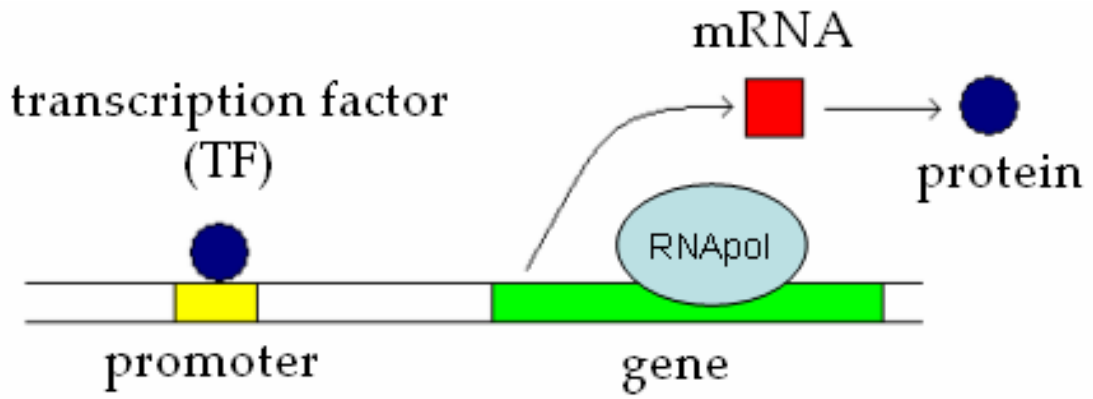
Inference in hierarchical transcriptional network motifs

Andrea Ocone and Guido Sanguinetti

Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh

MLSB workshop, Oct 2010

Outline of the talk



- TFs can be present in active/inactive state
- measure of active TFs is very hard
- gene expression levels (mRNA) easy to measure

Basic problem

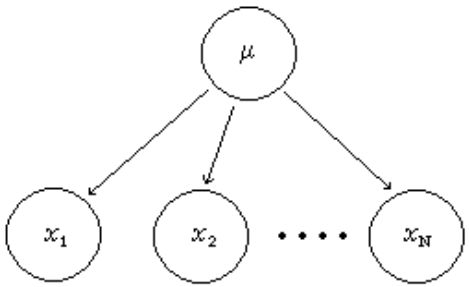
- Consider an ODE model of transcription dynamics

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

- 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 θ_i, b_i and λ_i
- Problem originally considered by Barenco *et al.*, and then Lawrence *et al.*, Khanin *et al.*, Rogers *et al.*,...

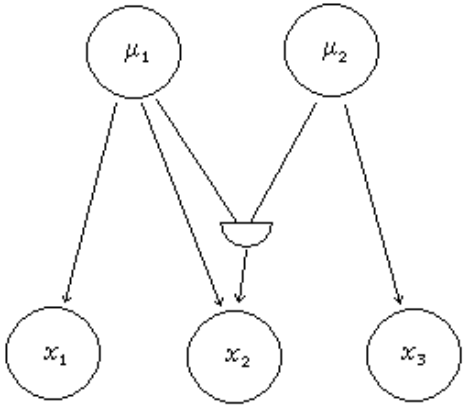
Network motifs

Single-Input Motif
SIM



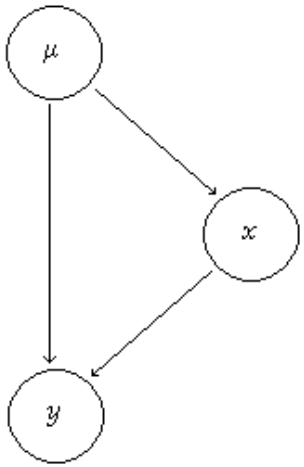
Barenco *et al.*, 2006
Rogers *et al.*, 2006
Lawrence *et al.*, 2006
Sanguinetti *et al.*, 2009

Dense Overlapping Regulons
DOR



Opper and Sanguinetti, 2010

Feed-Forward Loop
FFL



Nobody so far...

Feed-forward loop (FFL) network motif

- μ master transcription factor
- x slave transcription factor
- y target gene

- FFL can act as a biological filter

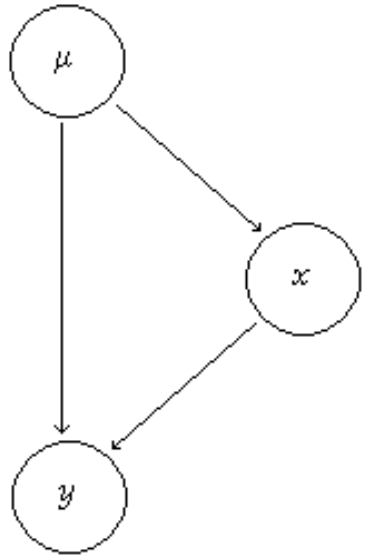
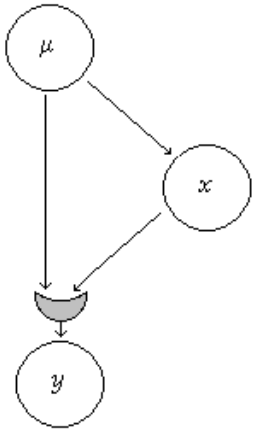


Figure: FFL network motif

FFL network motifs

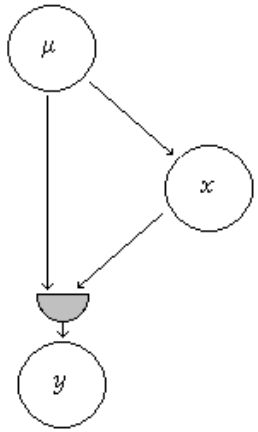
OR-gate FFL

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = A_1\mu(t) + b_1 - \lambda_1 x(t)$$
$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = A\mu(t) + b - \lambda y(t) + A_2\Theta[x(t) - c]$$



AND-gate FFL

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = A_1\mu(t) + b_1 - \lambda_1 x(t)$$
$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = A\mu(t)\Theta[x(t) - c] + b - \lambda y(t)$$



$\Theta[x(t) - c]$ represents the Heaviside step function (it is 1 if $x(t) > c$, zero otherwise)

What we are interested in

Given microarray observations $\hat{x}$ and $\hat{y}$ at discrete time points, the problems are

- state inference: when does TF is active and inactive?
- parameters estimation $(A, b, \lambda, A_1, b_1, \lambda_1, A_2, c)$
- model selection: AND gate FFL, OR gate FFL, mixture?

Bayesian inference framework

- Prior distribution
  - 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

$$\frac{dp_1(t)}{dt} = -(f_1 + f_0)p_1(t) + f_1(t) \tag{1}$$

- Likelihood
  - observations corrupted by Gaussian noise

$$p(\hat{x}_i|x_i) = \mathcal{N}(\hat{x}_i|x_i, \sigma_{x_i}) \qquad p(\hat{y}_i|y_i) = \mathcal{N}(\hat{y}_i|y_i, \sigma_{y_i})$$

Variational approach

- In principle, the posterior process can be obtained via Bayes' theorem

$$p_{post}(\mu_{0:T}|\hat{x},\hat{y}) = \frac{1}{Z}\, p(\hat{x}|\mu_{0:T})\, p(\hat{y}|\mu_{0:T},\hat{x})\, p_{prior}(\mu_{0:T}|f_{0,1})$$

- We will approximate the posterior with a Markov process
- We compute the *Kullback-Leibler (KL) divergence* between the posterior process and an approximating telegraph (Markov) process $q(\mu|g_{0,1})$

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

Variational approximation

- First and second moment for x and y can be calculated by solving iteratively a system of ODEs
- We still need to calculate some non trivial expectations under the approximating distribution q:
  - $\langle \Theta[x(s) - c] \rangle$
  - $\langle \Theta[x(s) - c] \mu(t) \rangle$
  - $\langle \Theta[x(s) - c] \Theta[x(t) - c] \rangle$

Assumptions

- $\langle \Theta[x(s) - c] \rangle = P(x(t) > c) = \int_c^\infty p(x(t)) \mathrm{d}x(t)$
  where $p(x(t)) \sim \mathcal{N}(x | \langle x(t) \rangle, \langle x(t)^2 \rangle - \langle x(t) \rangle^2)$

- $\langle \Theta[x(s) - c] \mu(t) \rangle \sim \langle \Theta[x(s) - c] \rangle \langle \mu(t) \rangle$

- $\langle \Theta[x(s) - c] \Theta[x(t) - c] \rangle$ decreases exponentially with the distance $t - s$, i.e.
  $\langle \Theta[x(s) - c] \rangle + (\langle \Theta[x(s) - c] \rangle \langle \Theta[x(t) - c] \rangle - \langle \Theta[x(s) - c] \rangle) \cdot (1 - e^{-\lambda_1 (t-s)})$

Optimisation

- ODEs for moments and master equation are included in the $KL[q||p_{post}]$ by using Lagrange multipliers λ_i
- Approximating process q found by gradient descent, solving forward and backward an iterative system

Algorithm

while $\Delta KL[q||p_{post}] > threshold$

solve forward: master equation, ODEs for moments

solve backward: $\left(\frac{\delta\mathcal{L}}{\delta q}, \frac{\delta\mathcal{L}}{\delta\langle x\rangle}, \frac{\delta\mathcal{L}}{\delta\langle x^2\rangle}, \frac{\delta\mathcal{L}}{\delta\langle y\rangle}, \frac{\delta\mathcal{L}}{\delta\langle y^2\rangle}\right) = 0 \longrightarrow \lambda_i(t)$

calculate gradients w.r.t. transition rates: $\left(\frac{\delta\mathcal{L}}{\delta g_+}, \frac{\delta\mathcal{L}}{\delta g_-}\right)$

calculate gradients w.r.t. parameters: $\left(\frac{\delta\mathcal{L}}{\delta A}, \frac{\delta\mathcal{L}}{\delta b}, \cdots\right)$

update transition rates $g_{0,1}$ and parameters

end

Results on simulated data set: state inference

Observations are given by adding Gaussian noise with SD of 0.03 to 10 discrete time points drawn from the model with a given TF activity (input μ) and known parameters. The inferred posterior TF activity is then compared with the true input

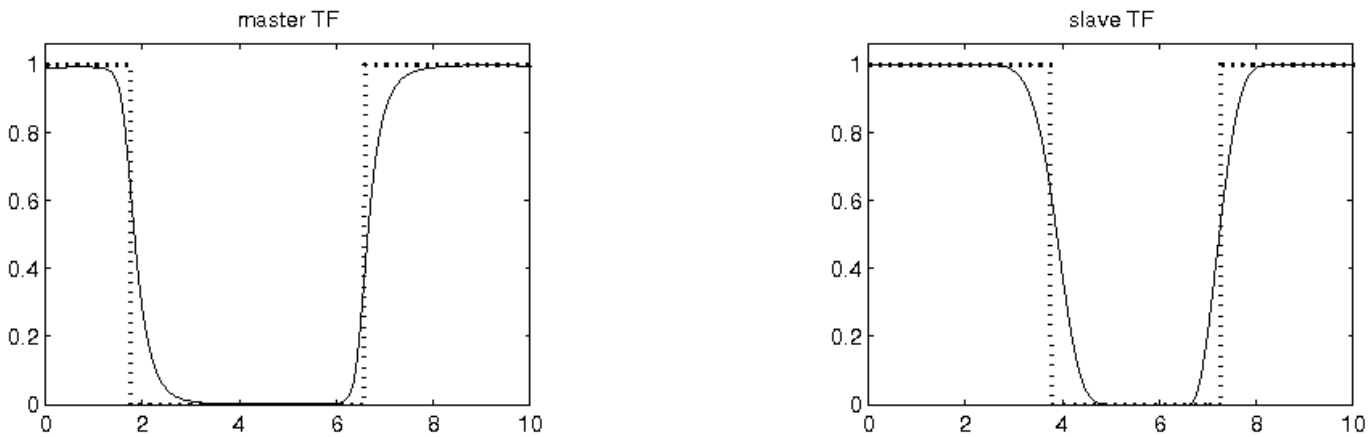


Figure: Inferred posterior mean activity (solid) versus true input impulse (dashed)

Results on simulated data set: parameters estimation

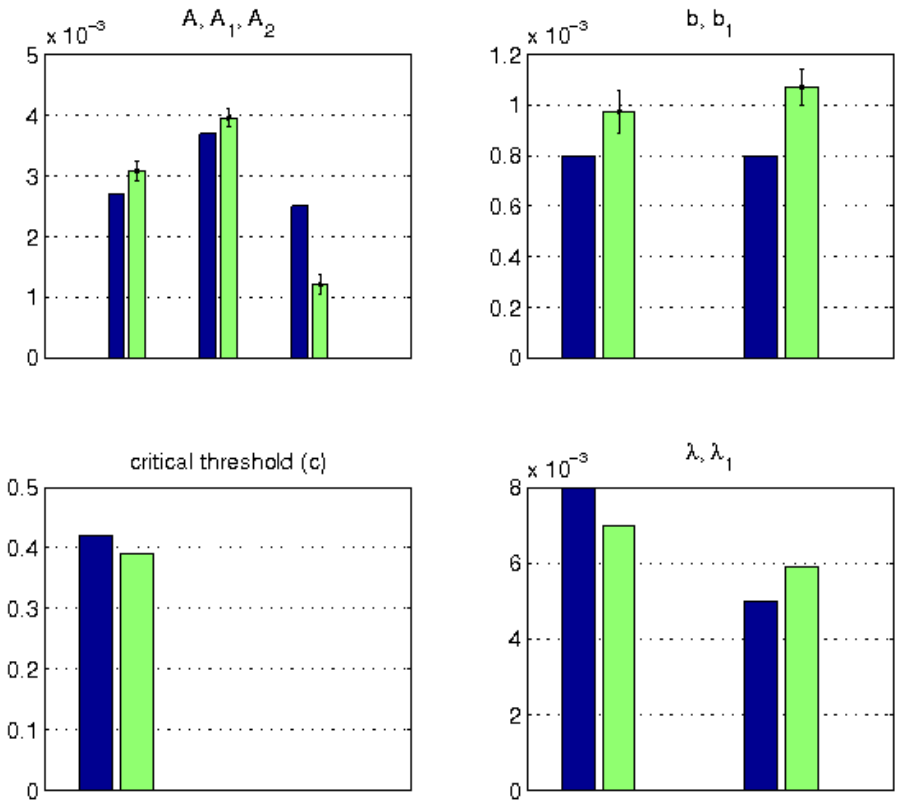


Figure: Inferred posterior parameters (green) versus true parameters (blue)

Results on simulated data set: moments reconstrunction

From inferred transcription factors activities and estimated model parameters we reconstruct first moment for x and y and compare with real observations $\hat{x}$ and $\hat{y}$, respectively

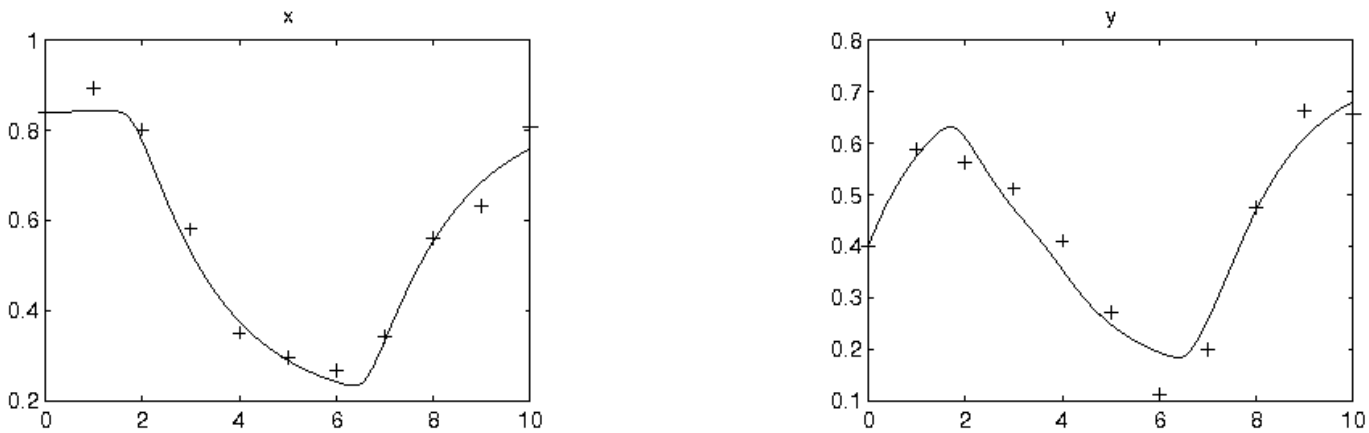
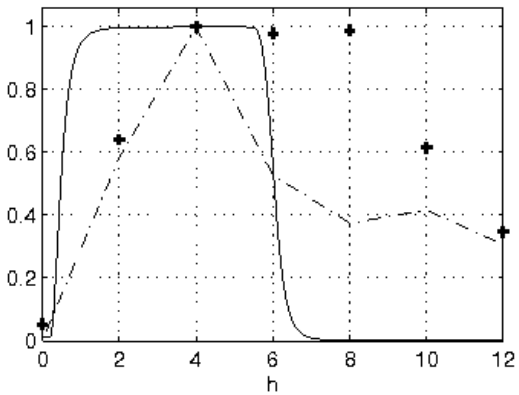
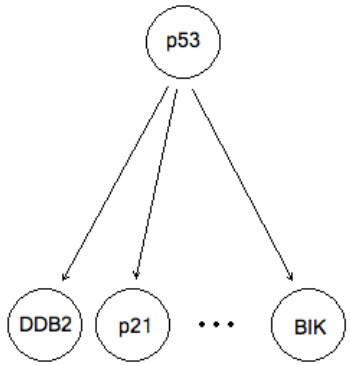


Figure: Inferred posterior first moments (solid) versus observations (crosses)

Results on p53 data set: SIM model

- Activity of p53 has been experimentally measured by Barenco et al. (Genome Biology, 2006) using western blots (semi-quantitative)
- Barenco (and later Lawrence et al., NIPS 2007) predicts p53 activity using a single-input motif (SIM) structure
- We compute inference on p53 activity using a SIM model and compare our results with Barenco's prediction

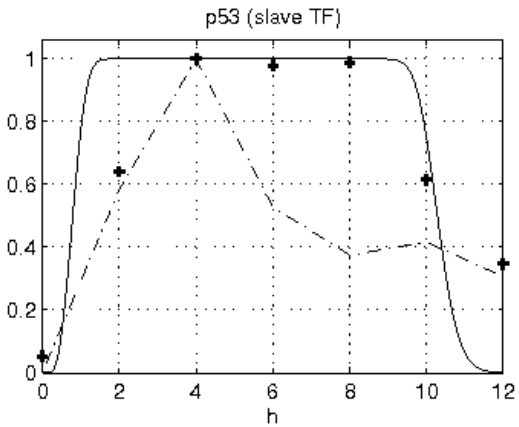


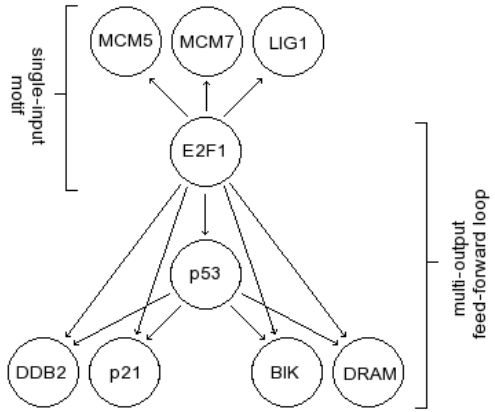


- p53 activity experimental measure (crosses)
- Barenco SIM prediction (dashed) compared with our SIM prediction (solid)

Results on p53 data set: FFL model

- p53 is involved in a FFL where it acts as a slave TF (Nature Reviews, 2009)
- E2F1 represents the master TF which activates p53 and p53 target genes
- We compute inference on p53 activity using a FFL model and compare our results with Barenco's prediction





- p53 activity experimental measure (crosses)
- Barenco SIM prediction (dashed) compared with our FFL prediction (solid)

Conclusion and future directions

- FFL models can explain biological data and give better predictions on TFAs, compared to SIM models
- multi-input FFL, multi-slave FFL, feedback loops
- stochastic versions (see Opper, Ruttor and Sanguinetti NIPS10)

References

- G. Sanguinetti, A. Ruttor, M. Opper and C. Archambeau, Switching Regulatory Models of Cellular Stress Response, Bioinformatics, 25(10):1280-1286 (2009)
- M. Opper and G. Sanguinetti, Learning combinatorial transcriptional dynamics from gene expression data, Bioinformatics, 26(13) 1623-1629 (2010)
- M. Opper, A. Ruttor and G Sanguinetti, Approximate inference in continuous time Gaussian-Jump processes, NIPS 2010

Acknowledgements

Jeff Green, University of Sheffield

Manfred Opper, TU Berlin