Time-dependent Gene Network Modelling by Sequential Monte Carlo

Ercan E. Kuruoglu[1,2], **Sergiy Ancherbak**[1], **Martin Vingron**[2]

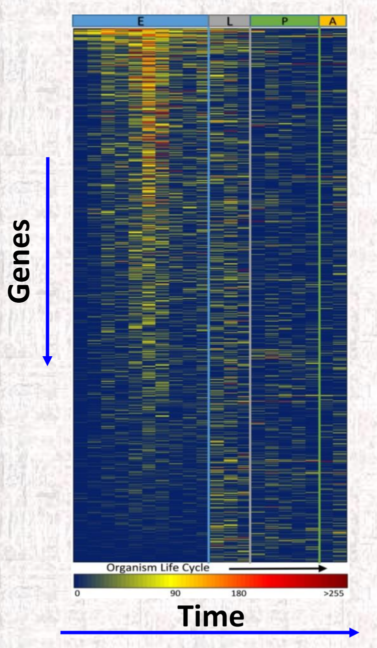
1 - CNR- ISTI, Pisa, Italy
2 - Max Planck Institute for Molecular Genetics, Berlin, Germany

Most existing methods used for gene regulatory network modeling are dedicated to inference of steady state networks, which are prevalent over all time instants. However, gene interactions evolve over time. Information about the gene interactions in different stages of the life cycle of a cell or an organism is of high importance for biology. In the statistical graphical models literature one can find a number of methods for network modelling while the study of time varying networks is rather recent. Using synthetic time series dataset for a gene network, we show that a sequential Monte Carlo method, namely Particle Filtering method is capable of tracking gene expression data and infer time-varying networks online.

Keywords: Bayesian network, gene expression data, gene network, particle filter, sequential Monte Carlo, time series

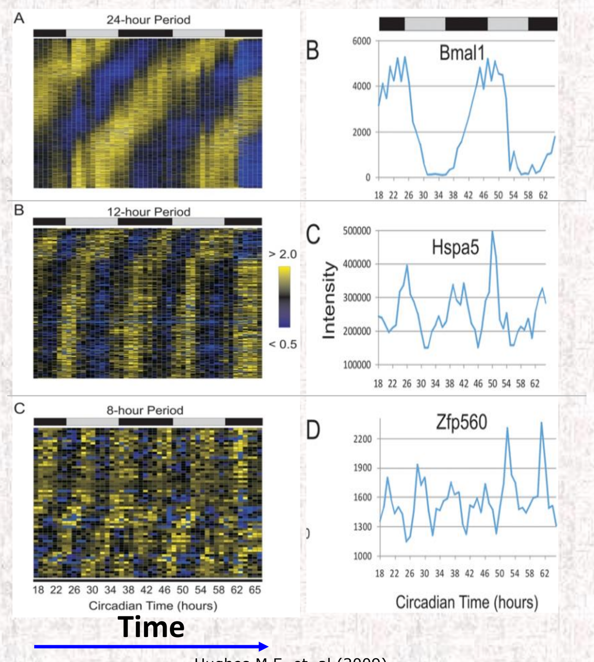
Gene expression data over time

Drosophila Melanogaster



Arbeitman M.N. et. al (2002)

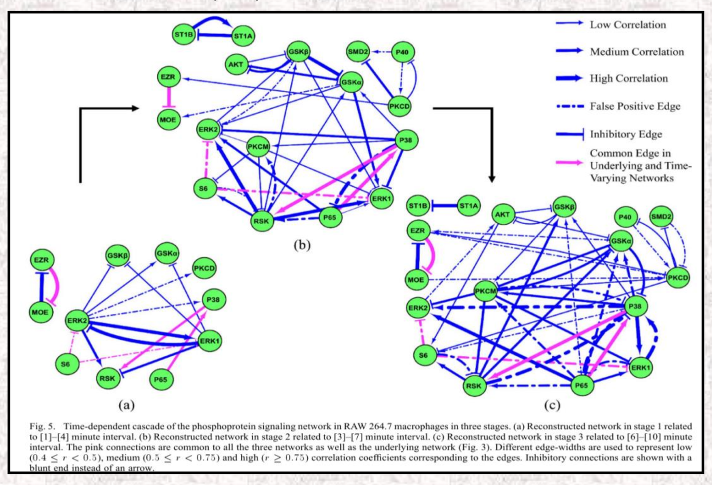
Mouse liver (cycling genes)



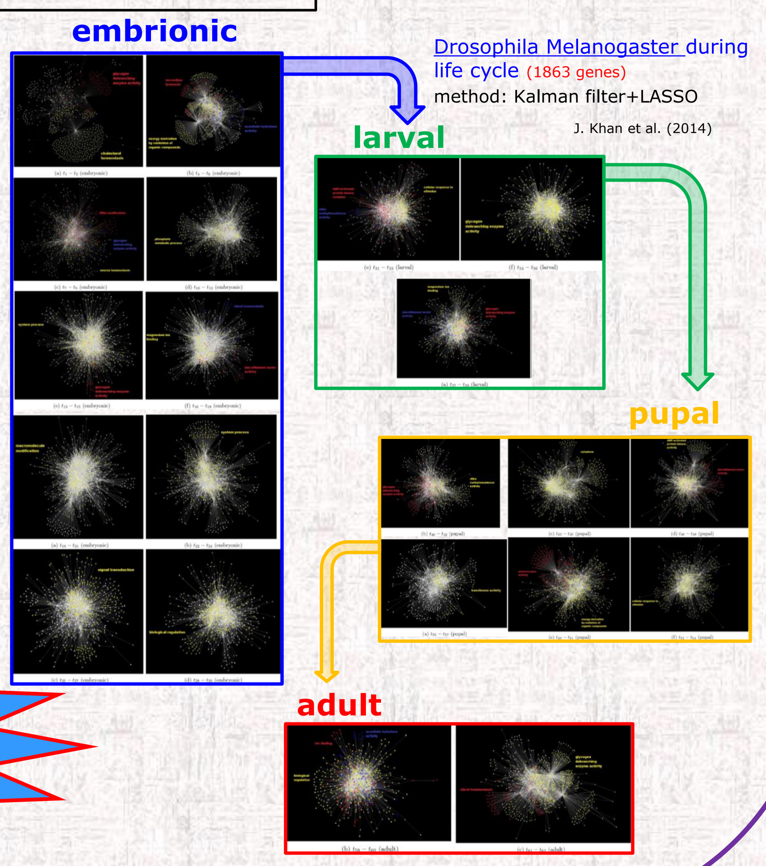
Hughes M.E. et. al (2009)

Gene networks over time

Phosphoprotein signalling network
method: least-squares regression + statistical hypothesis testing (t-test)
M. Masnadi-Shirazi (2014)



Drosophila Melanogaster during life cycle (1863 genes)
method: Kalman filter+LASSO
J. Khan et al. (2014)



Gene interaction network changes over time!

Model

We have time series of expression data for *N genes* and want to estimate *N-by-N coefficients* a_{ij} which model the regulatory relations among various genes:

$$x(t) = [x_1(t) \;\; x_2(t) \dots x_N(t)]^T$$

$$a_{ij}(t) = [a_{11}, \dots, a_{1N}; a_{21}, \dots, a_{2N}; \dots a_{N1} \dots a_{NN}]^T$$

Multivariate linear regression model relating the expression value of each gene at a given time to the gene expression values of the previous time instant:

gene expression values - $$x_{i,t} = a_{i1,t} \cdot x_{1,t-1} + a_{i2,t} \cdot x_{2,t-1} + \dots + a_{iI,t} \cdot x_{I,t-1} + \eta_t, \quad (1)$$

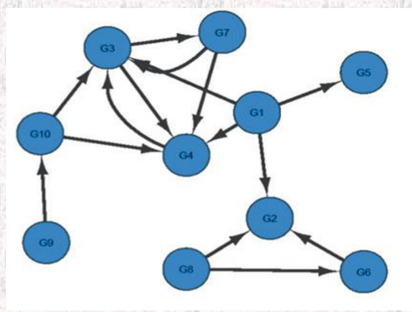
$i = 1, \dots, I; \quad I - total\ number\ of\ genes,$

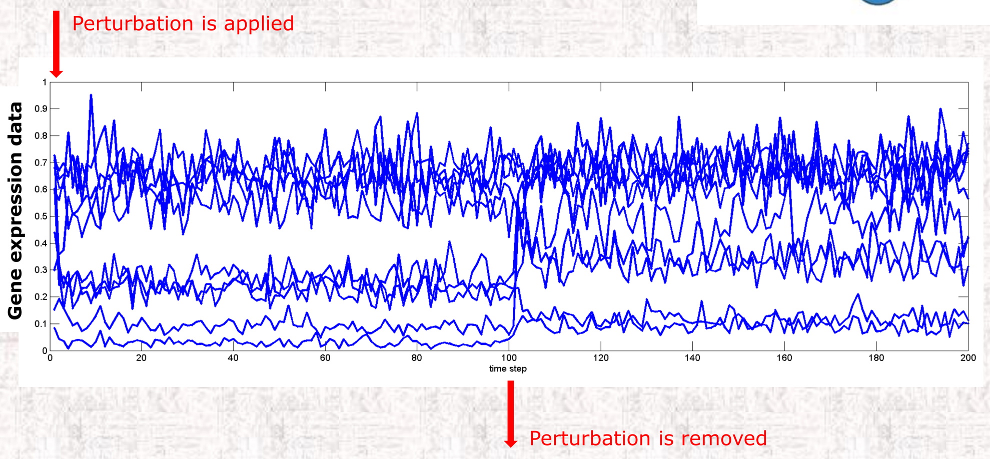
regulatory coefficients - $$a_{ij,t} = a_{ij,t-1} + v_t \quad (2)$$

Explicitly in a vector form $$\begin{bmatrix} x_{1,t} \\ x_{2,t} \\ \dots \\ x_{I,t} \end{bmatrix} = \begin{bmatrix} a_{11,t} & a_{12,t} & \cdots & a_{1I,t} \\ a_{21,t} & a_{22,t} & \cdots & a_{2I,t} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1,t} & a_{I2,t} & \cdots & a_{II,t} \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \\ \dots \\ x_{I,t-1} \end{bmatrix} + \begin{bmatrix} \eta_{1,t} \\ \eta_{2,t} \\ \dots \\ \eta_{I,t} \end{bmatrix} \quad (3)$$

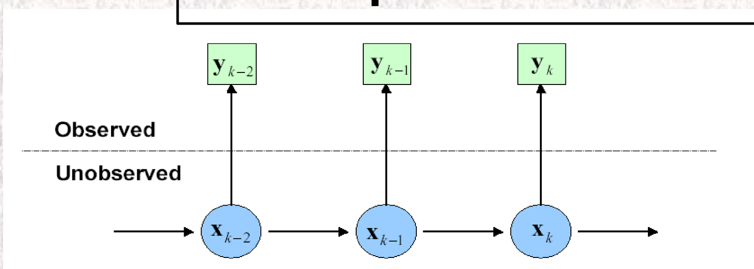
Data: DREAM 4, challenge 2

- Subnetwork (**10 genes**) from transcriptional regulatory networks of *S. cerevisiae.*
- Data are generated from known network topologies.
- Perturbation applied - the mRNA production of several genes is "slightly" perturbed at once.





Particle Filtering or Sequential Monte Carlo Method



Filtering is a problem of sequentially estimating the states (parameters or hidden variables) of a system as a set of observations become available on-line.

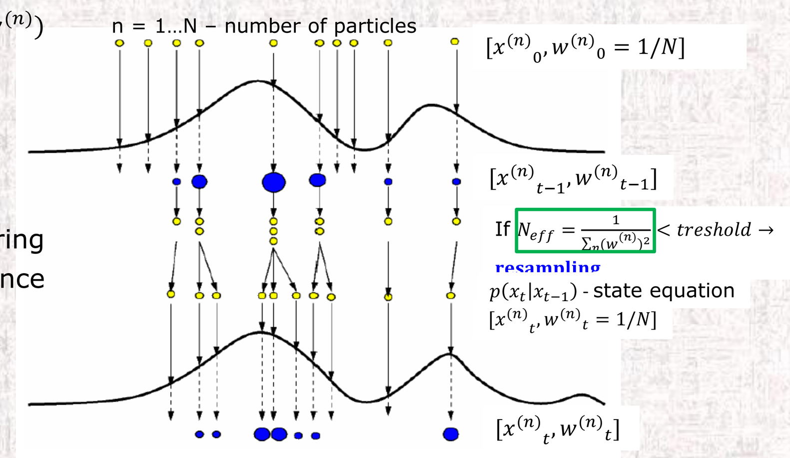
State equation (1st order Markov process): $p(x_t|x_{t-1})$ $\quad x_t = F * x_{t-1} + v_t$

Observation equation: $p(y_t|x_t)$ $\quad y_t = H * x_t + \sigma_t$

- **Weighted particle** presentation $(x^{(n)}, w^{(n)})$ of the posterior distribution:

$$p(x) = \sum_i w^{(n)} \delta(x - x^{(n)})$$

- Approximate Bayesian optimal filtering equations with sequencial importance sampling.



Algorithm

- **Initialization**: Draw *N* samples $x_0^{(n)}$ from the prior

$$x_0^{(n)} \sim p(x_0)$$

and set $w_0^{(n)} = 1/N$.

- **Prediction**: Draw *N* new samples $x_{t+1}^{(n)}$ from importance distributions

$$x_{t+1}^{(n)} \sim q\left(x_{t+1} \middle| x_t^{(n)}, y_{t+1}\right)$$

- **Update**: Calculate new weights according to

$$w_{t+1}^{(n)} \sim w_t^{(n)} \cdot \frac{p\left(y_{t+1} \middle| x_{t+1}^{(n)}\right) p\left(x_{t+1}^{(n)} \middle| x_t^{(n)}\right)}{q\left(x_{t+1}^{(n)} \middle| x_t^{(n)}, y_{t+1}\right)}$$

- A convenient choise for proposal density:

$$q\left(x_{t+1}^{(n)} \middle| x_t^{(n)}, y_{t+1}\right) \sim p\left(x_{t+1}^{(n)} \middle| x_t^{(n)}\right) \;\Rightarrow\; w_{t+1}^{(n)} \sim w_t^{(n)} \cdot \frac{p\left(y_{t+1} \middle| x_{t+1}^{(n)}\right) p\left(x_{t+1}^{(n)} \middle| x_t^{(n)}\right)}{q\left(x_{t+1}^{(n)} \middle| x_t^{(n)}, y_{t+1}\right)}$$

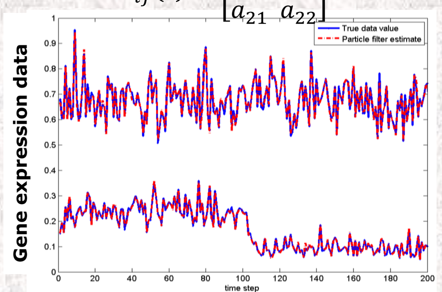
$$w_{t+1}^{(n)} \sim w_t^{(n)} \cdot p\left(y_{t+1} \middle| x_{t+1}^{(n)}\right)$$

Results

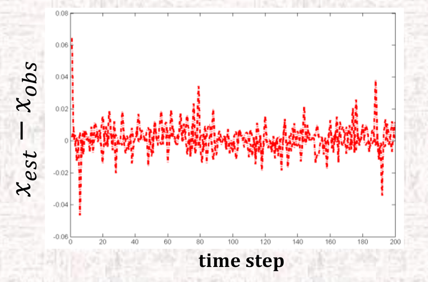
- For 2 genes network:

2-by-2 matrix of coefficients a_{ij}:

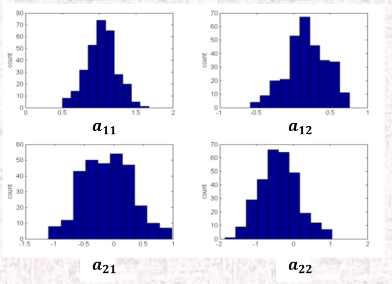
$$a_{ij}(t) = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

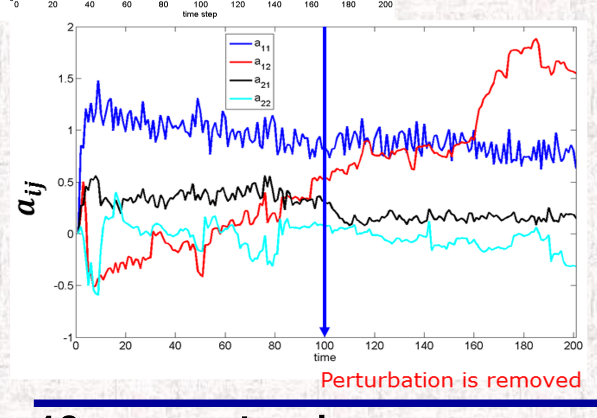


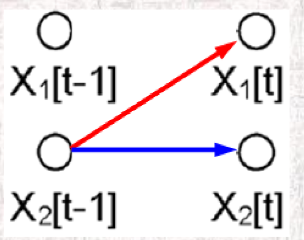
Residuals between synthetic gene expression data, x_{obs}, and estimated by Particle Filter, x_{est}:



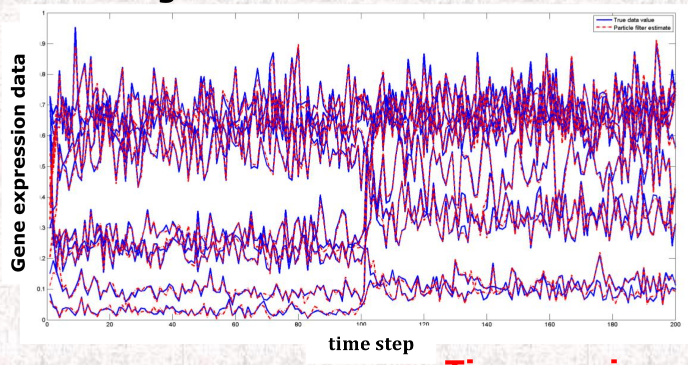
A histogram of estimated values of coefficients a_{ij} at $T = 201$:







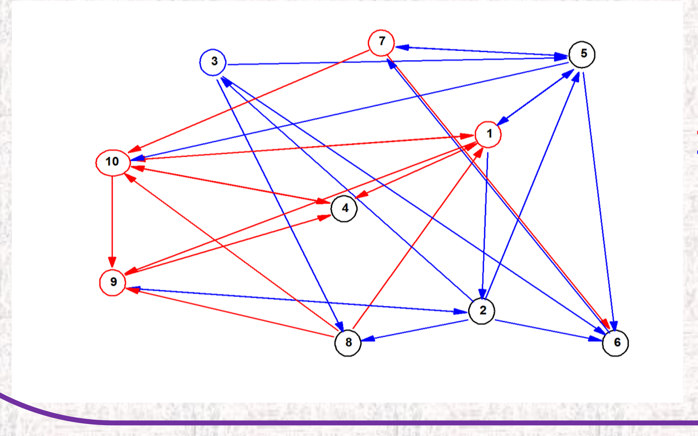
- For 10 genes network:



10-by-10 matrix of coefficients a_{ij}:

$$a_{ij}(t) = \begin{bmatrix} a_{11}, \dots, a_{1N} \\ a_{21}, \dots, a_{2N} \\ \dots \dots \dots \dots \dots \\ a_{N1}, \dots, a_{NN} \end{bmatrix}$$

Time varying network structure



expression and repression activities of one gene to another at adjacent time epochs

self-expression,
self-repression and
no expression of a gene in adjacent time epochs

ACKNOWLEDGMENTS: This work was funded by the CNR Flagship project InterOmics. Ercan E. Kuruoglu gratefully acknowledges partial support from Alexander von Humboldt Foundation in the form of an Experienced Research Fellowship which made his stay in the Max Planck Institute possible.