Large data sets and complex models:
A view from Systems Biology

Bärbel Finkenstädt, Department of Statistics, University of Warwick

OxWaSP workshop October 2015
Joint work with

**Mathematical and Statistical Modelling:**
Sylvia Calderazzo, Simone Tiberi, Kirsty Hey, Hiroshi Momiji, Dafyd Jenkins, George Minas, David Rand (University of Warwick)

**Experimental Collaboration:**
PRL expression in pituitary gland (Julien Davis, Mike White, University of Manchester)
Chronotherapy and Cancer Research Unit (Francis Lévi and group), Warwick Medical School and INSERM, Paris
Molecular Neurobiology of circadian timing (Michael Hastings and group), MRC Laboratory of Molecular Biology, Cambridge
PRESTA Consortium (University of Warwick)

**Funding:** BBSRC, EPSRC, MRC, EU (BIOSIM), Wellcome
Gene Expression

Diagram showing the process of gene expression:
- DNA molecule
- Gene 1, Gene 2, Gene 3
- Transcription
- mRNA
- Translation
- Protein
- Amino acids (Trp, Phe, Gly, Ser)
- Gene expression measurement

Transcriptome: Genes
Proteome: Proteins
Standard Model of Gene Expression (Single Gene)
Standard Model of Gene Expression  
(Single Gene, ODE Version)

\[
\frac{dm(t)}{dt} = \beta(t) - \delta_m m(t),
\]

\[
\frac{dp(t)}{dt} = \alpha m(t) - \delta_p p(t)
\]

where

\(m(t)\): concentration of mRNA  
\(p(t)\): concentration of protein  
\(\beta(t)\): transcription rate of mRNA  
\(\delta_m, \delta_p\): degradation rate of mRNA, protein
Standard Model of gene expression
(Stochastic version, single gene)

\[
X(t) = \begin{pmatrix}
X_m(t) \\
X_p(t)
\end{pmatrix}
\]

4 reactions (transcription, degradation mRNA, translation, degradation protein)

\[
v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad v_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad v_4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}
\]

A reaction of type \( j \) changes \( X(t) \) to \( X(t) + v_j \).
Each reaction occurs at a rate \( w_j(X(t)) \).
<table>
<thead>
<tr>
<th>Event</th>
<th>Effect</th>
<th>Transition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription</td>
<td>((X_m, X_p) \rightarrow (X_m + 1, X_p))</td>
<td>(w_1 = \beta(t))</td>
</tr>
<tr>
<td>Degradation of mRNA</td>
<td>((X_m, X_p) \rightarrow (X_m - 1, X_p))</td>
<td>(w_2 = \delta_m X_m(t))</td>
</tr>
<tr>
<td>Translation</td>
<td>((X_m, X_p) \rightarrow (X_m, X_p + 1))</td>
<td>(w_3 = \alpha X_m(t))</td>
</tr>
<tr>
<td>Degradation of protein</td>
<td>((X_m, X_p) \rightarrow (X_m, X_p - 1))</td>
<td>(w_4 = \delta_p X_p(t))</td>
</tr>
</tbody>
</table>

Table 1: Summary of reactions in the standard model of gene expression

Reaction networks constitute continuous time Markov jump processes and thus satisfy the Chapman-Kolmogorov equation for which one can obtain the forward form known as the master equation (ME) describing the evolution of the probability \(P(X_m = n_1, X_p = n_2; t)\).

Although an exact numerical simulation algorithm is provided (Gillespie, 1977) the ME is rarely tractable and hence an explicit formula for the exact likelihood is not available for parameter inference.
FIG. 2. Oscillations obtained by numerical simulation of the stochastic model for circadian rhythms with the Gillespie algorithm. The panels show oscillations of mRNA concentration, $M$ (left column), limit cycles (second column), autocorrelation function (third column), and the period distribution (fourth column), for (A) $\Omega = 1000$, (B) $\Omega = 100$, and (C) $\Omega = 10$. The autocorrelation and period histograms have been calculated on a time series of 25,000 h, i.e., more than 1000 cycles. The white curve on the phase plane corresponds to the deterministic limit cycle. The deterministic oscillations have a period of 22 h. Parameter values are: $v_s=1.6 \text{ nM h}^{-1}$, $K_t=1 \text{ nM}$, $n=4$, $v_m=0.905 \text{ nM h}^{-1}$, $K_m=0.5 \text{ nM}$, $k_s=0.5 \text{ h}^{-1}$, $v_d=1.4 \text{ nM h}^{-1}$, $K_d=0.13 \text{ nM}$, $k_1=0.5 \text{ nM h}^{-1}$, $k_2=0.6 \text{ nM h}^{-1}$. 
Gene Networks
Figure 1. A model for the mammalian circadian clock.


http://127.0.0.1:8081/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1002309
\[
\frac{dx}{dt} = k_{f_{1},x7} - k_{d_{1},x1} - d_{1},x1
\]

\[
\frac{dy}{dt} = \frac{1 + g \left( \frac{x_1}{k_{1}} \right)^y}{1 + \left( \frac{PC}{k_{3}} \right) \left( \frac{x_1}{k_{3}} \right)^y} - d_{3},y3
\]

\[
\frac{dz}{dt} = \frac{1 + g \left( \frac{x_1}{k_{1}} \right)^z}{1 + \left( \frac{PC}{k_{3}} \right) \left( \frac{x_1}{k_{3}} \right)^z} - d_{3},z4
\]

\[
\frac{dx_4}{dt} = \frac{1 + g \left( \frac{x_1}{k_{1}} \right)^x_4}{1 + \left( \frac{PC}{k_{3}} \right) \left( \frac{x_1}{k_{3}} \right)^x_4} - d_{3},x4
\]

\[
\frac{dx_5}{dt} = k_{f_{1},x5} - k_{d_{1},z6} - d_{1},z6
\]

\[
\frac{dx_6}{dt} = k_{f_{1},x6} - k_{d_{1},z7} - d_{1},z7
\]

\[
\frac{dx_7}{dt} = k_{f_{1},x7} - k_{d_{1},x5} - d_{1},x5
\]

\[
\frac{dx_8}{dt} = k_{f_{1},x8} - k_{d_{1},z6} - d_{1},z6
\]

\[
\frac{dx_9}{dt} = k_{f_{1},x9} - k_{d_{1},z7} - d_{1},z7
\]
Gene Expression Data
Oligonucleotide microarray

Cells of person 1/condition 1

RNA isolation

mRNA

Reverse transcriptase labelling

cDNA

“Green Fluorescent” Targets

Hybridise to microarray

Microarray with short ssDNA spanning the entire genome

---

cDNA microarray

Cancer cells

RNA isolation

mRNA

Reverse transcriptase labelling

cDNA

“Red Fluorescent” Targets

Combine targets

Microarray with long cDNA covering the transcriptional activity of the cell type

Normal cells
Plant Microarray time series (PRESTA project)
Searching for transcription factors (TFs) of gene regulation from microarray data

Scenario:
Have (replicate) time series microarray gene expression data across various experiments and a set of candidate parents from Y1H experiments
Modeling:

Transcription $\beta(t)$ of a gene (*Child gene*) is regulated by transcriptional activators and/or inhibitors. These are protein products, called Transcription Factors (TFs), of other genes (*Parent genes*).

Wish to draw inference about regulation of mRNA transcription of a child gene by an unknown subset

$$\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_\Gamma\}$$

of the candidate parents

$$\mathcal{G} = \{g_1, g_2, \ldots, g_\mathcal{G}\}$$

Switches occur at times where the expression, $P_\gamma(t)$ of a parent $\gamma \in \Gamma$ crosses a threshold level due to an increase or decrease.
Define activation function

\[ \alpha_{\Gamma}(t) = (\alpha_{\gamma_1}(t), \alpha_{\gamma_2}(t), \ldots, \alpha_{\gamma_{\Gamma}}(t)) \]

The set of switch times of the child’s transcription is the set of time-points \( \{s_1, s_2, \ldots, s_k\} \) at which at least one parent crosses its threshold.

The total time interval split into subintervals \([0, s_1], (s_1, s_2], \ldots, (s_k, L]\) where the transcription rates of the subintervals are the same if the expression of all parent genes remains at the same state.

For a given activation function \( \alpha_{\Gamma}(t) \) we then have a set of parental states, \( b_j, j = 1, 2, \ldots, \kappa \) observed in the union \( I_{b_j} \) of (at least one of) the above subintervals with the corresponding transcription rates \( \tau_{b_j}, j = 1, 2, \ldots, \kappa \)
\[
\frac{dM}{dt} = \begin{cases} 
\tau_{b_1} - \delta_M M(t), & \text{for } t \in I_{b_1} \\
\tau_{b_2} - \delta_M M(t), & \text{for } t \in I_{b_2} \\
\vdots & \vdots \\
\tau_{b_{\kappa}} - \delta_M M(t), & \text{for } t \in I_{b_{\kappa}}.
\end{cases}
\]
Bayesian approach: RJMCMC to infer plausible models for parents

Likelihood based on normal error with inhomogeneous variance

Weighted Least Squares solution to piecewise linear ODE makes algorithm fast

Delayed RNA as proxy for unobserved TF’s
Single cell imaging data
Reporter Gene constructs

DNA → mRNA → Protein

regulation → mRNA → degradation

translation → mRNA → degradation

proteins → degradation → 

transcription → mRNA → degradation

reporter mRNA → degradation

reporter Protein → degradation

Light Intensity → measurements
Modelling single cell data

The class of state space models provides a unifying framework for modelling SRNs

\[
Y_t \sim g(y_t | x_t, \theta)
\]

\[
X_{t+1} \sim h(x_{t+1} | \theta)
\]

\(h\) is the transition density of the approximating SRN

\(g\) is the density of the measurement process
Approximations

- ODE (neglects intrinsic noise)
- Chemical Langevin Equation (usually intractable)
- Linear Noise Approximation (linearization of the ME, tractable)
- Other approximations?
Parameter Inference

The data likelihood is given by the marginal density

\[ f(y|\theta) = \int_x f(y, x|\theta)dx \]

where the integrand can be factorized as

\[ f(y, x|\theta) = h(x_0|\theta)g(y_0|x_0, \theta) \prod_{t=1}^{T} h(x_t|x_{t-1}, \theta)g(y_t|x_t, \theta) \]
Parameter Inference

Under the LNA (with Gaussian measurement error):
Interval can be evaluated explicitly using Kalman methodology.
Inference can be achieved by sampling from the posterior $f(\theta|y)$

Under BDA (for example) : use 2-step Gibbs sampler:

1. Sample the parameter vector from $f(\theta|x, y)$
2. Sample the latent states $x$ from the filtering density $f(x|y, \theta)$
Challenges for statisticians in Systems Biology and Systems Medicine

• Complex and large networks of genes and their products

• Data from many sources:
  - Replicates and if so what type?
  - Many cells (Bayesian Hierarchical Modeling)
  - Various labs (Prior Distributions)
  - Destructive Sampling
  - ....

• Different Types of Experiments imply different modeling approaches
  - Information about intrinsic noise?
  - Extrinsic noise?
  - Destructive sampling?
  - Reporter gene constructs? Which reporter gene?
  - Type of camera used for imaging
• Not all state variables are observed
  Can use distributed delay functions as a proxy?

• Parameter Identifiability

• Spatio-temporal modelling and inference

• Collaboration between experimentalists and mathematician/statisticians
Research Questions

• Mammalian Circadian Oscillator under diseases and treatments. Exploitation for Chronotherapy

• Development of signal processing tools for circadian time series to be used in forecasting systems for patient care at home

• Development of Spatio-temporal models to investigate how cells synchronize in the SCN
Implantation pompé

Jour après inoculation tumorale
Implantation pump

DEN (15 mg/kg/j, ip)

CT490: KI/KI Per2::luc male mouse (B1)
Figure 1. A model for the mammalian circadian clock.

http://127.0.0.1:8081/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1002309