# Parameter estimation for gene regulatory networks defined by differential equations

Nadia Lalam, Chalmers University of Technology

Joint work with Chris Klaassen, University of Amsterdam and EURANDOM

# Outline

- Gene regulatory network
- Mathematical formalisms
- Statistical model
- Maximum likelihood based inference

# Gene Regulatory Network

- Collection of genes in a cell which interact with each other and with other substances in the cell
- Regulation of gene expressions governs intracellular and extracellular mechanisms
- Aim: Determine the interactions within a gene regulatory network

# Mathematical formalisms

- Many formalisms have been developed for qualitative or quantitative description of gene regulatory networks. They may be:
  - Discrete
  - Continuous
  - Stochastic
  - Deterministic

# Mathematical formalisms

- Boolean Networks: Qualitative discrete models
  - State of a gene is either on or off
  - Connections between genes defining the activation/repression of a gene product on another
  - Change of state defined by an activation function
    Example: $f_\ell(x) = \mathbf{1}(\sum_{k=1}^{N} W_{\ell k} x_k - h_\ell > 0)$,
    $x = (x_1, \ldots, x_N) \in \{0,1\}^N$
  - Synchronous or asynchronous

# Mathematical formalisms

- Bayesian Networks: Probabilistics models
  - Vertices $\leftrightarrow$ Random variables describing the gene expression levels
  - Conditional distributions of the vertices given their direct parents
  - Edges $\leftrightarrow$ Dependencies
  - Static or dynamic

# Mathematical formalisms

- Stochastic Equations:
  - $P(\mathbf{X}, t)$ proba. that $\mathbf{X} = (X_1, \ldots, X_N)$ molecules are inside the cell at time $t$, $r$ reactions

$$P(\mathbf{X}, t + \delta t) = P(\mathbf{X}, t)(1 - \sum_{\ell=1}^{r} \alpha_\ell \delta t) + \sum_{\ell=1}^{r} \beta_\ell \delta t$$

    $\alpha_\ell \delta t$ proba. that $\ell$ occurs during $[t, t + \delta t]$ given that $\mathbf{X}$ molecules at time $t$, and $\beta_\ell \delta t$ proba. that the system is one $\ell$ reaction removed from the state $\mathbf{X}$ at time $t$ and then undergoes $\ell$ in $[t, t + \delta t]$
  - $\delta t \to 0$ entails the Stochastic Master Equation

$$\frac{\partial P(\mathbf{X}, t)}{\partial t} = \sum_{\ell=1}^{r} (\beta_\ell - \alpha_\ell P(\mathbf{X}, t))$$

# ODE's formalism

- State variables $\leftrightarrow$ Gene product concentrations
- The gene product concentrations $g_{k\ell}(t;\theta)$ are assumed to satisfy a set of Ordinary Differential Equations

$$\frac{dg_{k\ell}(t;\theta)}{dt} = \psi_{k\ell}(\mathbf{g}(t;\theta),\theta)$$

$k$ the cell index, $1 \leq k \leq K$
$\ell$ the gene product (protein or mRNA) index, $1 \leq \ell \leq N$
$\mathbf{g}(t;\theta) = [g_{k\ell}(t;\theta)]_{k,\ell}$
$\mathbf{g}(t_0;\theta)$ initial condition
$\theta$ the unknown biological parameter of interest

# Example: Reaction-diffusion model

- Linear array of cells, $N$ interacting genes
- Let $\mathbf{W} = [W_{\ell\ell'}]_{\ell,\,\ell'}$ be the regulatory weight matrix s.t.
  $W_{\ell\ell'} > 0$ if gene $\ell'$ activates the synthesis of $\ell$
  $W_{\ell\ell'} < 0$ if gene $\ell'$ represses $\ell$
  $W_{\ell\ell'} = 0$ if $\ell'$ and $\ell$ do not interact
- Let $D_\ell$ be the diffusion parameter of gene product $\ell$
- Let $\phi_\ell(x) = x^2/(x^2 + h_\ell)$ (Hill function)

$$\begin{aligned}\frac{dg_{k\ell}(t;\theta)}{dt} &= \phi_\ell\Big(\sum_{\ell'=1}^{N} W_{\ell\ell'} g_{k\ell'}(t;\theta)\Big) \\ &\quad + D_\ell[g_{k-1\,\ell}(t;\theta) - 2g_{k\ell}(t;\theta) + g_{k+1\,\ell}(t;\theta)]\end{aligned}$$

- Biological parameter: $\theta = (\mathbf{W}, \mathbf{D}, \mathbf{h})$

# ODE's formalism

- $k$ cell index, $\ell$ gene product index, initial conditions $\mathbf{g}(t_0;\theta)$ and

$$\frac{dg_{k\ell}(t;\theta)}{dt} = \psi_{k\ell}(\mathbf{g}(t;\theta),\theta)$$

- Aim: Infer $\theta$ based on observations of the solutions of the ODE's at different times for different combinations of gene products

# Proteomic data

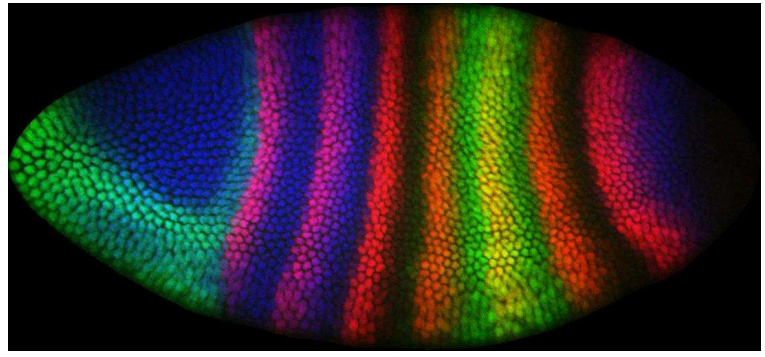
- Scans of stained fixed organisms obtained by confocal laser scanning microscopy
- At maximum three stained gene products at a time whereas the gene regulatory network may contain more than three genes

  $\Rightarrow$ 2D picture
- Stained intensity proportional to the protein concentration

  $\Rightarrow$ Numerical data

# Proteomic data

Image from confocal laser scanning microscopy...



... converted into numerical values

| cell | caudal | hunchback | Kruppel |
|---|---|---|---|
| 4113 | 33.57 | 91.41 | 9.28 |
| 4114 | 37.22 | 88.49 | 11.02 |
| 4115 | 44.17 | 83.57 | 12.91 |
| 4116 | 52.42 | 77.98 | 15.07 |
| 4117 | 59.76 | 71.47 | 17.71 |

...

Source: http://flyex.ams.sunysb.edu/flyex/

Poustelnikova et al., 2004

# Statistical model

- Data obtained by confocal laser scanning microscopy divided into $d$ samples: Each sample contains organisms stained for the same three gene products at the same intended time

- For an observation $j$ from sample $i$, the random variable $X_{i\,j}$ represents its scan converted into numerical values

- Aim: Infer $\theta$, where $g_{k\ell}(t;\theta)$ are the solutions of the ODE's

# Statistical model

- For a sample $i$ formed by $n_i$ observations, let $L_i$ be the subset of three stained proteins among the $N$ interacting genes, $t_i$ the intended observation time

- $X_{i\,j}$ is a table $\{X_{i\,j\,k\,\ell}\}_{k,\ell}$, where $k$ indexes the cells and $\ell \in L_i$ indexes the gene products, such that

$$X_{i\,j\,k\,\ell} = g_{k\ell}(t_i + \delta_{i\,j}; \theta) + \varepsilon_{i\,j\,k\,\ell}$$
$$1 \leq i \leq d,\ 1 \leq j \leq n_i$$

- Account for stochasticity in the time of observation (random error $\delta_{i\,j}$) and for stochasticity in the measurement of gene product (random error $\varepsilon_{i\,j\,k\,\ell}$)

# Statistical model

- Taylor approximation

$$g_{k\ell}(t_i + \delta_{i\,j}; \theta) \approx g_{k\ell}(t_i; \theta) + \frac{\partial g_{k\ell}(t; \theta)}{\partial t}|_{t=t_i} \delta_{i\,j}$$

with $\partial g_{k\ell}(t_i; \theta)/\partial t = \psi_{k\ell}(\mathbf{g}(t_i; \theta), \theta) = \psi_{k\,\ell}(t_i; \theta)$ the RHS of the ODE at time $t = t_i$

- Then

$$X_{i\,j\,k\,\ell} = g_{k\ell}(t_i; \theta) + \psi_{k\,\ell}(t_i; \theta)\delta_{i\,j} + \varepsilon_{i\,j\,k\,\ell}$$

- Assume all $\varepsilon_{i\,j\,k\,\ell}$ i.i.d. with density $f(\cdot)$, mean 0 and variance $\sigma^2_\varepsilon$, all $\delta_{i\,j}$ i.i.d. with density $g(\cdot)$, mean 0 and variance $\sigma^2_\delta$, and $\varepsilon_{i\,j\,k\,\ell}$ independent of $\delta_{ij}$

# Maximum Likelihood Estimation

- Assume $f(\cdot)$ and $g(\cdot)$ Gaussian densities
- Parameter $\gamma = (\theta, \sigma_\varepsilon^2, \sigma_\delta^2)$ estimated by MLE: Maximization in $\gamma$ of the likelihood

$$\prod_{i=1}^{d} \prod_{j=1}^{n_i} \int_{\mathbb{R}} \left[ \prod_{k=1}^{K} \prod_{\ell \in L_i} \frac{1}{\sigma_\varepsilon} \phi \left( \frac{X_{ijk\ell} - g_{k\ell}(t_i;\theta) - \psi_{k\ell}(t_i;\theta)\sigma_\delta y}{\sigma_\varepsilon} \right) \right] \phi(y) dy$$

with $\phi(y) = 1/\sqrt{2\pi} \exp\{-0.5y^2\}$

# Maximum Likelihood Estimation

- The log-likelihood normalized by $-2/n$, with $n = \sum_{i=1}^{d} n_i$ the number of observations, is

$$
\begin{aligned}
M_n(\gamma) \;=\; & \frac{1}{n}\sum_{i,\,j}\{K|L_i|\log\sigma_\varepsilon^2 + \log\left(1+\frac{\sigma_\delta^2}{\sigma_\varepsilon^2}\sum_{k,\,\ell}\psi_{k\,\ell}^2(t_i;\theta)\right) \\
& + \frac{1}{\sigma_\varepsilon^2}\sum_{k,\,\ell}\left(X_{i\,j\,k\,\ell} - g_{k\ell}(t_i;\theta)\right)^2 \\
& - \frac{\sigma_\delta^2\left[\sum_{k,\,\ell}(X_{i\,j\,k\,\ell} - g_{k\ell}(t_i;\theta))\psi_{k\,\ell}(t_i;\theta)\right]^2}{\sigma_\varepsilon^2(\sigma_\varepsilon^2 + \sigma_\delta^2\sum_{k,\,\ell}\psi_{k\,\ell}^2(t_i;\theta))}\}
\end{aligned}
$$

with $\gamma = (\theta, \sigma_\varepsilon^2, \sigma_\delta^2)$

# Maximum Likelihood Estimation

- Particular simple case: No time error ($\sigma_\delta = 0$). Then the MLE in the Gaussian case is the Least Squares Estimator
- Asymptotics: $n \to \infty$ s.t. $\lim_{n\to\infty} n_i/n = p_i > 0$
- The SLLN yields $\lim_{n\to\infty} M_n(\gamma) \stackrel{a.s.}{=} M(\gamma)$
- Let $\Gamma_0 = \{\gamma_* \in \Gamma : \gamma_* = \text{argmin}_{\gamma\in\Gamma} M(\gamma)\}$ be the set of minimizers of the asymptotic criterion $M(\cdot)$
  $\Rightarrow \Gamma_0$ contains the true parameter value $\gamma_0$

# Asymptotic properties

- $\Gamma_0 = \{\gamma_* \in \Gamma : \gamma_* = \text{argmin}_{\gamma \in \Gamma} M(\gamma)\}$, and $n \to \infty$ s.t. $\lim_{n\to\infty} n_i/n = p_i > 0$

- Under smoothness conditions of the RHS of the ODE's, the MLE $\widehat{\gamma}_n$ is consistent in the sense that, for all $\varepsilon > 0$, $\lim_{n\to\infty} P(\inf_{\gamma_* \in \Gamma_0} d(\widehat{\gamma}_n, \gamma_*) \geq \varepsilon) = 0$

- If $\Gamma_0 = \{\gamma_0\}$, then $\widehat{\gamma}_n$ is weakly consistent, i.e. for all $\varepsilon > 0$, $\lim_{n\to\infty} P(d(\widehat{\gamma}_n, \gamma_0) \geq \varepsilon) = 0$

Lalam and Klaassen, 2006

# Asymptotic properties

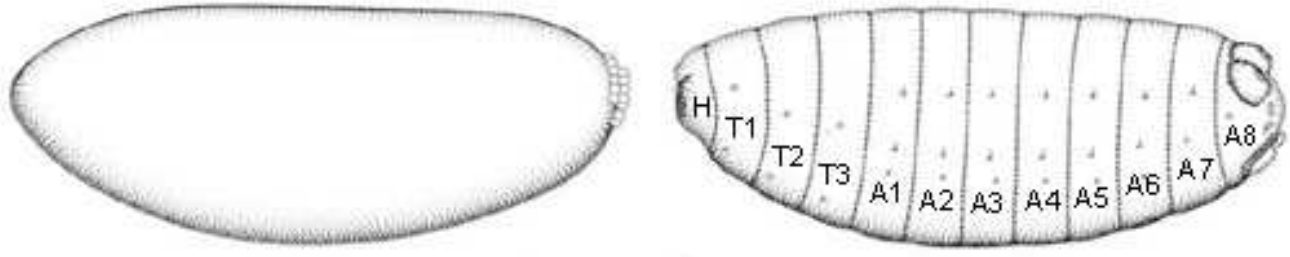
- $\Gamma_0 = \{\gamma_* \in \Gamma : \gamma_* = \text{argmin}_{\gamma \in \Gamma} M(\gamma)\}$, and $n \to \infty$ s.t. $\lim_{n\to\infty} n_i/n = p_i > 0$

- If $\widehat{\gamma}_n$ is weakly consistent, and under regularity assumptions of the RHS of the ODE's and of the expectation of $M_n(\gamma)$, then $\widehat{\gamma}_n$ is root-$n$ consistent, that is $\lim_{M\to\infty} \text{lim sup}_{n\to\infty} P(\sqrt{n} d(\widehat{\gamma}_n, \gamma_0) > M) = 0$

Lalam and Klaassen, 2006

# Drosophila case

- Work in progress: Application to the MLE approach to experimental data
- Gene regulatory network responsible of a particular early stage of Drosophila embryo development: Segmentation



# Drosophila case

- $N = 6$ genes in the network responsible of the early embryo segmentation
- $n = 954$ observations between cleavage cycle 13 (after egg fertilization) and cleavage cycle 14
- Observations: Protein concentrations
- 9 time classes
- $K = 58$ cells along the embryo antero-posterior axis

Mjolsness et al., 1991, Jaeger et al., 2004

# Drosophila case

- Initial conditions at the onset of cleavage cycle 13 and

$$\begin{aligned}\psi_{k\ell}(\mathbf{g}(t;\theta),\theta) &= R_\ell \Phi\Big(\sum_{\ell'=1}^{N_g} W_{\ell\ell'} g_{k\ell'}(t) + m_\ell g_{k\,bcd}(t) + h_\ell\Big) \\ &\quad + D_\ell[g_{k-1\,\ell}(t) - 2g_{k\ell}(t) + g_{k+1\,\ell}(t)] \\ &\quad - \lambda_\ell g_{k\ell}(t)\end{aligned}$$

with $\Phi(x) = 0.5[(x/\sqrt{x^2+1}) + 1]$

- Gene regulation and synthesis/Diffusion/Decay
- $\theta = ((R_\ell, m_\ell, h_\ell, D_\ell, \lambda_\ell)_{1\le\ell\le N}, (W_{\ell\ell'})_{1\le\ell,\ell'\le N})$

Mjolsness et al., 1991, Jaeger et al., 2004

# Summary and Perspectives

- Summary
  - Gene regulatory networks modelled with ODE's
  - Statistical model for data corrupted by two sources of noise in state variable and time observation
  - Inference of a parameter of ODE's by MLE
  - Asymptotic properties and ongoing work to apply the MLE procedure to data
- Perspectives
  - Remove the assumption of Gaussian errors
  - Extend the approach to Partial Differential Equations
  - Account for differences in shape between organisms

# Bibliography


Bickel, P. J., Klaassen, C. A. J., Ritov, Y., Wellner, J. (1998) Springer

Bolouri, H., Davidson, E. H. (2002) BioEssays, 24, 1118-1129

Hasty, J., McMillen, D., Isaacs, F., Collins, J. J. (2001) Nature Reviews Genetics, 2, 268-279

Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K., Manu, Myasnikova, E., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H., Reinitz, J. (2004) Nature, 430, 368-371

de Jong, H. (2002) Journal of Computational Biology, 9, 67-103

Kosman, D., Small, S., Reinitz, J. (1998) Development Genes and Evolution, 208, 290-294

Lalam N., Klaassen C. (2006) EURANDOM Preprint 2006-018

Mjolsness, E., Sharp, D. H., Reinitz, J. (1991) Journal of Theoretical Biology, 152, 429-453

Poustelnikova, E., Pisarev, A., Blagov, M., Samsonova, M., Reinitz, J. (2004) Bioinformatics, 20, 2212-2221

# Acknowledgements


- Joke Blom, CWI Amsterdam
- Richard Gill, University of Leiden and EURANDOM
- Jaap Kaandorp, University of Amsterdam
- Members of the Reinitz Fly Lab, Stony Brook University, New York