2.5 Time-series modelling in bioinformatics

Bayesian methods are well-suited for analysis of molecular biology data as the data sets practically always consist of very few samples with a high noise level. We have studied models of gene transcription regulation based on time series gene expression data in collaboration with the Machine Learning and Optimisation group at the University of Manchester. This is a very challenging modelling task as the time series are very short, typically at most a dozen time points.

In [22], we have developed a method of modelling single input motif systems, where a single transcription factor regulates a number of genes. This is achieved by imposing a Gaussian process prior on the latent regulator (transcription factor protein) activity, which under a linear ODE transcription model leads to a joint Gaussian process model for all observable gene expression values. The model can further be extended by incorporating the transcription factor expression levels through a translation model. It is also possible to consider nonlinear models by using approximate inference. A sample model of p53 activation is illustrated in Fig. 2.8.

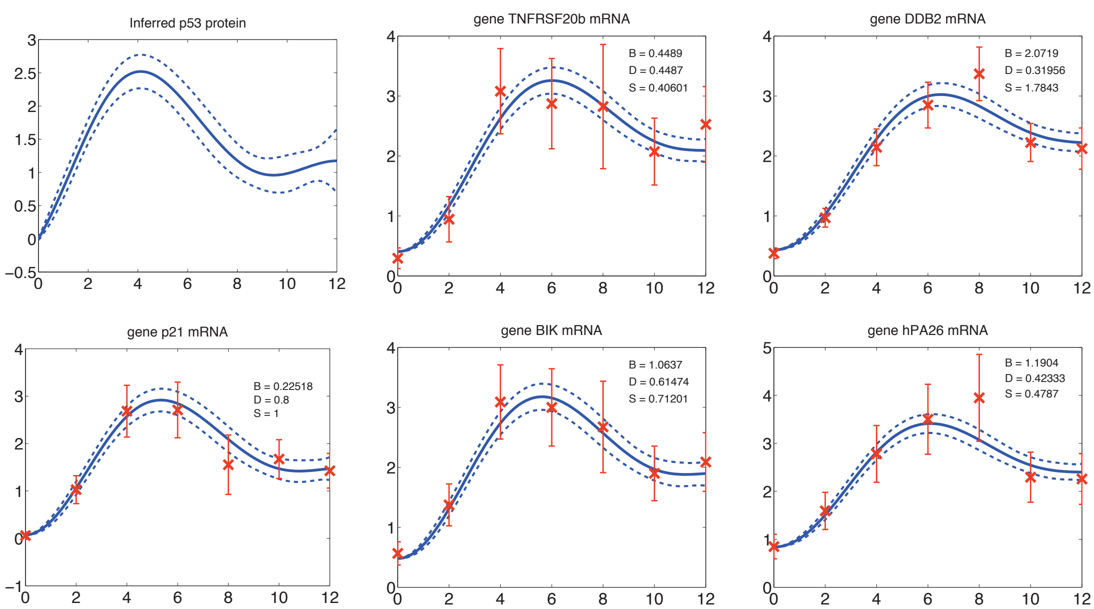

Figure 2.8: An inferred model of transcription factor p53 activation based on five known target genes. Red marks denote observed gene expression values while blue curves are inferred by the model along with 2 standard deviation error bars.

We have applied the model to genome-wide ranking of potential target genes of transcription factors. In experiments with key regulators of *Drosophila* mesoderm and muscle development, this has lead to extremely promising results in terms of enrichment of differential expression in loss-of-function mutants as well as ChIP-chip binding near the predicted target genes [23].

[17] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. Tech. report TKK-ICS-R6, Helsinki University of Technology, TKK reports in information and computer science, Espoo, Finland, 2008.

[18] L. Kozma, A. Ilin, and Tapani Raiko. Binary principal component analysis in the Netflix collaborative filtering task. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, Grenoble, France, September 2009.

[19] J. Luttinen, A. Ilin, and Juha Karhunen. Bayesian robust PCA for incomplete data. In *Proc. of the 8th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2009)*, pp. 66–73, Paraty, Brazil, March 2009.

[20] J. Zhao, Q. Jiang. Probabilistic PCA for t distributions. *Neurocomputing*, 69:2217–2226, 2006.

[21] C. Archambeau, N. Delannay, M. Verleysen. Robust probabilistic projections. In *Proc. of the 23rd International Conference on Machine Learning (ICML 2006)*, pp. 33-40, New York, NY, USA, 2006.

[22] P. Gao, A. Honkela, M. Rattray, and N. D. Lawrence. Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics* 24(16):i70–i75, 2008.

[23] A. Honkela et al. A model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci U S A*, 2010. doi:10.1073/pnas.0914285107

[24] M. Harva. Algorithms for approximate Bayesian inference with applications to astronomical data analysis. *TKK Dissertations in Information and Computer Science*, TKK-ICS-D3, Espoo, Finland, 2008. Available at `http://lib.tkk.fi/Diss/2008/isbn9789512293483/`.

[25] S. Refsdal. On the possibility of determining Hubble's parameter and the masses of galaxies from the gravitational lens effect. *Monthly Notices of the Royal Astronomical Society*, 128:307–310, 1964.

[26] J. C. Cuevas-Tello, P. Tino, and S. Raychaudhury. How accurate are the time delay estimates in gravitational lensing? *Astronomy & Astrophysics*, 454:695–706, 2006.

[27] M. Harva and S. Raychaudhury. Bayesian estimation of time delays between unevenly sampled signals. *Neurocomputing*, 72(1-3):32–38, 2008.

[28] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated Model-based Bias Field Correction of MR Images of the Brain. *IEEE Transactions on Medical Imaging*, 18(10):885–896, 1999

[29] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated Model-based Tissue Classification of MR Images of the Brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908, 1999

[30] T. Autti, M. Mannerkoski, J. Hämäläinen, K. Van Leemput, and L. Åberg. JNCL patients show marked brain volume alterations on longitudinal MRI in adolescence. *Journal of Neurology*, 255(8):1226–1230, 2008

[31] K. Van Leemput Encoding Probabilistic Brain Atlases Using Bayesian Inference. *IEEE Transactions on Medical Imaging*, 28(6):822–837, 2009