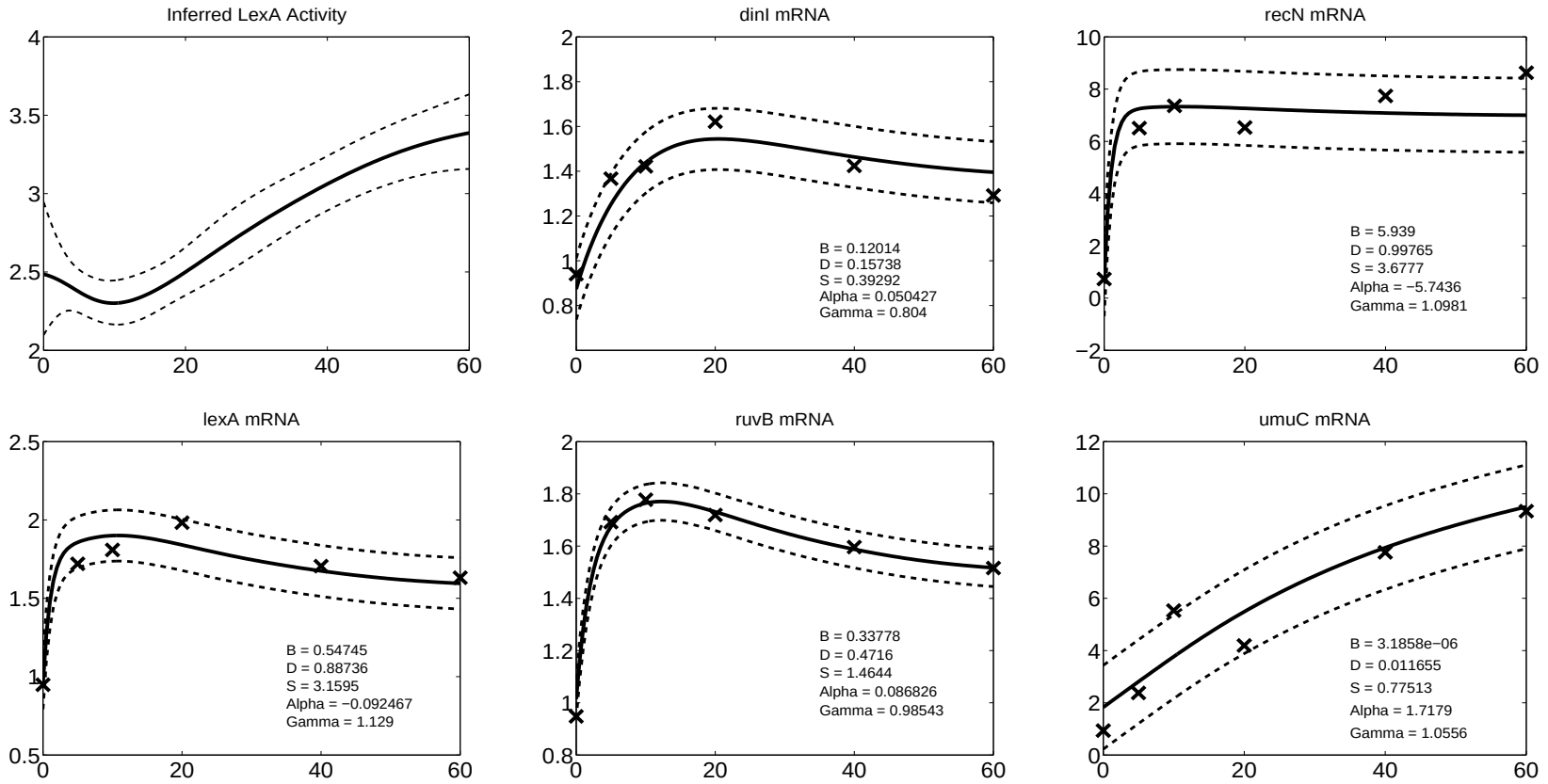
$$x_j\left(t\right) = \alpha_j e^{-D_j t} + \frac{B_j}{D_j} + S_j \int_0^t e^{-D_j(t-u)} g_j(f\left(u\right))\mathrm{d}u \tag{8}$$

Example: Inferring the Repressor LexA Activity

Experiment Settings

- In the context of the SOS system in *E. coli*, genes are controlled by the transcriptional repressor protein LexA.
- Six time points of the mRNA measurements for the target genes are collected.
- We follow Khanin et al.[2006] to use 14 target genes to train the model.
- The log of the TF is assigned as a GP with no other hard constraints.
- Gene-specific noise variance parameter is also estimated for each target gene.

Results for the repressor LexA



Cascaded Differential Equations

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

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

The solution for $f(t)$, setting transient terms to zero, is

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

Example: Inferring the Mef2 Activity

Experiment Settings

- Focusing on the TF Mef2 (Myocyte enhancing factor 2).
- Six targets of Mef2 are selected. They are identified by ChIP-chip assays and were observed to be up-regulated after Mef2 is expressed.
- Affymetrix time course microarray data from wild-type embryos [Tomancak et al., 2002] provide us with observations of Mef2 expression ($y(t)$, the driving input) and expressions of the target genes at hourly intervals in three replicas.

Results for Mef2 by using the Cascaded Model

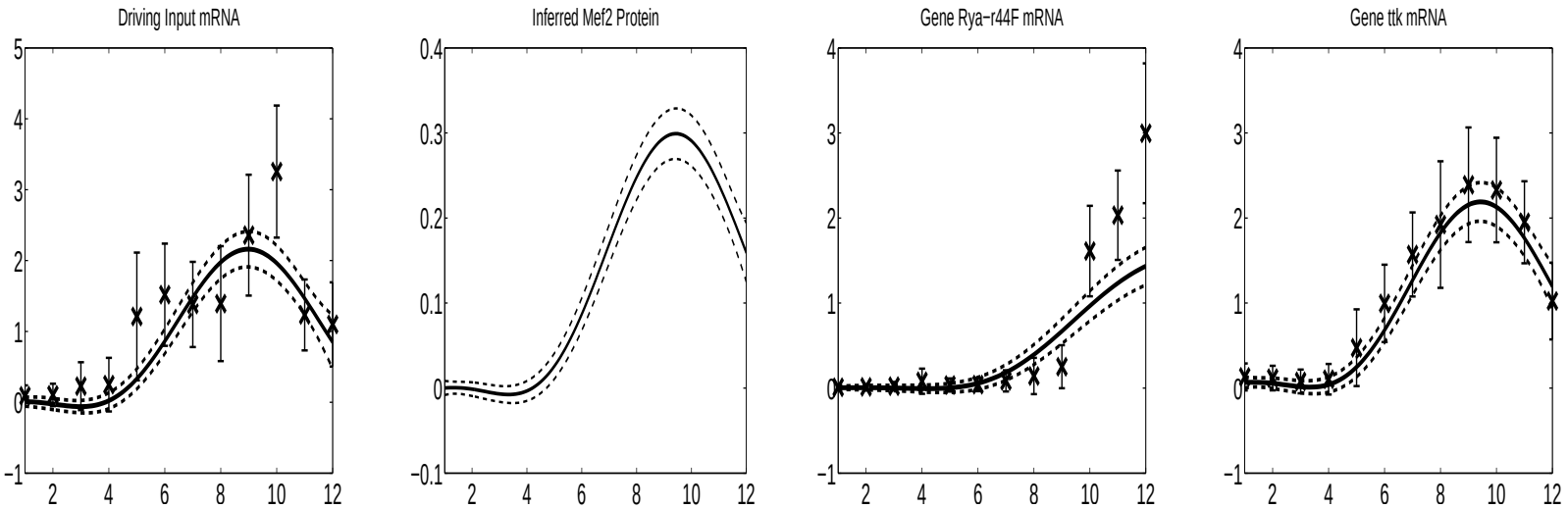


Figure: Results for Mef2 from one replicate. Data points and error bars are obtained from probe-level analysis of the Affymetrix data using the puma package [Liu et al., 2005].

Discussion and Future Work

- Applications to target identification.
- Scaling up to larger systems.
- Applications to other types of system, *e.g.* non-steady-state metabolomics, spatial systems etc.

Acknowledgement

- Neil Lawrence and Magnus Rattray.
- Antti Honkela from Helsinki University of Technology.
- 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).
- Raya Khanin and Ernst Wit of the University of Glasgow and the University of Lancaster (*E. coli* repressor system).