Heterogeneity, stochasticity and uncertainty in systems biology models

Darren Wilkinson

tinyurl.com/darrenjw

School of Mathematics & Statistics and Centre for Integrated Systems Biology of Ageing & Nutrition (CISBAN)
Newcastle University, UK

UCM 2010
Sheffield
13th July 2010
Overview

- Example: Regulation of motility in *Bacillus subtilis*
- Stochastic kinetic models
- Bayesian inference for stochastic kinetic models
- Emulating complex stochastic models
- Modelling data: single-cell time-lapse microscopy and GFP reporters
- Layers of uncertainty
- Summary
**Bacillus subtilis**

- *Bacillus subtilis* is the most widely studied model gram positive bacterium.
- It is easy to culture and highly genetically tractable (naturally competent for transformation).
- Relatively interesting life cycle — must make some expensive cellular decisions.
- Can grow and divide, become competent, sporulate, or become motile.
- The motility regulation network is large and complex — for illustrative purposes, just focus on one aspect — repression of fla/che by CodY.
Motility regulation

- The $\sigma$ factor $\sigma^D$ (SigD) is key to the regulation of motility — many genes and operons encoding motility related proteins are governed by this $\sigma$ factor
- $hag$ (encoding flagellin) is one gene with a $\sigma^D$ promoter
- $sigD$ is embedded in a large operon — the fla/che operon
- fla/che strongly repressed by CodY, and CodY is up-regulated in good nutrient conditions

\[ \text{CodY} \quad \longrightarrow \quad \text{fla/che-SigD} \quad \downarrow \]
\[ \text{Hag} \]
Stochasticity in bacterial decision making

- Cellular processes are not deterministic, and the motility decision is no exception.
- There is considerable cell to cell variability in motility onset, even for isogenic batch cultures.
- Even in good nutrient conditions, a small fraction of cells will turn on motility function, and this is possibly evolutionarily beneficial.
- Although there are many other sources of heterogeneity, **intrinsic stochasticity** in gene expression is of fundamental importance to understanding and modelling this process.
Stochastic chemical kinetics

- $u$ species: $X_1, \ldots, X_u$, and $v$ reactions: $\mathcal{R}_1, \ldots, \mathcal{R}_v$
- $\mathcal{R}_i: p_{i1}X_1 + \cdots + p_{iu}X_u \longrightarrow q_{i1}X_1 + \cdots + q_{iu}X_u, \ i = 1, \ldots, v$
- In matrix form: $PX \longrightarrow QX$ ($P$ and $Q$ are sparse)
- $S = (Q - P)'$ is the stoichiometry matrix of the system
- $X_{jt}$: \# molecules of $X_j$ at time $t$. $X_t = (X_{1t}, \ldots, X_{ut})'$
- Reaction $\mathcal{R}_i$ has hazard (or rate law, or propensity) $h_i(X_t, c_i)$, where $c_i$ is a rate parameter, $c = (c_1, \ldots, c_v)'$, $h(X_t, c) = (h_1(X_t, c_1), \ldots, h_v(X_t, c_v))'$ and the system evolves as a Markov jump process
- For mass-action stochastic kinetics,

$$h_i(X_t, c_i) = c_i \prod_{j=1}^{u} \left( \frac{X_{jt}}{p_{ij}} \right), \quad i = 1, \ldots, v$$
Motility model

\begin{align*}
codY & \rightarrow codY + CodY \\
CodY & \rightarrow \emptyset \\
flache & \rightarrow flache + SigD \\
SigD & \rightarrow \emptyset \\
SigD_{hag} & \rightarrow SigD + hag + Hag \\
Hag & \rightarrow \emptyset \\
SigD + hag & \rightarrow SigD_{hag} \\
SigD_{hag} & \rightarrow SigD + hag \\
CodY + flache & \rightarrow CodY_{flache} \\
CodY_{flache} & \rightarrow CodY + flache \\
CodY + hag & \rightarrow CodY_{hag} \\
CodY_{hag} & \rightarrow CodY + hag
\end{align*}
Sample model trajectory

Composition against time

- CodY
- Hag
- SigD

Number of molecules against time.
Bayesian inference

- In principle it is possible to carry out rigorous exact Bayesian statistical inference for the parameters of stochastic kinetic models — some clues in the paper
- Fairly detailed experimental data are required — e.g., quantitative single-cell time-course data derived from live-cell imaging
- The standard procedure uses GFP labelling of key reporter proteins together with time-lapse confocal microscopy, but other approaches are also possible
- Global MCMC algorithms for exact inference for the true discrete model (Boys, W, Kirkwood 2008) do not scale well to problems of realistic size and complexity, due to the difficulty of efficiently exploring large complex integer lattice state spaces
Likelihood-free MCMC approach

Let \( \pi(x|c) \) denote the (complex) likelihood of the simulation model.
Let \( \pi(D|x, \tau) \) denote the (simple) measurement error model.
Put \( \theta = (c, \tau) \), and let \( \pi(\theta) \) be the prior for the model parameters.
The joint density can be written

\[
\pi(\theta, x, D) = \pi(\theta) \pi(x|\theta) \pi(D|x, \theta).
\]

Interest is in the posterior distribution \( \pi(\theta, x|D) \).
Posterior sampling

- Posterior distribution $\pi(\theta, x|D)$
- Propose a joint update for $\theta$ and $x$ as follows:
  - Current state of the chain is $(\theta, x)$
  - First sample $\theta^* \sim f(\theta^*|\theta)$
  - Then sample a new path, $x^* \sim \pi(x^*|\theta^*)$
  - Accept the pair $(\theta^*, x^*)$ with probability $\min\{1, A\}$, where

\[
A = \frac{\pi(\theta^*)}{\pi(\theta)} \times \frac{f(\theta|\theta^*)}{f(\theta^*|\theta)} \times \frac{\pi(D|x^*, \theta^*)}{\pi(D|x, \theta)}.
\]

- Note that choosing a prior independence proposal of the form $f(\theta^*|\theta) = \pi(\theta^*)$ leads to the simpler acceptance ratio

\[
A = \frac{\pi(D|x^*, \theta^*)}{\pi(D|x, \theta)}
\]
"Likelihood-free" MCMC

- Crucially, because the proposal exploits a forward simulation, the acceptance probability does not depend on the likelihood of the simulator output — important for complex stochastic models.

- This scheme is completely general, and works very well provided that $|\mathcal{D}|$ is small.

- **Problem**: If $|\mathcal{D}|$ is large, the MCMC scheme will mix very poorly (very low acceptance rates).

- **Solution**: Exploit the Markovian structure of the process, and adopt a sequential approach, updating one (or a small number of) observation(s) at a time...
Sequential likelihood-free algorithm

- Data $\mathcal{D}_t = \{d_1, \ldots, d_t\}$, $\mathcal{D} \equiv \mathcal{D}_n$. Sample paths 
  $x_t \equiv \{x_s | t - 1 < s \leq t\}$, $t = 2, 3, \ldots, n$, so that
  $x \equiv \{x_2, \ldots, x_n\}$.

1. Assume at time $t$ we have a (large) sample from $\pi(\theta, x_t | \mathcal{D}_t)$ (for time 0, initialise with sample from prior)
2. Run an MCMC algorithm which constructs a proposal in two stages:
   1. First sample $(\theta^*, x^*_t) \sim \pi(\theta, x_t | \mathcal{D}_t)$ by picking at random and perturbing slightly (sampling from the kernel density estimate)
   2. Next sample $x^*_{t+1}$ by forward simulation from $\pi(x^*_{t+1} | x^*_t, \theta^*)$
   3. Accept/reject $(\theta^*, x^*_{t+1})$ with
      $$A = \frac{\pi(d_{t+1} | x^*_{t+1}, \theta^*)}{\pi(d_{t+1} | x_{t+1}, \theta)}$$
3. Output $\pi(\theta, x_{t+1} | \mathcal{D}_{t+1})$, put $t := t + 1$, return to step 2.
Advantages of the sequential algorithm

- In the presence of measurement error, the sequential likelihood-free scheme is effective, and is much simpler than a more efficient MCMC approach.
- The likelihood-free approach is easier to tailor to non-standard models and data, and synthesis of data from multiple distinct models with many shared parameters.
- The essential problem is that of calibration of complex stochastic computer models.
- For slow stochastic models, there is considerable interest in developing fast emulators and embedding these into MCMC algorithms (as millions of forward-simulations from the model will typically be required).
Building emulators for slow simulators

- Use **Gaussian process regression** to build an emulator of a slow deterministic simulator
- Obtain runs on a carefully constructed set of design points (e.g., a Latin hypercube) — easy to exploit **parallel computing** hardware here
- For a stochastic simulator, many approaches are possible
  - (Mixtures of) Dirichlet processes (and related constructs) are potentially quite flexible
  - Can also model output **parametrically** (say, Gaussian), with parameters modelled by (independent) Gaussian processes
  - Will typically want more than one run per design point, in order to be able to estimate distribution
- In principle, could (and should) allow for the imperfection of the emulator, but difficult without evaluating likelihoods
Accounting for emulator inaccuracy

- Suppose that instead of forward-simulation from the “true” model, $\pi(x|\theta)$, we instead simulate from a fast stochastic emulator, $\tilde{\pi}(x|\theta)$

- We can account for this in the MCMC acceptance ratio, which now becomes

$$A = \frac{\pi(D|x^*, \theta^*)}{\pi(D|x, \theta)} \times \frac{\pi(x^*|\theta^*)}{\tilde{\pi}(x^*|\theta^*)} \times \frac{\tilde{\pi}(x|\theta)}{\pi(x|\theta)}$$

- Note that the awkward likelihood term no longer drops out. Evaluating the likelihoods is (at least) as complex as forwards simulation...

- Potential for modelling/emulation of the the log-likelihood ratio, and plugging in — but then the acceptance ratio still won’t be perfect...
Limitations and alternatives

- The LF-MCMC approach is simple and very general, but not without problems
- Scales badly as number of unknowns to be estimated increases (curse of dimensionality) and as the number of sequential steps increases (particle degeneracy)
- Can we construct global and general MCMC algorithms with better stability and numerical efficiency?
  - **Particle MCMC** one possibility — perhaps more stable, if not computationally cheaper (Andreiu et al, 2010)
  - **Diffusion approximation** — approximate, but can then use techniques of inference for SDE models (Golightly and W, 2005, 2006, 2008, 2010)
  - **Approximate MCMC** exploiting (semi-)tractable stochastic emulators (Henderson et al, 2010)
  - **(Sequential) ABC** also possible — but requires approximations, also not cheap, and not trivial to automate (Toni et al, 2009)
CaliBayes

- **CaliBayes** is a generic software framework for Bayesian calibration of stochastic kinetic (or deterministic) models (encoded in SBML) using time course data.
- Implements **likelihood-free sequential MCMC** based on forwards-simulation using third-party simulators (such as FERN, COPASI or BASIS).
- Can use exact stochastic simulation, or hybrid, or diffusion — anything supported by the simulator.
- **CaliBayesR** — R client software for accessing the CaliBayes system directly from R.
- You provide the model, data, and (a sample from) the prior and it returns (a sample from) the posterior.

www.calibayes.ncl.ac.uk
Simulation study

- 3 “unknowns”
- True values:

$$k_{SigDprod} = 1, \quad k_{flacherep} = 0.02, \quad k_{flacheunrep} = 0.1$$

These correspond to the maximal rate of production of SigD, and the binding and unbinding of CodY to the fla/che operon.

- Prior distributions:

$$\log(k_{SigDprod}) \sim \text{Unif}(\log\{0.01\}, \log\{100\})$$

$$\log(k_{flacherep}) \sim \text{Unif}(\log\{0.0002\}, \log\{2\})$$

$$\log(k_{flacheunrep}) \sim \text{Unif}(\log\{0.001\}, \log\{10\})$$

These priors cover two orders of magnitude either side of the true value, and hence represent very vague prior knowledge.
Observation of $\sigma^D$

- Observe levels of $\sigma^D$ every 5 minutes for 2 hours (24 observations) subject to low measurement error (standard deviation 10 molecules)
Observation of $\sigma^D$

Contours of bivariate marginal posterior

Posterior predictive intervals

Number of molecules

Time
Observation of $Hag$

- $\sigma_D$ turns out to be difficult to measure directly — what if we could measure $Hag$ instead?

![Density plots](Darren Wilkinson — UCM 2010, 13/7/2010)
Observation of Hag

Contours of bivariate marginal posterior

Posterior predictive intervals

Number of molecules

Time

Darren Wilkinson — UCM 2010, 13/7/2010
Uncertainty in systems biology models
GFP reporters

- In fact, it is not straightforward to measure any native protein directly — instead reporters are used
- Example: Green fluorescent protein — GFP
- The gene for GFP, gfp has to be integrated into the host genome in such a way as to try and make the levels of mature GFP correlate strongly with the levels of the target protein of interest — not straightforward...
- For $\sigma^D$, standard approach is to form a fusion of the hag promoter, $P_{hag}$ to gfp to get $P_{hag}$–gfp and then integrate this construct at a convenient location in the genome, which is often at amyE
- Genotype of resulting mutant written: $amyE::P_{hag}$–gfp
The addition species and reactions can be incorporated into the model and shows the relationship of $GFP$ to the other model species.

Composition against time

![Composition graph](attachment:image.png)
Fluorescence microscopy for $amyE::P_{hag}-gfp$
GFP data
Now consider the observation of $GFP$ subject to small measurement error.
Observation of \( \textit{GFP} \)

Contour plots of bivariate marginal posterior:

- Posterior predictive intervals
  - Time
  - Number of molecules

- SigD
- Hag
- CodY

Uncertainty in systems biology models
Confusion of *GFP* with $\sigma^D$

- Now consider what happens when we observe *GFP* subject to small measurement error but enter these measurements into the inference algorithm as $\sigma^D$.

![Density plots for kSigDprod, kflacherep, and kflacheunrep](image_url)
Confusion of \textit{GFP} with $\sigma^D$

Contours of bivariate marginal posterior

Posterior predictive intervals

Number of molecules

Time

Time

Darren Wilkinson — UCM 2010, 13/7/2010
Population data and knockout variants

- The inference techniques presented are extremely flexible due to their sequential nature.
- Even in the scenario of perfect models and well-calibrated experimental data (unrealistic!), we get limited information about parameters from a single time course.
- Data from multiple cells can be incorporated by taking the posterior from one cell as the prior for the next.
- Knockout variants and different reporters can be handled similarly — relevant aspects of the posterior from one mutant model can be fed in as the prior for another.
- Similar Bayesian methods can be used for single cell measurements on cell populations, such as that obtained from flow cytometry techniques.
Layers of uncertainty

- This talk has focused mainly on two layers of uncertainty:
  - *intrinsic noise* in the biological system
  - complex stochastic measurement process
- Other important sources — including *extrinsic noise*
  - Genuine cell–cell variability/heterogeneity
  - Uncertain initial conditions
  - Model inadequacy - including *noisy, time-varying parameters*, capturing the effect of unmodelled processes on modelled processes
Time-varying parameters

- Consider a particular reaction rate, \( c_i \) — this is the hazard of a type \( i \) reaction
- We imagine that there is a “true value” for this
- This concept can be made reasonable precise \textit{in vitro}, but we are only interested in the effective value \textit{in vivo}, and this is rather more subtle
- Effective rate likely to be moderated by various (unmodelled) cellular processes
- Can model rate \( c_i \) with a time-varying stochastic process \( c_i(t) \) — say, a geometric Gaussian OU process — and take (say) the stationary mean (or stationary \textit{distribution}) to be the “true value”
Summary and conclusion

- Bacterial decision processes are noisy and complex with many layers of uncertainty — stochastic modelling can help to understand observed behaviour.
- Experimental data is often very partial, and typically needs to be modelled explicitly.
- Formal Bayesian methods for inferring parameters and unobserved processes from experimental data are challenging but just about possible.
- We need to have flexible, generic, scalable automated Bayesian inference software and computing infrastructure.
- Systems (and synthetic) biology is a new and interesting area of scientific research — more statisticians required!
Acknowledgements

- **Modelling B. subtilis decisions:** Leendert Hamoen, Jan-Willem Veening, Pete Milner
- **Likelihood-free modelling:** Richard Boys, Jeff Chen, Colin Gillespie, Daniel Henderson, Eryk Wolski, Jake Wu

This work was funded by the BBSRC through grants BBF0235451, BBSB16550 and BBC0082001.


**Contact details...**

email: d.j.wilkinson@ncl.ac.uk

www: tinyurl.com/darrenjw