Stochastic modelling and Bayesian inference for biochemical network dynamics

Darren Wilkinson
tinyurl.com/darrenjw
School of Mathematics & Statistics and
Centre for Integrated Systems Biology of Ageing & Nutrition (CISBAN)
Newcastle University, UK

Warwick EPSRC Symposium 2009
Challenges in Scientific Computing
30th June–3rd July 2009

Overview

- Stochastic modelling in systems biology
- Quick guide to stochastic chemical kinetics
- Bayesian parameter inference for stochastic kinetic models
- MCMC for the chemical Langevin equation (CLE)
- Sequential likelihood-free MCMC for Markov process models
- Inference for network structure from HTP data

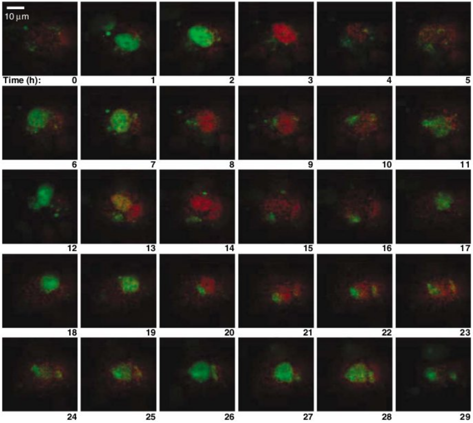
Pathways to senescence

- The mammalian group within CISBAN is interested in cell ageing and many aspects of the processes which lead to cellular senescence, and study this using immortalised human and rodent cell lines
- DNA damage and repair processes are one important component of this large and complex system, and therefore molecules involved in damage signalling and repair are of direct interest
- Considerable interest in the role of p53 ("the guardian of the genome") in this context, and the development of models for p53 regulation
- p53 has many important functions, but of most relevance to this discussion is its ability to activate DNA repair proteins in response to DNA damage

Single cell fluorescence microscopy

Loading movie...

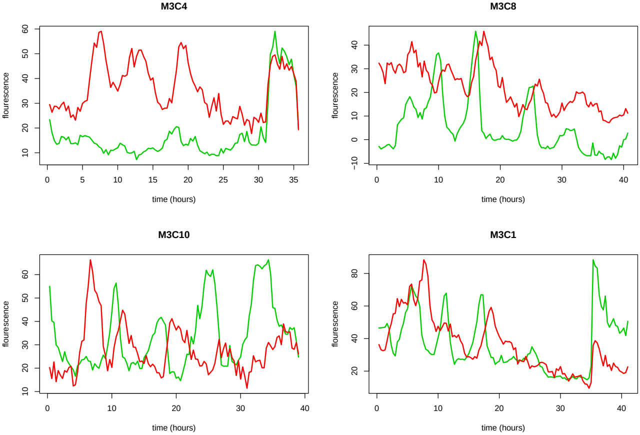
Single cell fluorescence microscopy



p53-CFP and Mdm2-YFP

p53/Mdm2 oscillations subsequent to gamma irradiation

Single cell time course data



Geva-Zatorsky et al (2006), *Mol. Sys. Bio.* [Uri Alon's lab]

Stochastic chemical kinetics

- u species: $\mathcal{X}_1, \ldots, \mathcal{X}_u$, and v reactions: $\mathcal{R}_1, \ldots, \mathcal{R}_v$
- $\mathcal{R}_i : \; p_{i1}\mathcal{X}_1 + \cdots + p_{iu}\mathcal{X}_u \longrightarrow q_{i1}\mathcal{X}_1 + \cdots + q_{iu}\mathcal{X}_u, \; i = 1, \ldots, v$
- In matrix form: $P\mathcal{X} \longrightarrow Q\mathcal{X}$ (P and Q are sparse)
- $S = (Q - P)'$ is the stoichiometry matrix of the system
- X_{jt}: # molecules of $\mathcal{X}_j$ 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,

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

Time change representation

- R_{it}: $\#$ reactions of type $\mathcal{R}_i$ in $(0,t]$, $R_t = (R_{1t}, \ldots, R_{vt})'$
- $X_t - X_0 = SR_t$ (state updating equation)
- For $i = 1, \ldots, v$, $N_i(t)$ are the count functions for independent unit Poisson processes, so

$$R_{it} = N_i\left(\int_0^t h_i(X_\tau, c_i)d\tau\right)$$

- Putting $N(t_1, \ldots, t_v) = (N_1(t_1), \ldots, N_v(t_v))'$, we can write $R_t = N\left(\int_0^t h(X_\tau, c)d\tau\right)$ to get:

Time-change representation of the Markov jump process

$$X_t - X_0 = S\,N\left(\int_0^t h(X_\tau, c)d\tau\right)$$

The Gillespie algorithm

1. Initialise the system at $t = 0$ with rate constants $c_1, c_2, \ldots, c_v$ and initial numbers of molecules for each species, $x = (x_1, x_2, \ldots, x_u)'$.
2. For each $i = 1, 2, \ldots, v$, calculate $h_i(x, c_i)$ based on the current state, x.
3. Calculate $h_0(x, c) \equiv \sum_{i=1}^{v} h_i(x, c_i)$, the combined reaction hazard.
4. Simulate time to next event, τ, as an $Exp(h_0(x, c))$ random quantity, and put $t := t + \tau$.
5. Simulate the reaction index, j, as a discrete random quantity with probabilities $h_i(x, c_i) \ / \ h_0(x, c)$, $i = 1, 2, \ldots, v$.
6. Update x according to reaction j. That is, put $x := x + S^{(j)}$, where $S^{(j)}$ denotes the jth column of the stoichiometry matrix S.
7. Output x and t.
8. If $t < T_{max}$, return to step 2.

Modelling large biological systems

BASIS — Biology of Ageing e-Science Integration and Simulation

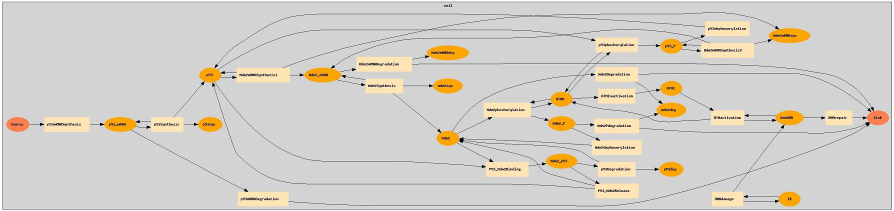
BBSRC Bioinformatics and e-Science grant, and UK e-Science GRID pilot project, now incorporated into CISBAN — `http://www.basis.ncl.ac.uk/`

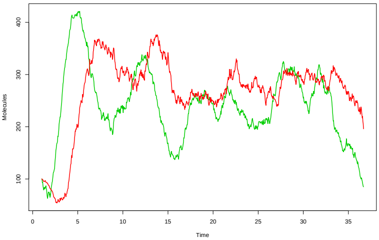
- Modelling large complex systems with many interacting components (with emphasis on biological mechanisms relating to ageing)
- SBML model database
- Discrete stochastic simulation service running on a large cluster (and a results database)
- Distributed computing infrastructure for routine use (web portal and web-service interface for GRID computing)

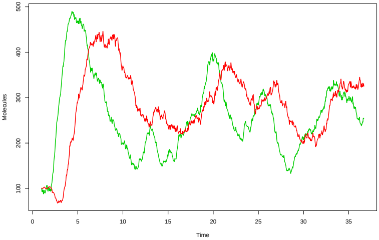
Stochastic kinetic model

- Discrete stochastic kinetic model developed at Newcastle (by Carole Proctor) for the key biomolecular interactions between p53, Mdm2 and their response to DNA damage induced by irradiation
- More complex than a simple Lotka-Volterra system (17 species and 20 reactions), but essentially the same regulatory feedback mechanism (Mdm2 synthesis depends on the level of free p53, and Mdm2 encourages degradation of p53)
- Some information about most kinetic parameters, but considerable uncertainty for several — ideal for a Bayesian analysis

Model structure and sample output







Benefits of stochastic modelling

- Several (essentially) deterministic models have been proposed
- Only a stochastic model can mimic the behaviour of single cells (observed individually, or at the level of a cell population model, using FACS) and the average behaviour of the cell population (using time-course microarrays — single peak)
- Many possible sources of heterogeneity in the cell population (though genetic differences should be minimal) - eg. cell size, cell cycle phase
- This discrete molecular-level model shows that intrinsic stochasticity in gene expression is sufficient to explain the observed heterogeneity (but does not rule out other sources), and requires no artificial modelling devices such as time-delays

Bayesian inference

Tuning model parameters so that output from the model “better matches” experimental data is a standard optimisation problem, but is problematic and unsatisfactory for a number of reasons:

- Defining an appropriate “objective function” is not straightforward if the model is stochastic or the measurement error has a complex structure (not IID Gaussian)
- The statistical concept of likelihood provides the “correct” way of measuring the evidence in favour of a set of model parameters, but typically requires computationally intensive Monte Carlo procedures for evaluation in complex settings
- Simple optimisation of the likelihood (the maximum likelihood approach) is also unsatisfactory, as there are typically many parameter combinations with very similar likelihoods (and the likelihood surface is typically multi-modal, making global optimisation difficult)

Markov chain Monte Carlo (MCMC)

- Additionally, likelihood ignores any existing information known about likely parameter values *a priori*, which can be very useful for regularising the inference problem — better to base inference on the posterior distribution
- MCMC algorithms can be used to explore plausible regions of parameter space in accordance with the posterior distribution — these provide rich information
- eg. rather than simple point estimates for parameter values, can get plausible ranges of values, together with information on parameter identifiability and confounding
- MCMC algorithms are computationally intensive, but given that evaluation of the likelihood is typically computationally intensive anyway, nothing to lose and everything to gain by doing a Bayesian analysis

Bayesian inference for stochastic models

- Bayesian inference techniques can be used to estimate the parameters of non-linear stochastic process models from data
- As well as giving insight into plausible parameter values and the extent to which these are identified by the data, they also allow one to asses the extent to which the stochastic model fits the data at all
- Ultimately, predictive quantitative statements can be made about the behaviour of individual cells and cell populations under a range of experimental conditions

Bayesian inference

- In principle it is possible to carry out rigorous Bayesian statistical inference for the parameters of stochastic kinetic models
- Fairly detailed experimental data are required — eg. 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

The chemical Langevin equation (CLE)

- The CLE is a diffusion approximation to the true Markov jump process
- Start with the time change representation

$$X_t - X_0 = S\,N\left(\int_0^t h(X_\tau, c)d\tau\right)$$

and approximate $N_i(t) \simeq t + W_i(t)$, where $W_i(t)$ is an independent Wiener process for each i

- Substituting in and using a little stochastic calculus gives:

The CLE as an Itô SDE:

$$dX_t = Sh(X_t, c)\,dt + \sqrt{S\,\mathrm{diag}\{h(X_t, c)\}S'}\,dW_t$$

MCMC-based Bayesian inference for the CLE

- Inference for a non-linear multivariate stochastic differential equation model observed partially, at discrete times and most likely with error
- This also turns out to be a rather challenging problem, due to the intractability of the discrete-time transition densities, but it is possible to develop computationally intensive MCMC algorithms that are very effective (Golightly & W, 05, 06a, 06b, 08, 09)
- However, the global MCMC algorithms (05, 08, 09) are very computationally intensive, rely on the CLE being a reasonable approximation, and are non-trivial to adapt to realistic scenarios (mutiple data sets on different species and different model variants) — sequential MCMC algorithms (06a, 06b) are more flexible, and are not limited to the CLE

MCMC-based fully Bayesian inference for *fast* computer models

- Before worrying about the issues associated with slow simulators, it is worth thinking about the issues involved in calibrating fast deterministic and stochastic simulators, based only on the ability to forward-simulate from the model
- In this case it is often possible to construct MCMC algorithms for fully Bayesian inference using the ideas of likelihood-free MCMC (Marjoram et al 2003)
- Here an MCMC scheme is developed exploiting forward simulation from the model, and this causes problematic likelihood terms to drop out of the M-H acceptance probabilities

Generic problem

- Model parameters: c
- (Stochastic) model output: $\mathbf{x}$
- (Noisy and/or partial) data: $\mathcal{D}$
- For simplicity suppose that $c \perp\!\!\!\perp \mathcal{D}|\mathbf{x}$ (but can be relaxed)
- We wish to treat the model as a "black box", which can only be forward-simulated
- We are thinking about data relating to a single realisation of the model (so no need to explicitly treat initial conditions), but replicate runs and multiple conditions can be handled sequentially (as will become clear)

MCMC-based Bayesian inference

- Target: $\pi(c|\mathcal{D})$
- Specify a “measurement error model”, $\pi(\mathcal{D}|\mathbf{x})$ — eg. just a product of Gaussian or t densities
- Generic MCMC scheme:
  - Propose $c^\star \sim f(c^\star|c)$
  - Accept with probability $\min\{1, A\}$, where

$$A = \frac{\pi(c^\star)}{\pi(c)} \times \frac{f(c|c^\star)}{f(c^\star|c)} \times \frac{\pi(\mathcal{D}|c^\star)}{\pi(\mathcal{D}|c)}$$

- $\pi(\mathcal{D}|c)$ is the “marginal likelihood” (or “observed data likelihood”, or...)

Special case: deterministic model

- Deterministic function $g(\cdot)$ such that $\mathbf{x} = g(c)$
- Then

$$\begin{aligned}\pi(\mathcal{D}|c) &= \pi(\mathcal{D}|c, g(c)) \\ &= \pi(\mathcal{D}|c, \mathbf{x}) \\ &= \pi(\mathcal{D}|\mathbf{x})\end{aligned}$$

- Here $\pi(\mathcal{D}|\mathbf{x})$ is just the “measurement error model” — eg. simple product of Gaussian or t densities
- This setup is somewhat simplistic for the deterministic case, but we are really more concerned with the stochastic case...

Stochastic model

- Can't get at the marginal likelihood directly, so make the target $\pi(c, \mathbf{x}|\mathcal{D})$, where $\mathbf{x}$ is the "true" simulator output which led to the observed data...
- Clear that we can marginalise out $\mathbf{x}$ if necessary, but typically of inferential interest anyway
- Use ideas from "likelihood-free MCMC" (Marjoram et al, 2003)
- Propose $(c^\star, \mathbf{x}^\star) \sim f(c^\star|c)\pi(\mathbf{x}^\star|c^\star)$, so that $\mathbf{x}^\star$ is a forward simulation from the (stochastic) model based on the proposed new $c^\star$

$$A = \frac{\pi(c^\star)}{\pi(c)} \times \frac{f(c|c^\star)}{f(c^\star|c)} \times \frac{\pi(\mathcal{D}|\mathbf{x}^\star)}{\pi(\mathcal{D}|\mathbf{x})}$$

“Likelihood-free” MCMC

- Again $\pi(\mathcal{D}|\mathbf{x})$ is a simple measurement error model...
- 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 $\mathbf{x}_t \equiv \{x_s | t-1 < s \leq t\}$, $t = 2, 3, \ldots, n$, so that $\mathbf{x} \equiv \{\mathbf{x}_2, \ldots, \mathbf{x}_n\}$.

1. Assume at time t we have a (large) sample from $\pi(c, x_t | \mathcal{D}_t)$ (for time 0, initialise with sample from prior)
2. Run an MCMC algorithm which constructs a proposal in two stages:
    1. First sample $(c^\star, x_t^\star) \sim \pi(c, x_t | \mathcal{D}_t)$ by picking at random and perturbing slightly (sampling from the kernel density estimate)
    2. Next sample $\mathbf{x}_{t+1}^\star$ by forward simulation from $\pi(\mathbf{x}_{t+1}^\star | c^\star, x_t^\star)$
    3. Accept/reject $(c^\star, x_{t+1}^\star)$ with

$$A = \frac{\pi(d_{t+1} | x_{t+1}^\star)}{\pi(d_{t+1} | x_{t+1})}$$

3. Output $\pi(c, x_{t+1} | \mathcal{D}_{t+1})$, put $t := t + 1$, return to step 2.

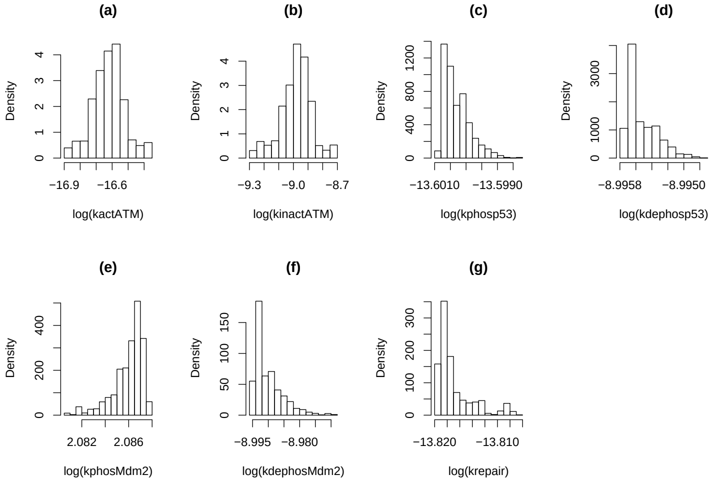
Advantages of the sequential algorithm

- 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
- 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)

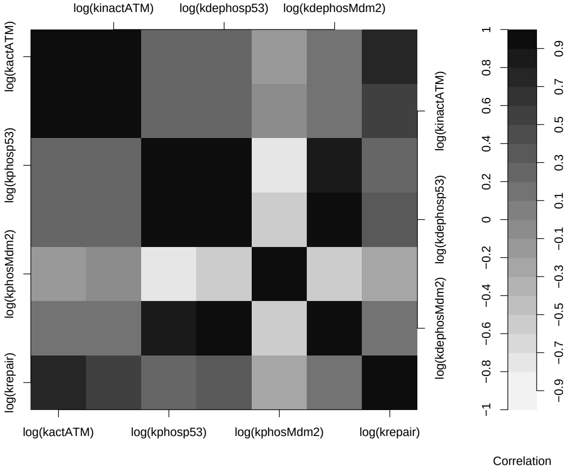
Building emulators for *slow* simulators

- Use Gaussian process regression to build an emulator of a slow deterministic simulator
- Obtain runs on a carefully constructed set of design points (eg. 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

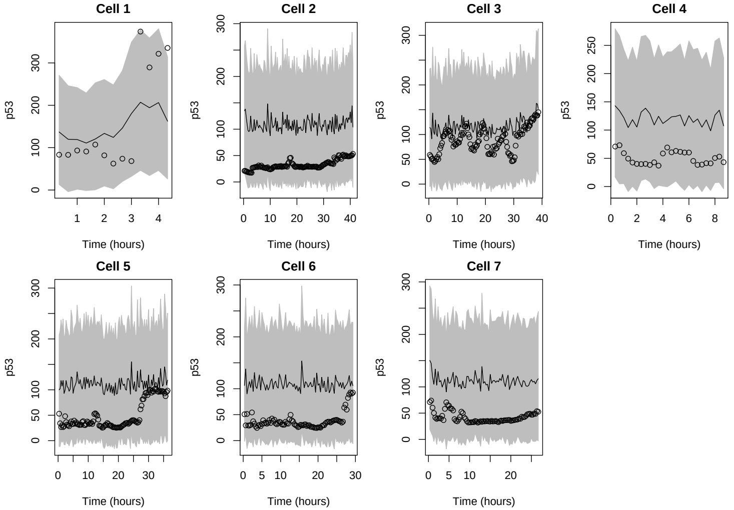
Parameter inference for the p53 model



Posterior correlations for the p53 model



Predictive fit for the p53 model



Why sequential rather than global MCMC?

- Can develop global algorithms for single time course (single cell), but need to condition on data from multiple time courses (multiple cells) and multiple model variants
- In principle this could be handled by developing a hierarchical model framework, but this will be extremely difficult and time-consuming in practice
- Alternatively, can use the sequential MCMC methods previously described — it is then easy to handle multiple cells by taking the posterior distribution from one cell as the prior distribution for the next
- Model variants (such as gene knockouts) can be handled similarly

Calibration of complex simulation models

CaliBayes — Integration of GRID-based post-genomic data resources through Bayesian calibration of biological simulators

BBSRC Bioinformatics and e-Science II project

`http://www.calibayes.ncl.ac.uk/`

- Bayesian model calibration is concerned with the problem of parameter estimation, model validation, design and analysis based only on the ability to forward simulate from the model
- It is particularly appropriate for slow and/or complex models and/or data, where likelihood-based methods are computationally infeasible
- Provides a flexible and generic framework for parameter inference problems in Systems Biology

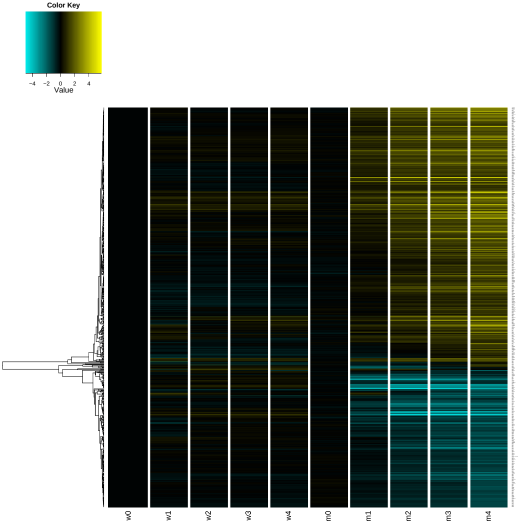
CaliBayes service-oriented architecture

- CaliBayes simulator interface — a standard SOAP web-services interface to an SBML-compliant simulator (eg. BASIS, COPASI, etc.). Could be Gillespie, Langevin, deterministic, hybrid, or an emulator...
- CaliBayes calibration engine — the main back-end computational service implementing the Bayesian sequential MCMC algorithm for model calibration based on updating a single block.
- CaliBayes data integrator — the main user-level calibration service. This service allows the calibration of a model based on multiple time series, which may consist of measurements of different species or other model components, and at different time points.

High throughput (HTP) data

- Although we would prefer to use high-resolution single-cell time course data for all of our statistical modelling, such data is difficult to obtain in a high throughput (HTP) fashion for large numbers of proteins
- We therefore wish to integrate HTP data into our modelling approach. Such data is usually of lower resolution and possessing relatively poor dynamic range, but provides (simultaneous) measurement of very large numbers of biological features
- HTP data is potentially useful for uncovering network structure

Time course microarray data



The sparse VAR(1) model approximates the CLE

- We have already seen how the true Markov jump process can be approximated by the CLE
- We can go further and linearise the CLE to get a multivariate Gaussian Ornstein-Uhlenbeck (OU) process
- This OU process can be time-discretised exactly to give a VAR(1) model with sparse auto-regressive matrix (the sparsity of this matrix derives from the sparsity of the stoichiometry matrix of the CLE)
- This suggests that the sparse VAR(1) model might be a good top-down model for inferring the underlying structure of biochemical networks from dynamic HTP data

Sparse VAR(1) model

- Observe a p-dimensional vector X_t, at each of n time points, $t = 1, \ldots, n$ (with $p >> n$)

$$X_{t+1} = \mu + A(X_t - \mu) + \epsilon_t, \quad \epsilon_t \sim N(0, V)$$

- The $p \times p$ matrix A is assumed to be sparse (ie. most elements are expected to be exactly zero)
- Sparsity can be modelled in many ways. Simplest:

$$\Pr(a_{ij} \neq 0) = \pi, \ \forall i, j, \qquad a_{ij} | a_{ij} \neq 0 \sim N(0, \sigma^2), \ \forall i, j$$

- The non-zero structure of A can be associated with a graph (network) of dynamic interactions (non-zero a_{ij} implies arc from node j to node i)

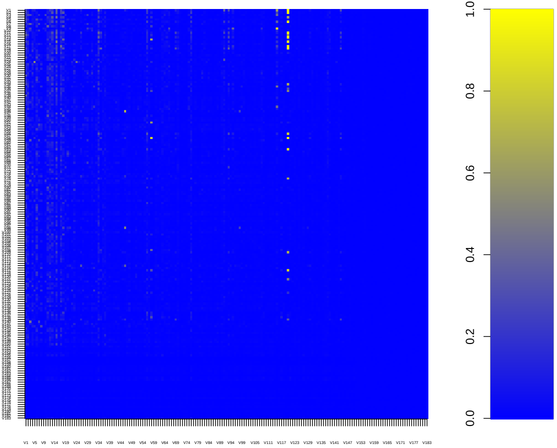
Inference for model parameters and structure from data

- Can get a point estimate for the network structure by computing a shrinkage estimate of A and then thresholding (Opgen-Rhein & Strimmer, 2007)
- Can also use Bayesian MCMC methods to explore the space of plausible interaction graphs
- MCMC methods allow computation of useful quantities such as $\Pr(a_{ij} \neq 0|\mathcal{D})$
- Inference for graphs is a hard problem...

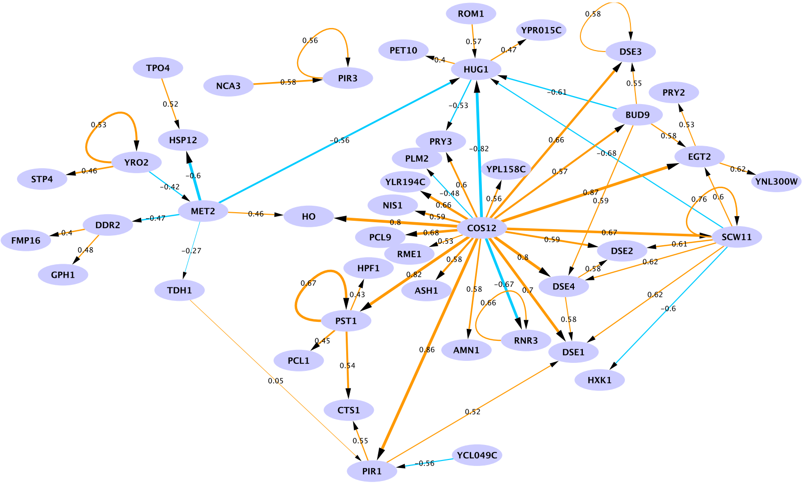
MCMC for sparse VAR(1) models

- RJ-MCMC algorithm to explore both graphical structure and model parameters (auto-regressive coefficients, mean vector, variance components) — routine to develop and implement, but exhibits poor mixing in high-dimensional settings
- Conditional on the graphical structure, possible (but messy) to develop a variational algorithm which gives an approximate marginal log-likelihood for the model after a few iterations — can embed this in a very simple MCMC algorithm to explore just the graphical structure
- Even this algorithm mixes poorly for large p (say, $p > 200$), but there are 2^{p^2} graphs, after all...
- Could probably get reasonable speed-up by using (parallel) sparse matrix algorithms

Connectivity matrix for the yeast data



Inferred network for the yeast data



Summary

- Stochastic models are useful in many areas of systems biology, due to intrinsic stochasticity of intra-cellular processes, but are especially relevant in the context of modelling damage and repair processes associated with ageing
- Fitting stochastic models to data is challenging due primarily to the difficulty of evaluating the likelihood of the data for a given parameter set
- Bayesian methods can be used for parameter estimation, and provide much richer information than other approaches
- It is possible to develop inferential algorithms which rely only on the ability to forward simulate from the model
- For slow simulation models, it can be useful to develop fast emulators of the process to be used in place of the full model

Acknowledgements

- Stochastic kinetic models: Richard Boys, Leendert Hamoen, Tom Kirkwood, Colin Gillespie, Andrew Golightly, Conor Lawless, Pete Milner, Daryl Shanley, Carole Proctor, Wan Ng, Jan-Willem Veening (funding from BBSRC, EPSRC, MRC, DTI, Unilever)
- Likelihood-free modelling: Richard Boys, Jeff Chen, Colin Gillespie, Daniel Henderson, Eryk Wolski, Jake Wu (funding from BBSRC)
- Inference for diffusions: Andrew Golightly (funding from EPSRC)
- Sparse VAR(1) modelling: Richard Boys, Colin Gillespie, Amanda Greenall, Adrian Houghton, Guiyuan Lei, David Lydall (funding from BBSRC and EPSRC)

More acknowledgements...

Funding: BBSRC, EPSRC, Newcastle University (CISBAN)

People: The work described here was contributed to by more or less everyone associated with CISBAN:

www.cisban.ac.uk

Boys, R. J., D. J. Wilkinson and T. B. L. Kirkwood (2008) Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing,* **18**(2), 125–135.

Golightly, A. and D. J. Wilkinson (2008) Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics and Data Analysis,* **52**(3), 1674–1693.

Henderson, D. A., Boys, R. J., Krishnan, K. J., Lawless, C., Wilkinson, D. J. (2009) Bayesian emulation and calibration of a stochastic computer model of mitochondrial DNA deletions in substantia nigra neurons, *Journal of the American Statistical Association.* **104**(485):76-87.

Henderson, D. A., Boys, R. J., Proctor, C. J., Wilkinson, D. J. (2009) *Linking systems biology models to data: a stochastic kinetic model of p53 oscillations*, A. O'Hagan and M. West (eds.) Handbook of Applied Bayesian Analysis, Oxford University Press, in press.

Wilkinson, D. J. (2009) Stochastic modelling for quantitative description of heterogeneous biological systems, *Nature Reviews Genetics.* **10**(2):122-133.

Wilkinson, D. J. (2006) *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC Press.

tinyurl.com/darrenjw