

# Gaussian Processes and Gene Regulation

## Multiple Transcription Factors, Networks and Model Organisms

15.6.08

**Background and Motivation.** Gaussian Processes (Rasmussen and Williams, 2005) can provide a convenient framework for the analysis of gene regulation as shown in two recent papers (Lawrence et al., 2007; Gao et al., 2008). In these two papers, the expression levels of a set of genes were known and they were governed by a common regulatory factor (TF). The unobserved concentration of TF was described by a Gaussian Process. This approach has several advantages and should be explored to the full. Relative to the papers (some of which are clearly being worked on by Lawrence and colleagues) above three useful generalisations immediately springs to mind:

- Multiple TFs. If the number ( $k$ ) is known, then their unobserved concentration would be described by a vector valued GP. If the number is not known, it would be natural to give this a prior. Even if this is satisfactorily solved, two questions immediately arise: Firstly, are the concentrations of the TFs correlated? The simplifying answer is no. Secondly, how do multiple TFs govern transcription of individual genes? There exist very complex models for this, including any Boolean function from  $\{0,1\}^k$  to  $\{0,1\}$  and highly parametrized kinetic models. The simplifying answer is that the effects of TFs are additive and independent.
- Network Models of GR in a single organism, will allow the TFs to be products of the genes whose expression levels are observed.
- Network Models and Model Organisms will in the simplest case occur, if experiments in for instance human and mouse have been performed on homologous sets of molecules. A model of the evolution of GR would be needed.

Exploring these extensions together with suitable data sets would be very interesting. The proposed extensions clearly are relevant for coming data. Several methodological questions should be addressed:

- What is the computational complexity of the proposed models? As models become more complex it will be increasingly challenging, to explore the parameter/solution space. All three extensions also introduces uncertainty of the structure of the model, which implies that this would have to be given or a search over possible models would have to be performed.
- For increasingly complex (and realistic models) more and more data will be needed and it could be very well be possible that it could be shown that is impossible to reach biological conclusions for more complex models.
- As formulated these models do not use biological knowledge, which is a serious restriction as we move towards increasingly complex models. Adding biological knowledge could seriously restrict parameter/solution space and extend how complex situations can be handled.

Several points were not addressed in the original papers, such as how to deal with error in measurement times.

Kalman Filters (Cui et al, 2005), Dynamical Bayesian Networks (Huang et al., 2007) and ODE inference are the main competitors to GP. Any data set should be probably be analyzed by all methods.

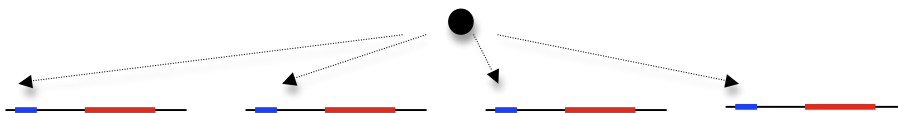
**The Basic Model.** (for details see Lawrence et al., 2007 and Gao et al., 2008)

The concentration of gene product  $j$  is  $x_j(t)$  at time  $t$ . There is a basic production  $B_j$ , a linear response  $S_j$  to the concentration of the hidden TF having concentration  $f(t)$  and there is a decay,  $D_j$ , proportional to the concentration of the of the gene  $j$ . The initial condition of the resulting ODE is taken to be the equilibrium level without the presence of TF, the result sets of ODEs will be:

$$\frac{dx_j}{dt} = B_j + S_j f(t) - D_j x_j(t), \quad x_j(0) = \frac{B_j}{D_j}$$

The concentration of the TF -  $f(t)$  - will be described by a Gaussian Process. The resulting model is simple, nice, tractable and avoids reliance on computationally intensive procedures. Due to the linearity of the model, the concentrations of the gene products will also follow a Gaussian process, and the distribution of  $\{x_j(t), f(t)\}$  is fully characterized by knowing the means and covariances for pairs of concentrations and time points. Posterior distribution of  $f(t)$  given the  $x_j(t)$  are easily obtained.

The first natural extension is to relax the assumption that regulation depends linearly on  $f(t)$ , but use a more realistic model like Michaelis-Menten. A second extension would be to let the process be observed with error, ie  $y_j(t) = x_j(t) + \varepsilon_j(t)$ , where  $\varepsilon_j(t)$  are independently  $N(0, \sigma^2)$  distributed. Both extensions were investigated by Lawrence and colleagues. We will only add any level of detail to the multiple TF extension, but continue to elaborate on this as the work proceeds.



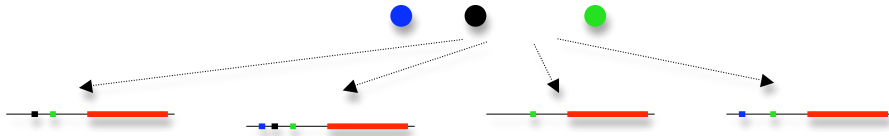
One TF (transcription factor - black ball) whose concentration fluctuates over times influences  $k$  genes (four in this illustration) through their TFBS (transcription factor binding site - blue). The strength of its influence is described through a gene specific sensitivity,  $S_j$ .

**Project 1: Multiple TFs.**

If the transcription levels of  $n$  genes are governed by  $k$  TFs, assuming additivity and independence of the effect of transcription factors, the basic equation will become

$$\frac{dx_j}{dt} = B_j + \sum_i S_{ij} f_i(t) - D_j x_j(t), \quad x_j(0) = \frac{B_j}{D_j}$$

Even for a small number of TFs, this will lead to large growth in parameters. In the simple linear model, it will be necessary to address: Firstly, that the number of transcription factors is unknown and secondly, that a given TF only interact with a subset of the genes (ie  $S_{ij}$  is often zero). The natural approach here is Bayesian where  $k$  is given a prior and given  $k$  the interactions ( $S_{ij}$  is not zero) is also given a prior with a bias towards a few interactions. The model can then be investigated using MCMC.



Several TFs (balls) whose concentrations fluctuates over times influences  $k$  (four in this illustration) through their TFBS (matching colours). The strength of their influences is described through a TF-gene specific sensitivities,  $S_{jk}$ .

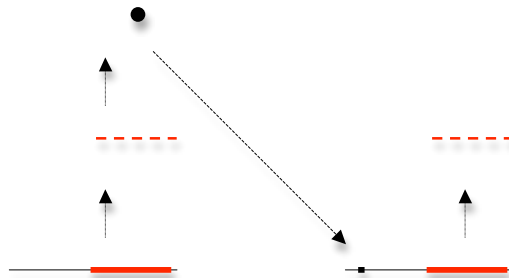
Simulated Data: Assume 10 genes, 5 TFs, at most 3 TFs can act on a gene simultaneously. Implement a simulator that can generate data from this model. Investigate the ability to recover the underlying model as a function of i) the underlying parameters, ii) number and time density of observations.

**Project 2: Network Models.**

Now a series of the gene products will also function as TFs, which will lead to a modification of the basic equation. In vector/matrix form

$$\mathbf{x}' = \mathbf{b} + \mathbf{S}^f \mathbf{f} + \mathbf{S}^x \mathbf{x} - \mathbf{D}\mathbf{x}$$

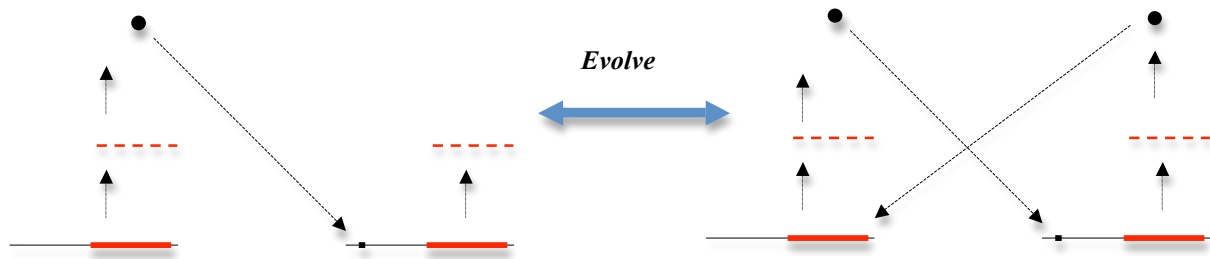
Where the time  $t$  has been suppressed.  $\mathbf{S}^f$  is a square matrix of dimension number of TFs and  $\mathbf{S}^x$  is a square matrix of dimension number of genes.  $\mathbf{b}$  is a vector and  $\mathbf{D}$  is a diagonal matrix. The  $\mathbf{S}^x \mathbf{x}$  term stems from the genes that code for TFs. This term also assumes that there is a linear and deterministic relationship between mRNA level and the produced TF.



Now we have two genes that illustrate the simplest conceivable network that would have two nodes and one arrow. The left gene produces RNA (red dashed lines) that produces a TF (black ball) that regulates the second gene. Now we have direct knowledge of the concentration of the TF since we can measure the RNA that produces it and is not a hidden process any more.

**Project 3: Model Organisms and Network Models.**

The set of genes and TFs might both change and so might  $\mathbf{B}$ ,  $\mathbf{D}$ ,  $\mathbf{S}^f$  and  $\mathbf{S}^x$ . There are natural models for each of these, both the set of genes and the set of TFs can evolve by adding-deleting elements. For continuous quantities, it is natural to use an Ornstein-Uhlenbeck process and when a continuous quantity is born it is sampled from the equilibrium distribution thereof.



Now we have observed two systems (in two species), each with two genes. The systems will be related, so observing one system will give information on the other system and visa versa. However, the two systems might not be identical. Here illustrated by that on the right side, the second gene also relates the first.

**Comments.** The projects sketched here, clearly are part of a much larger complex, where the next steps would be using phylogenetic footprinting to make statements about which TFs interacts with which genes. A step further would be to use mapping (combined genotype-phenotype observations) to select which genes to include in the set to be modelled.

Joanna Davies, Christo Fogelberg, Aziz Mithani, Marton Munz and Jotun Hein plan to work on Project 1 May-July 2008.

### **References.**

- Barenco, Tomescu, Brewer, Callard, Stark and Hubank (2006) "Ranked prediction of p53 targets using hidden variable dynamic modelling" *Genome Biology* 7.3.R53
- Cui, Liu, Jian and Ma (2005) "Characterizing the dynamic connectivity between genes by variable parameter regression and Kalman filtering based on temporal gene expression" *Bioinformatics* 21.8.1538-41.
- Gao, Hinkela, Rattray and Lawrence (2008) "Gaussian Process Modelling of Latent Chemical Species: Applications to Inferring Transcription Factor Activities"
- Huang, Wang, Zhang, Sanchez and Wang (2007) "Bayesian Inference of Genetic Regulatory Networks from Time Series Microarray Data Using Dynamic Bayesian Networks" *J. Multimedia* 2.3.46-56
- Lawrence, Sanguinetti, and Rattray (2007) "Modelling transcriptional regulation using Gaussian processes"
- Quayle, A. and Bullock, S. (2006) "Modelling the evolution of genetic regulatory networks." *Journal of Theoretical Biology*, 238 (4), pp. 737-753
- Torsten Reil (1999) "Dynamics of Gene Expression in an Artificial Genome - Implications for Biological and Artificial Ontogeny" *ECAL* 457-466
- Rasmussen and Williams (2005) "Gaussian Processes for Machine Learning" MIT