Switching Regulatory Models

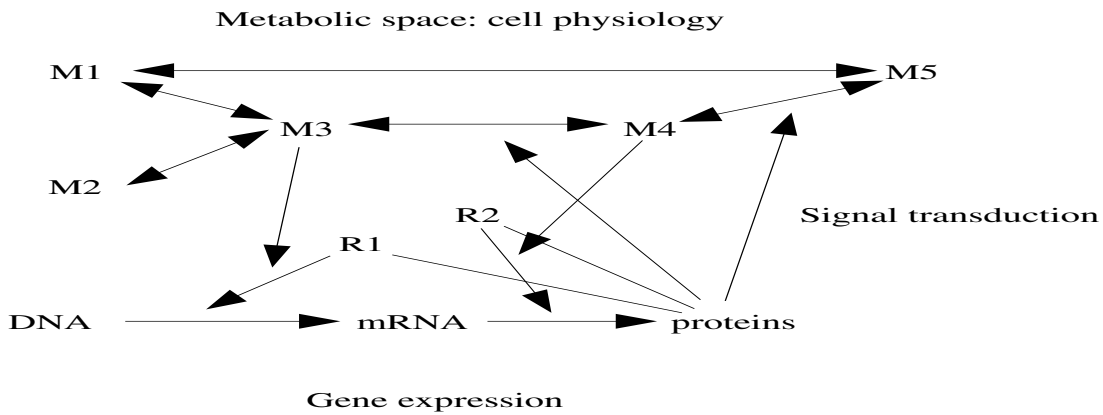
Guido Sanguinetti, Andreas Ruttor, Manfred Opper and Cedric Archambeau

Department of Computer Science - The University of Sheffield

Whistler, Dec 2008

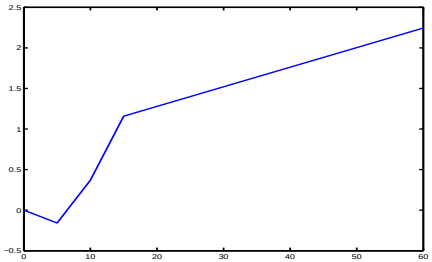
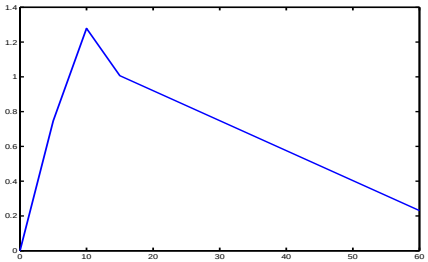
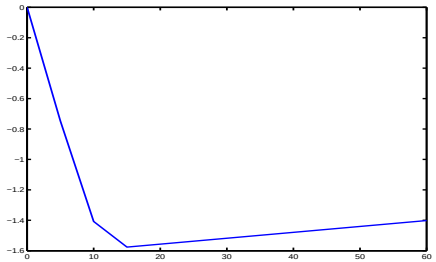
Outline of the talk

Basic problem



- We wish to predict dynamics of signalling components (transcription factors, TFs) during adaptation to stress, based on time-course mRNA profiles
- Example: *E. coli* transition between aerobic and anaerobic states
- Not just interested in steady states

Basic problem

- Various possible dynamics: *left* poxB (pyruvate metabolism), *centre* ygbD, *right* hybC (anaerobic respiration). Data from Partridge *et al.* 2006.
- Presence of many poorly understood transient expression profiles suggest tightly controlled timings of regulatory signals.

Is this a problem?

- TFs are often post-transcriptionally and post-translationally regulated: mRNA poor proxy
- Stoichiometry is a problem: small changes in TFs are greatly amplified at mRNA level. Therefore, many TFs are low expressed
- TFs activation state changes quickly: need high sampling rates
- *Main idea*: treat TF activity profile as a latent function of time and use inference (Liao *et al.* 2003, Barenco *et al.* 2006, *etc*)

Single Input Motif

- In general, most genes have complex promoter structures with several TFs interacting
- The single input motif (SIM) is a specific network motif where several genes are controlled by a single TF
- The TF input to the SIM are generally *Master regulators*, TFs who control hundreds of genes and generally are associated with large shifts in cellular behaviour

Basic problem

- Consider an ODE model of SIM dynamics

$$\frac{dx_i(t)}{dt} = A_i \frac{f(t)}{\kappa_i + f(t)} + b_i - \lambda_i x_i(t)$$

- Given time course observations of the expression levels of the target genes x_i, infer the profile of the transcription factor f and the model parameters A_i, b_i and λ_i
- Problem originally considered by Barenco *et al.* (linear dependence on the TF), and then Lawrence *et al.*, Khanin *et al.*, Rogers *et al.*,...

A prior problem

- When approaching the inference problem from a Bayesian point of view, one needs to specify priors for the function $f(t)$
- Most approaches discretise time and take piecewise constant f, assuming independence between the values of f in different intervals
- Lawrence *et al.* performed inference in continuous time, placing a Gaussian Process (GP) prior over the function f

Time is of the essence

- Retaining continuous time seems a more realistic approach
- GPs however make strong continuity assumptions, preventing sudden changes in TF activity
- "Transcription factors are usually designed to transit rapidly between active and inactive molecular states" (U. Alon, Introduction to Systems Biology, p 7)
- In practice, cells need to respond to some stimuli extremely quickly, so they keep a reservoir of inactive TF proteins that can be activated quickly and simultaneously through dimerisation, phosphorylation, etc.

Swicthing latent process

- We assume TF activity jumps quickly from zero to saturation level

$$\frac{dx_i(t)}{dt} = A_i \mu(t) + b_i - \lambda_i x_i(t) \tag{1}$$

  where $\mu(t) \in \{0, 1\}$
- The driving process $\mu(t)$ is modelled as a two-states Markov jump process, also known as a *telegraph process*
- Given transition rates $f_{0,1}(t)$ for the process, the probability $p_1(t)$ of $\mu(t) = 1$ at a given time is given by the following Master equation

$$\frac{dp_1(t)}{dt} = -(f_1 + f_0)p_1(t) + f_1(t). \tag{2}$$

Exact inference

- We consider the *joint* process $(\mu(t), x(t))$
- The joint posterior process can be shown to be still Markovian. Its time marginals factorises

$$q_\mu(x,t) = \frac{1}{Z} p_\mu(x,t)\Psi_\mu(x,t)\,. \tag{3}$$

  This is the product of forward and backward passing messages.
- The forward-backward recursion relations become in continuous time the Chapman-Kolmogorov equation. The backward equation is

$$\frac{\partial \Psi_\mu}{\partial t} + \sum_{i=1}^{m} (A_i \mu + b_i - \lambda_i x_i) \frac{\partial \Psi_\mu}{\partial x_i} = f_{1-\mu}(\Psi_\mu(x,t) - \Psi_{1-\mu}(x,t))\,. \tag{4}$$

- This is solved backwards in time with the observations providing jump conditions in (4)

Exact inference

- The posterior marginal satisfies a forward Chapman-Kolmogorov equation

$$\frac{\partial q_\mu}{\partial t} + \sum_{i=1}^{m} \frac{\partial}{\partial x_i}(A_i\mu + b_i - \lambda_i x_i) q_\mu(x,t) = \tag{5}$$
$$g_\mu(x,t)\, q_{1-\mu}(x,t) - g_{1-\mu}(x,t)\, q_\mu(x,t)\,,$$

  where

$$g_\mu(x,t) = \frac{\Psi_\mu(x,t)}{\Psi_{1-\mu}(x,t)}\, f_\mu \tag{6}$$

- In the case of a single target gene numerical integration of the PDEs (4) and (5) is computationally feasible

Variational approach

- We will approximate the posterior distribution over the latent switching process
- In principle, the posterior process can be obtained via Bayes' theorem

$$p_{post}\left(\mu_{0:T}|\hat{x}\right) = \frac{1}{Z} p\left(\hat{x}|\mu_{0:T}\right) p_{prior}\left(\mu|f_{0,1}\right). \tag{7}$$

- The general solution of equation (1) is obtained using Laplace transforms as

$$x\left(t\right) = \exp\left(-\lambda t\right)\left\{x\left(0\right) + \int_0^t \exp\left(\lambda s\right)\left[A\mu\left(s\right) + b\right] ds\right\}. \tag{8}$$

As this depends on the whole history of the process μ, the posterior process will not be Markovian

Variational approach

- We will approximate the posterior with a Markov process
- We compute the *Kullback-Leibler (KL) divergence* between the posterior process in (7) and an approximating Markov process $q(\mu|g_{\pm})$

$$KL[q\|p_{post}] = \ln Z + KL[q\|p_{prior}] - \sum_{j=1}^{N} E_q[\ln p(\hat{x}_j|x(t_j))].$$

- The KL divergence between two telegraph processes is given by

$$KL[q\|p_{prior}] = \int_0^T dt q_1(t)\left[g_-(t)\ln\frac{g_-(t)}{f_0(t)} + f_0(t) - g_-(t)\right] + \int_0^T dt\,[1-q_1(t)]\left[g_+(t)\ln\frac{g_+(t)}{f_1(t)} + f_1(t) - g_+(t)\right].$$

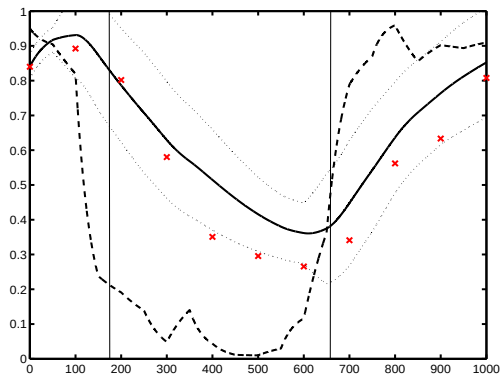
Variational approach

- Under the assumption of Gaussian noise, the estimation of the likelihood requires the estimation of the first two moments of the random variable $x(t)$ under the approximating process q.
- To do this one needs to compute integrals of the form

$$
\begin{aligned}
I_1^i &= \int_0^{t_i} \exp(\lambda s)\, q_1(s)\, ds \\
I_2^i &= \int_0^{t_i} \int_0^{t_i} \exp[\lambda(t+s)]\, q_1(t,s)\, dtds.
\end{aligned}
\tag{9}
$$

- Assuming that the rates of the approximating process are piecewise constant, these integrals can be computed analytically

Toy example



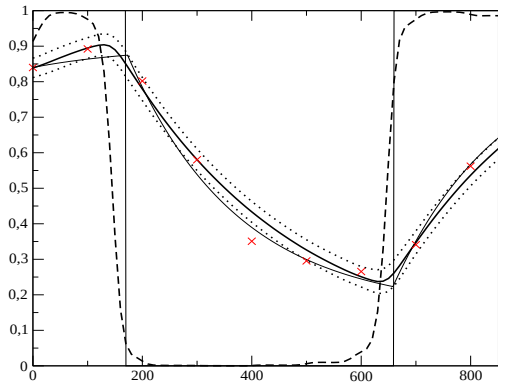


Figure: Left: variational inference $A = 2.3 \pm 0.5 \times 10^{-3}$, $b = 1.0 \pm 0.2 \times 10^{-3}$, $\lambda = 4 \pm 0.3 \times 10^{-3}$. Right: exact inference $A = 3.2 \pm 1.1 \times 10^{-3}$, $b = 0.08 \pm 0.6 \times 10^{-3}$, $\lambda = 3.1 \pm 1.3 \times 10^{-3}$. True values $A = 3.7 \times 10^{-3}$, $b = 0.8 \times 10^{-3}$, $\lambda = 5 \times 10^{-3}$.

FNR regulation

- As a real example on which to test our approach, we considered transcriptomic measurements of the reaction of *E.coli* to sudden oxygen starvation
- When oxygen is removed, Fe-S clusters are generated which dimerise and activate the master regulator FNR
- FNR activation is thought to be the main channel used by the bacterium to switch between aerobic and nitrate metabolism

FNR regulation

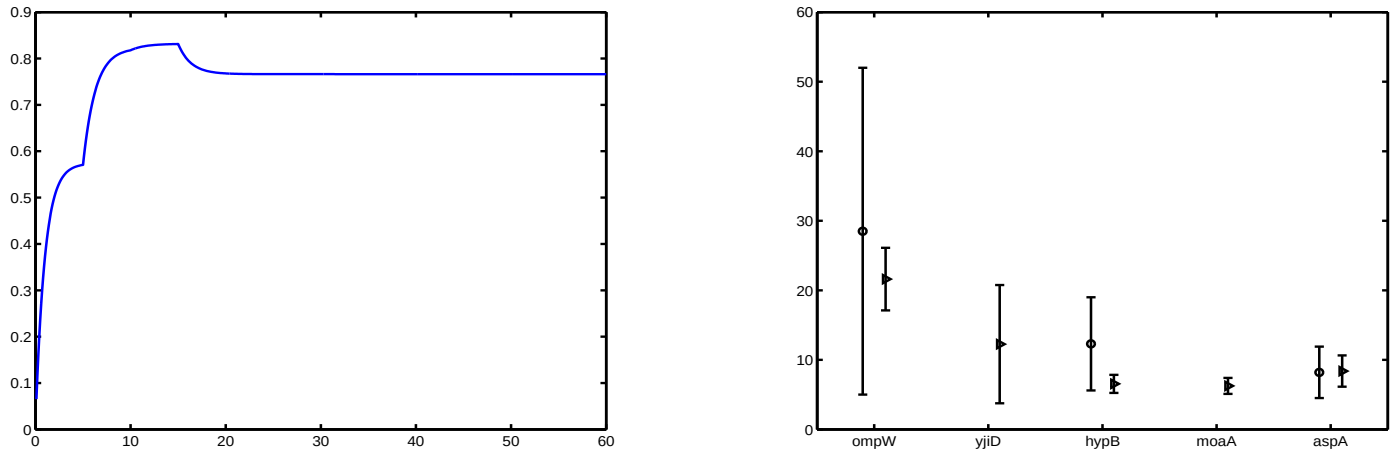


Figure: Results on *E.coli* data: (a) posterior mean FNR profile; (b) half lives of targets (in minutes) with uncertainty, inferred (triangles on the right) *versus* experimentally measured. No measurement of the half life of yjiD or moaA is available.

Conclusions

- We have proposed a novel TF inference framework which arguably could describe better some biological conditions
- It is of interest in its own right as an example of hybrid discrete-continuous models
- Potentially many other application domains

Future directions

- Determining optimal Markov approximations (removing piecewise constant rates)
- Generalise to bistable systems governed by SDEs
- Move beyond SIMs