Bayesian Inference for Systems Biology Models via a Diffusion Approximation

Andrew Golightly
School of Mathematics & Statistics

Darren Wilkinson
School of Mathematics & Statistics
and
Centre for Integrated Systems Biology of Ageing and Nutrition
Newcastle University, UK

Parameter Estimation in Systems Biology,
University of Manchester 28-29th March 2007

Overview

- Introduction
- Markov process models of biochemical network dynamics
- A diffusion approximation
    - Estimating diffusion parameters
- Application: Toy prokaryotic auto-regulatory network
- Summary & future directions

Computational Systems Biology (CSB)

- Concerned with building models of complex biological pathways, then validating and analysing those models using a variety of methods, including time-course simulation
- The traditional approach involves working with continuous deterministic models (e.g.coupled ODEs)
- There is increasing evidence that much intra-cellular behaviour (including gene expression) is intrinsically stochastic, and that systems cannot be properly understood unless stochastic effects are incorporated into the models
- Stochastic models are harder to build, estimate, validate, analyse and simulate than deterministic models...

Modelling

- Start with a set of (pseudo-)biochemical reactions
- Specify the rate laws and rate parameters of the reactions
- Run some stochastic or deterministic computer simulator of the system dynamics
- Straightforward using the Gillespie algorithm. The reverse problem is trickier – given time course data, and a set of reactions, can we recover the rates?

Mass Action Kinetics

Second Order Reaction

$$Y_1 + Y_2 \longrightarrow Y_3$$

This will occur when a molecule of Y_1 collides with a molecule of Y_2

- For a small, fixed volume (V) and assuming thermal equilibrium, the hazard of molecules colliding is constant (Gillespie, 1992).
- We assume the law of mass action such that the hazard of the above reaction $\propto Y_1 Y_2$.

Mass Action Kinetics (2)

Generically

k species and r reactions with a typical reaction

$$R_i : \quad u_{i1}Y_1 + \ldots + u_{ik}Y_k \quad \longrightarrow \quad v_{i1}Y_1 + \ldots + v_{ik}Y_k$$

- Each R_i has a stochastic rate constant, c_i and hazard $h_i(Y, c_i)$ where $Y = (Y_1, \ldots, Y_k)'$ is the current state of the system.
- Every system has a $r \times k$ net effect matrix, $A = (a_{ij})$ where

$$a_{ij} = v_{ij} - u_{ij}$$

Markov Process Models

Traditionally based on solving the “chemical master equation” for

$$\mathrm{P}(Y;t) = \mathrm{P}\left(Y_1, \ldots, Y_k \text{ molecules in } V \text{ at time } t\right)$$

Derive the M-eq. by noting that

$$\mathrm{P}(Y;t+\Delta t) = \sum_{i=1}^{r} h_i(Y - A_i', c_i)\mathrm{P}(Y - A_i';t)\Delta t + \left\{1 - \sum_{i=1}^{r} h_i(Y, c_i)\Delta t\right\}\mathrm{P}(Y;t)$$

which leads to the M-eq.

$$\frac{\partial}{\partial t}\mathrm{P}(Y;t) = \sum_{i=1}^{r}\{h_i(Y - A_i', c_i)\mathrm{P}(Y - A_i';t) - h_i(Y, c_i)\mathrm{P}(Y;t)\}$$

However

- M-eq is only tractable for a handful of cases
- Therefore stochastic models are typically examined using the Gillespie algorithm

The Gillespie algorithm

1. Initialise the system at $t = 0$ with rate constants $c_1, c_2, \ldots, c_r$ and initial numbers of molecules for each species, $Y_1, Y_2, \ldots, Y_k$.
2. Calculate $h_0(Y, c) \equiv \sum_{i=1}^{r} h_i(Y, c_i)$, the combined reaction hazard.
3. Simulate time to next event, $t' \sim Exp(h_0(Y, c))$ random quantity, and put $t := t + t'$.
4. Simulate the reaction index, j, as a discrete random quantity with probabilities $h_i(Y, c_i) \;/\; h_0(Y, c)$, $i = 1, 2, \ldots, r$.
5. Update Y according to reaction j. That is, put $Y := Y + A_j'$, where A_j denotes the jth row of the net effect matrix A.
6. Output Y and t.
7. If $t < T_{max}$, return to step 2.

Example: Lotka-Volterra

Reactions

$$\begin{array}{llcll} R_1: & Y_1 & \longrightarrow & 2Y_1 & \text{Prey reproduction} \\ R_2: & Y_1 + Y_2 & \longrightarrow & 2Y_2 & \text{Predator eats prey} \\ R_3: & Y_2 & \longrightarrow & \emptyset & \text{Predator dies} \end{array}$$

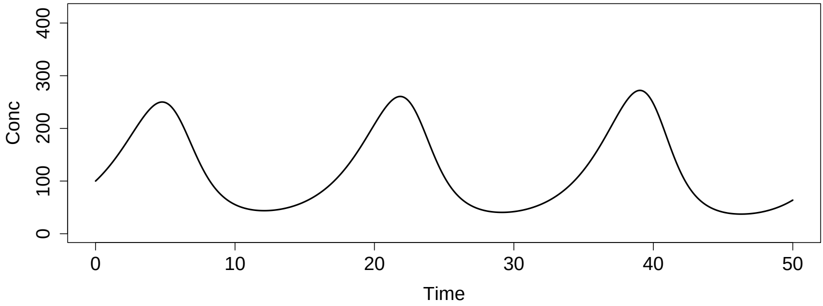
- If the discreteness and stochasticity are ignored, then it is straightforward to deduce the mass-action ODE system:

Lotka-Volterra: ODE Model

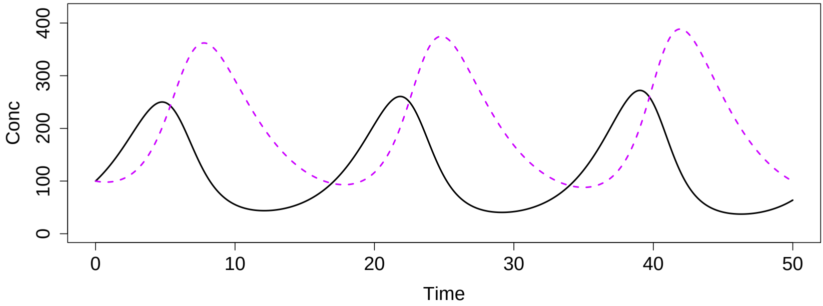
$$\begin{aligned} \frac{dY_1}{dt} &= c_1 Y_1 - c_2 Y_1 Y_2 \\ \frac{dY_2}{dt} &= c_2 Y_1 Y_2 - c_3 Y_2 \end{aligned}$$

- Analytic solutions are rarely available, but good numerical solvers can generate time course behaviour

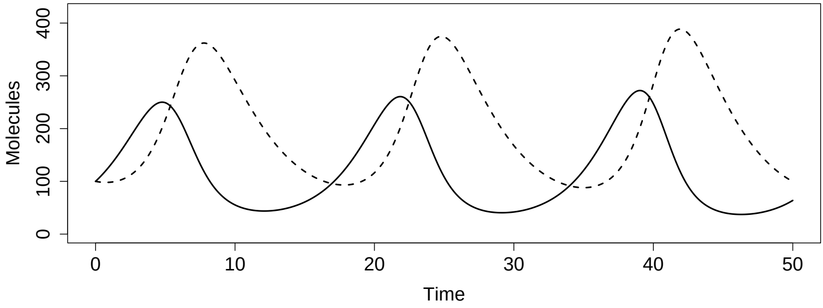
The Lotka-Volterra model



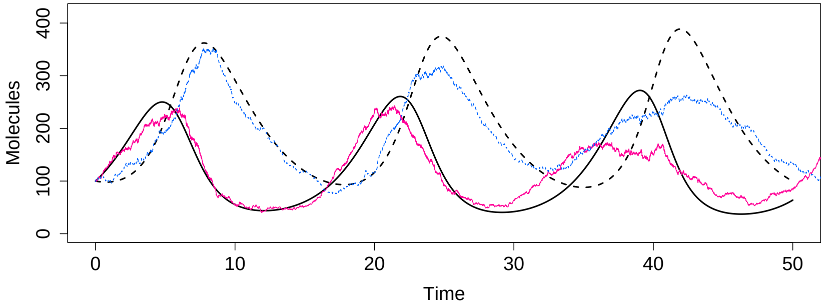
The Lotka-Volterra model



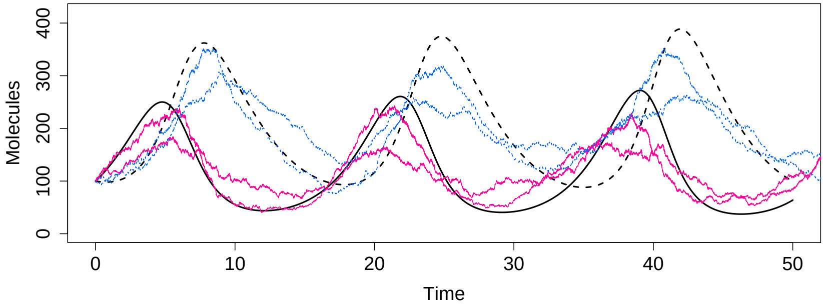
The Lotka-Volterra model



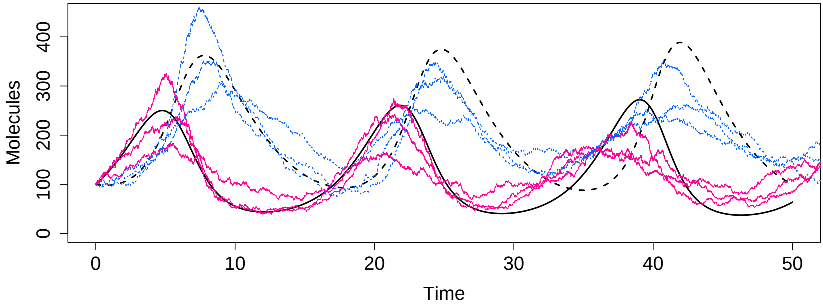
The Lotka-Volterra model



The Lotka-Volterra model



The Lotka-Volterra model



Key differences

- Deterministic solution is exactly periodic with perfectly repeating oscillations, carrying on indefinitely
- Stochastic solution oscillates, but in a random, unpredictable way
- Stochastic solution will end in disaster! Either prey or predator numbers will hit zero...
- Either way, predators will end up extinct, so expected number of predators will tend to zero — qualitatively different to the deterministic solution
- So, in general the deterministic solution does not provide reliable information about either the stochastic process or its average behaviour

Fully Bayesian inference

- In principle it is possible to carry out rigorous statistical inference for the parameters of the stochastic process model
- Techniques for exact inference for the true discrete model (Boys, Wilkinson, Kirkwood 2004) do not scale well to problems of realistic size and complexity
- True process is discrete and stochastic — stochasticity is vital — what about discreteness?
- Apply the Fokker-Planck equation to the Master equation for the true process to obtain an SDE known as the Chemical Langevin Equation (CLE)

The Stochastic-Kinetic Diffusion Approximation

Chemical Langevin Equation (Itô SDE)

$$dY_t = A'h(Y_t, c)dt + [A' \operatorname{diag}\{h(Y_t, c)\}A]^{1/2}\, dW_t$$

- Fairly general class of non-linear multivariate SDEs
- The net effect matrix A is typically rank-degenerate, which complicates things slightly
- A is known and Y (or a subset) is observed at discrete times (subject to error)
- Inference is for c (the vector of rate constants parameterising the reaction rate vector, $h(\cdot, \cdot)$)

Inference for Diffusions

- Set $\mu(Y_t, c) = A'h(Y_t, c)$, $\beta(Y_t, c) = A' \, \mathrm{diag}\{h(Y_t, c)\}A$
- Need to consider the general problem of inferring parameters c governing

$$dY_t = \mu(Y_t, c)dt + \beta^{\frac{1}{2}}(Y_t, c)dW_t$$

  using observations (that may be incomplete and subject to error) at discrete times
- Problem: For μ and β nonlinear, analytic solutions rarely available
    - Can't obtain underlying transition densities!
    - Likelihood inference non-trivial

Bayesian Imputation approach

Work with the Euler discretisation

$$\Delta Y_t = \mu(Y_t, c)\Delta t + \beta^{\frac{1}{2}}(Y_t, c)\Delta W_t\,, \qquad \Delta W_t \sim N_d(0, I\Delta t)$$

- Inter-obs. time, Δ^*, usually too big to use as Δt!
- Set $\Delta t = \Delta^*/m$, choose m large so that Δt is small
- Gives $m-1$ latent values between every pair of obs
- Augmented data in matrix form,

$$\hat{Y} = \begin{pmatrix} y_{t_0} & Y_{t_1} & \cdots & Y_{t_{m-1}} & y_{t_m} & Y_{t_{m+1}} & \cdots\cdots & Y_{t_{n-1}} & y_{t_n} \end{pmatrix}$$

- For data, D_n, formulate joint posterior for c and missing values $\hat{Y}\backslash\{D_n\}$

$$\pi(c, \hat{Y}\backslash\{D_n\}|D_n) \propto \pi(c) \times \prod_{i=0}^{n-1} \pi(Y_{i+1}|Y_i, c)$$

- Integrate over our uncertainty for $\hat{Y}$ using MCMC

Gibbs Sampling

Could sample $\pi(c, \hat{Y} \backslash \{D_n\} | D_n)$ by alternating between

- draws of missing data (e.g. one column at a time) conditional on c and D_n (Metropolis step)
- draws of c conditional on augmented data, $\hat{Y}$ (Metropolis step)

However, if the diffusion coefficient is not free of c, the algorithm is *reducible*

- For $m \to \infty$, there is an infinite amount of information in the augmented sample $\hat{Y}$

Solution (due to Roberts & Stramer, '01): Find an analytic transformation of the diffusion to constant volatility

- Typically impossible to implement for interesting nonlinear diffusions

Irreducible Global MCMC Schemes

- Idea (Chib, Pitt & Shephard, '06). Gibbs sampler: Draw from $c|\hat{W}$ rather than $c|\hat{Y}$ thereby breaking the problematic dependence. Target:

$$\pi(c|\hat{W}) \propto \pi(c)\,\pi(g(\hat{W},c)|c) \;\times\; \text{Jacobian}$$

- Conditional on c, there is a one-to-one relationship between $\hat{Y}$ and $\hat{W}$ – the skeleton of the driving B.M.
- Numerically map between the diffusion sample paths and the corresponding sample paths of the driving Brownian motion, for example using the Euler-Maruyama discretisation

$$\Delta Y_t = \mu(Y_t, c)\Delta t + \beta^{\frac{1}{2}}(Y_t, c)\Delta W_t$$
$$\Rightarrow \Delta W_t = \beta^{-\frac{1}{2}}(Y_t, c)[\Delta Y_t - \mu(Y_t, c)\Delta t]$$

- Problem: unless the diffusion is observed very indirectly, changing the parameters causes the sample paths to “miss” the data points, rendering it impractical

Modified Innovation Scheme

- (Golightly & Wilkinson, ’06): Use the modified diffusion bridge MDB construct of Durham and Gallant ’02 as a template for building sample paths, and use the Wiener processes driving the MDB as our sampler components
- Thinking just about a discretisation of $[0, 1]$ and the fully observed case, we can map back and forth using the deterministic transformations

$$\Delta Y_t = \frac{y_1 - Y_t}{1-t}\Delta t + \left(\frac{1-t-\Delta t}{1-t}\beta(Y_t, c)\right)^{\frac{1}{2}} \Delta W_t$$

$$\Rightarrow \Delta W_t = \left(\frac{1-t}{1-t-\Delta t}\right)\beta^{-\frac{1}{2}}(Y_t, c)\left[\Delta Y_t - \frac{y_1 - Y_t}{1-t}\Delta t\right]$$

- Crucially, there is no problem with failing to “hit” data points after transforming back to the observed diffusion

Algorithm

1. Initialise parameters c, and latent data $\hat{Y} \backslash \{D_n\}$
2. For times $t_0, t_m, \ldots, t_{n-m}$ update latent data in blocks of size $m-1$ using the MDB, and accept/reject with a M-H step
3. Map from $\hat{Y}$ to $\hat{W}$ using the MDB transformation on each interval
4. Propose a new parameter $c^\star$. Using $c^\star$ with fixed $\hat{W}$, deterministically construct the corresponding sample path $\hat{Y}^\star$, and accept/reject the pair jointly with a M-H step
5. Output state and return to step 2

Generalisations to noisy/imperfect observations are straightforward

Acceptance probabilities

- Let $\mathbf{Y}_m$ denote all latent values in (t_j, t_{j+m})
- The acceptance probability for a single interval path update on (t_j, t_{j+m}) takes the form

$$A = \frac{\pi(\mathbf{Y}_m^*|c, y_j, y_{j+m})}{\pi(\mathbf{Y}_m|c, y_j, y_{j+m})} \times \frac{q(\mathbf{Y}_m|c, y_j, y_{j+m})}{q(\mathbf{Y}_m^*|c, y_j, y_{j+m})}$$

- The acceptance probability for a proposed update to c^* takes the form

$$A = \frac{\pi(c^*)}{\pi(c)} \times \frac{f(c|c^*)}{f(c^*|c)} \times \frac{\frac{\pi(\hat{Y}^*|c^*)}{q(\hat{Y}^*|c^*)}}{\frac{\pi(\hat{Y}|c)}{q(\hat{Y}|c)}}$$

Toy Application: Prokaryotic Auto-Regulation

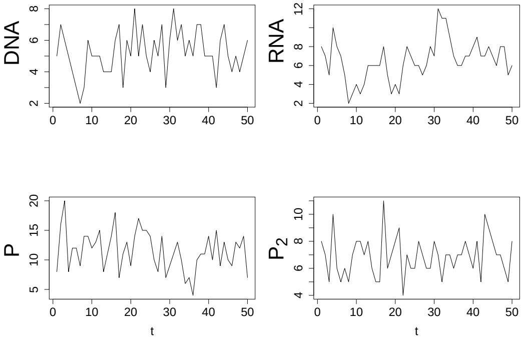
Reaction list:

R_1 :	$\mathrm{DNA} + \mathrm{P}_2$	$\longrightarrow$	$\mathrm{DNA} \cdot \mathrm{P}_2$	Repression
R_2 :	$\mathrm{DNA} \cdot \mathrm{P}_2$	$\longrightarrow$	$\mathrm{DNA} + \mathrm{P}_2$	
R_3 :	DNA	$\longrightarrow$	$\mathrm{DNA} + \mathrm{RNA}$	Transcription
R_4 :	RNA	$\longrightarrow$	$\mathrm{RNA} + \mathrm{P}$	Translation
R_5 :	$2\mathrm{P}$	$\longrightarrow$	P_2	Dimerisation
R_6 :	P_2	$\longrightarrow$	$2\mathrm{P}$	
R_7 :	RNA	$\longrightarrow$	$\emptyset$	Degradation
R_8 :	P	$\longrightarrow$	$\emptyset$	

- 5 species $\mathrm{DNA}, \mathrm{DNA} \cdot \mathrm{P}_2, \mathrm{RNA}, \mathrm{P}, \mathrm{P}_2$ and 8 reactions with rate constants $c = (c_1, \ldots, c_8)'$
- Note that DNA and $\mathrm{DNA} \cdot \mathrm{P}_2$ are deterministically related
- Induces a 4-dimensional diffusion process parameterised by c

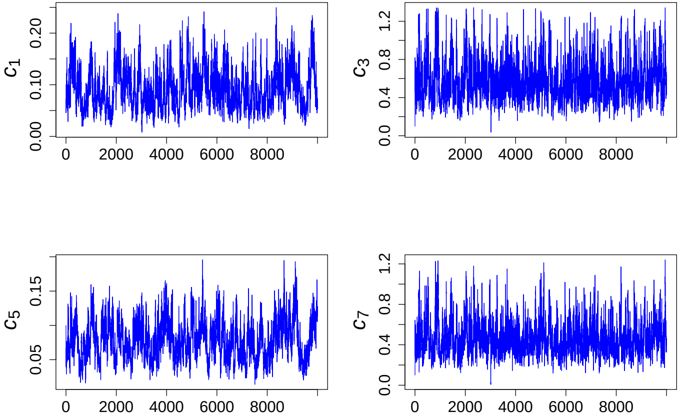
Simulation Study

- 50 obs simulated using the Gillespie algorithm

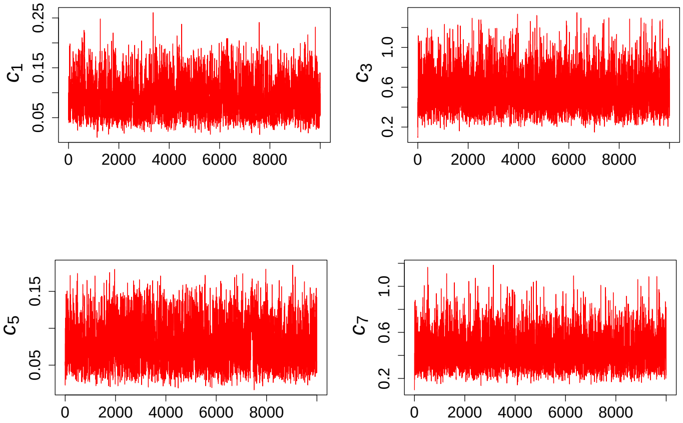


- Rate constants $c = (0.1, 0.7, 0.35, 0.2, 0.1, 0.9, 0.3, 0.1)'$
- Run the innovation scheme to recover these values

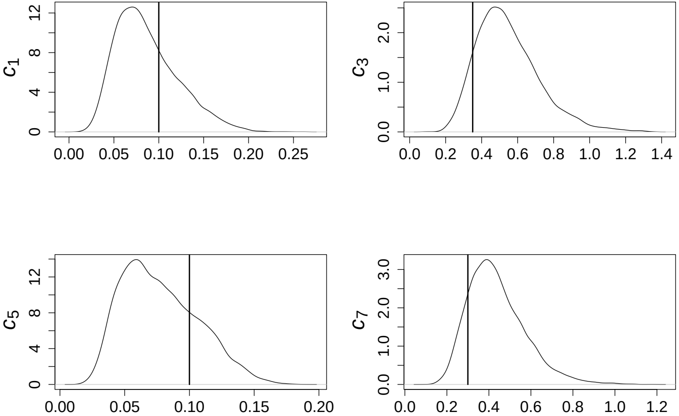
Results, $m = 10$, Gibbs Sampler



Results, $m = 10$, Innovation Scheme



Results, $m = 10$, Innovation Scheme



Results, $m = 10$, Innovation Scheme

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
True Values								
	0.1	0.7	0.35	0.2	0.1	0.9	0.3	0.1
Observe (DNA, RNA, P, P_2)								
Mean	0.087	0.655	0.547	0.055	0.078	0.758	0.437	0.038

Results, $m = 10$, Innovation Scheme

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
True Values								
	0.1	0.7	0.35	0.2	0.1	0.9	0.3	0.1
Observe (DNA, RNA, P, P_2)								
Mean	0.087	0.655	0.547	0.055	0.078	0.758	0.437	0.038
Observe (RNA, P, P_2)								
Mean	0.061	0.451	0.497	0.024	0.072	0.702	0.393	0.020

Results, $m = 10$, Innovation Scheme

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
				True Values				
	0.1	0.7	0.35	0.2	0.1	0.9	0.3	0.1
				Observe (DNA, RNA, P, P_2)				
Mean	0.087	0.655	0.547	0.055	0.078	0.758	0.437	0.038
				Observe (RNA, P, P_2)				
Mean	0.061	0.451	0.497	0.024	0.072	0.702	0.393	0.020
				Observe (RNA)				
Mean	0.047	0.262	0.540	0.049	0.023	0.153	0.461	0.034

Summary

- Systems Biology and post-genomics are full of interesting (hard) statistical problems
- It appears promising to consider the problem of understanding biochemical network dynamics in terms of inference for the Chemical Langevin Equation
- Inference for arbitrary multivariate diffusions observed partially, discretely and with error is non-trivial
- It is possible, however, to implement global MCMC schemes which do not break down for large amounts of augmentation

- Boys, R. J., Wilkinson, D.J. and T.B.L. Kirkwood (2004). Bayesian inference for a discretely observed stochastic kinetic model. In submission.
- Golightly, A. and D. J. Wilkinson (2006). Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology. 13*(3), 838–851.
- Golightly, A. and D. J. Wilkinson (2006). Bayesian inference for nonlinear multivariate diffusion models observed with error. In submission.
- Wilkinson, D. J. (2006). *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC Press.

Contact details...

email: a.golightly@ncl.ac.uk
www: http://www.mas.ncl.ac.uk/~nag48/