Gaussian Process Toolkit for Modelling the Dynamics of Transcriptional Regulation

Pei Gao, Michalis K. Titsias, Neil D. Lawrence, Magnus Rattary
School of Computer Science, University of Manchester

The complex dynamics of cells and tissues are regulated in part by networks of interacting genes and proteins. The structure of the interaction network dictates which genes are regulated by which transcription factors (TFs). Particularly, a number of experimental and computational methods have been proposed to explore the mechanisms of transcriptional regulation. A key problem with the analysis of transcription network is that the concentration of the activated TF is difficult to measure directly whereas the target mRNA quantities are relatively easy to obtain with a microarray. Therefore, we focus on the problem of inferring the transcription factor activity given the mRNA expression level data.

Mathematical modelling provides a powerful tool with which to use empirical data on network topology and time-course expression profiles to gain insight into functional forms for network interactions. Differential equations provide an ideal framework for such integrative studies, since they specifically relate expression profiles to functional forms for interactions that depend on a small number of parameters that are amenable to experimental determination (such as binding affinities, degradation rates, etc.). Mostly inspired by Barenco et al. (2006), we aims to infer the protein concentrations through differential equation models of the system in combination with probabilistic process prior distributions. We consider the case of a single input module (SIM) network motif, the mRNA concentration of a given gene has been modelled by the following differential equation

$$\frac{dx_j(t)}{(t)} = B_j + S_j g\left(f(t)\right) - D_j x_j(t) \tag{1}$$

where $f(t)$ is the concentration of the TF, $g(.)$ is a non-linear function which can take different forms for the activation or repression response, B_j is the basal transcription rate for gene j, S_j is the sensitivity to the TF and D_j is the linear degradation rate of $x_j(t)$.

We advocate the use of Gaussian processes to define prior distributions over the latent chemical species, i.e. $f(t)$. Gaussian process modelling provides not only the inference of continuous quantities without discretization but also the natural capability of handling uncertainty. This also 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.

Results

We demonstrate our general approach on three different biological examples of single input motifs, including both activation and repression of transcription. 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), approximate (MAP-Laplace method for non-linear regulation) and the sampling-based inference scheme and therefore provide us with a practical toolkit for model-based inference. For example in 1 we show the results from the same microarray data as Khanin et al. (2006) for the transcription factor LexA in *E. Coli* using the MAP-Laplace and the sampling-based methods.

References

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. *Genome Biology*, 7(3):R25, 2006.
R. Khanin, V. Viciotti, and E. Wit. *Proc. Natl. Acad. Sci. USA*, 103(49):18592–18596, 2006.
X. Liu, M. Milo, N.D. Lawrence, and M. Rattray. *Bioinformatics*, 21:3637–3644, 2005.

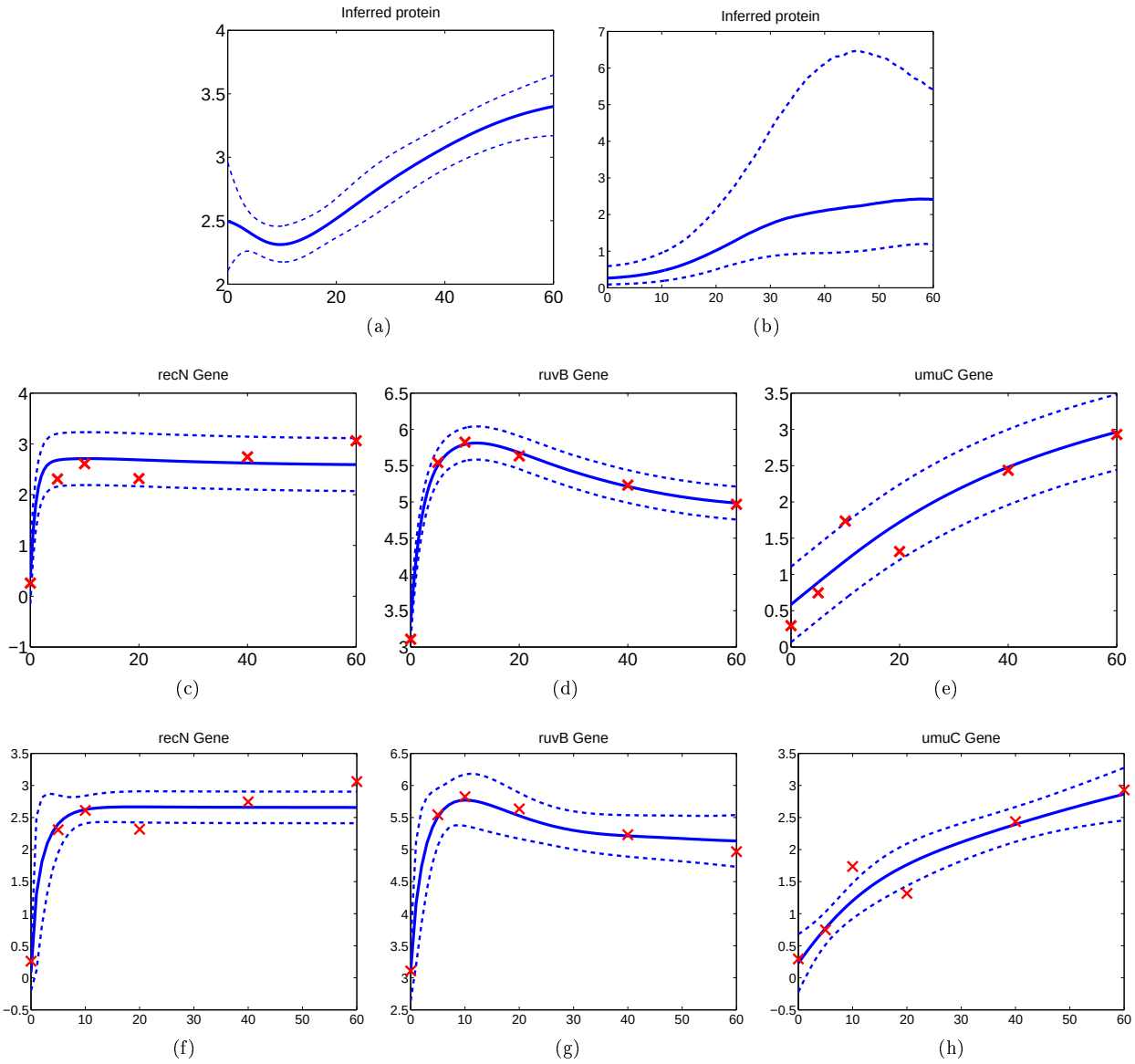

Figure 1: Example of the transcriptional factor activity inference, showing the results for the repressor LexA. (a) predicted LexA protein concentration using the MAP-Laplace method. (b) predicted LexA protein concentration using the sampling method. (c) predicted expression level for recN using the MAP-Laplace method. (d) predicted expression level for ruvB using the MAP-Laplace method. (e) predicted expression level for umuC using the MAP-Laplace method. (f) predicted expression level for recN using the sampling method. (g) predicted expression level for ruvB using the sampling method. (h) predicted expression level for umuC using the sampling method. 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.