

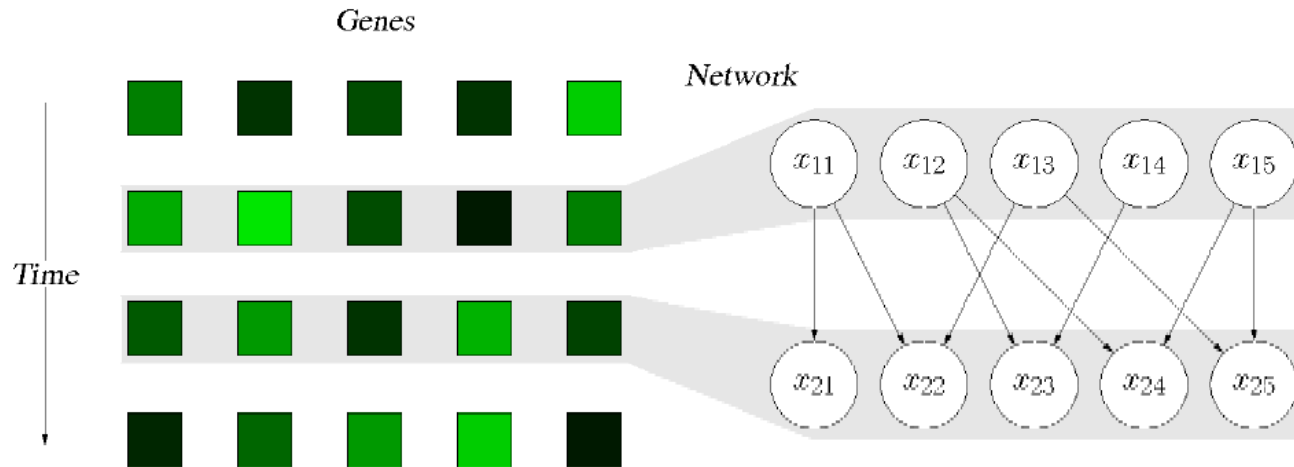
Bayesian Learning of Genetic Network Structure in the Space of Differential Equation Models

Daniel Peavoy & Ata Kaban
School of Computer Science
The University of Birmingham

Intro

- Reverse-engineering gene regulatory networks from expressions of thousand genes over time
 - Numerous approaches (biology, statistics, machine learning)
 - Difficulty lies with large number of variable & unknowns
 - Simplification of the true complexity is inevitable
 - Gain understanding of the mechanisms of the system by a more detailed description of dynamics [a feasibility study]

Graphical models (Bayes nets)



$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1, x_3)p(x_3|x_2, x_4)p(x_4|x_2, x_5)p(x_5|x_3)$$

Nodes: random variables (genes or proteins)

Edges: conditional probabilities

Overall model: Joint density

Drawbacks:

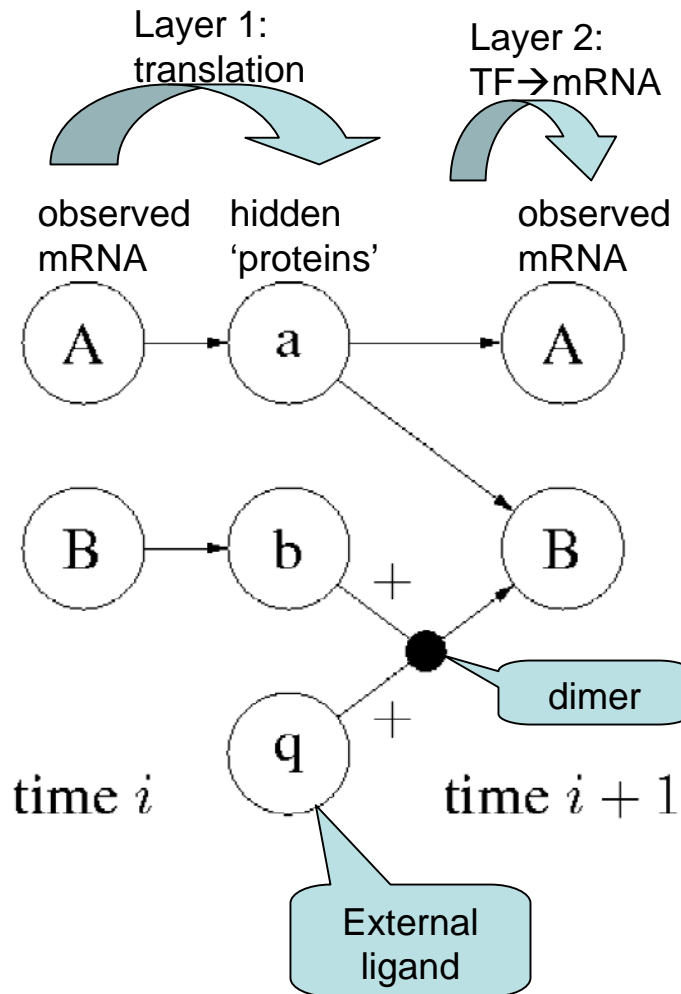
- need to specify the form of all these conditional distributions

Gaussian? – restricted to linear relations

Multinomial? – data discretisation problematic

- structure learning is hard

Graphical representation for our modelling



Nodes:

genes (capital letters)
proteins (lower case)

Edges: reactions, modelled as ordinary differential equations

Overall model: coupled ODEs
- unknown structure
- unknown parameters of constituent ODEs

Task: infer structure (& parameters) from data

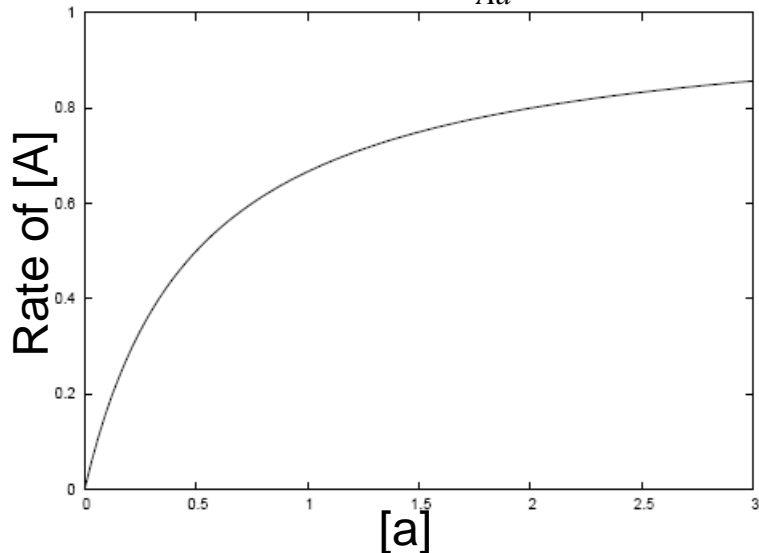
Synthetic data = simulated from such a model, with superimposed additive noise.

Nonlinear ODE building blocks

Basic building block:

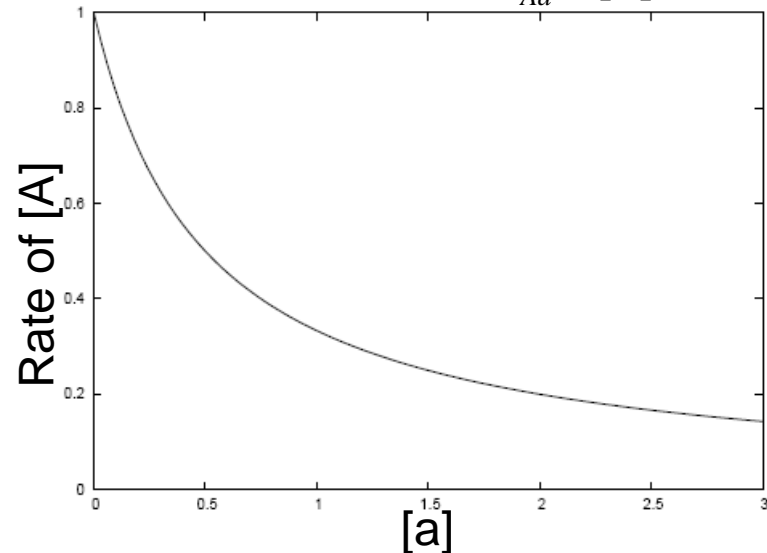
Michaelis-Menten eq, from the biological literature

$$\frac{d[A]}{dt} = V_{Aa} \frac{[a]}{K_{Aa} + [a]}$$



Promotory response curve.

$$\frac{d[A]}{dt} = V_{Aa} \left(1 - \frac{[a]}{K_{Aa} + [a]}\right)$$



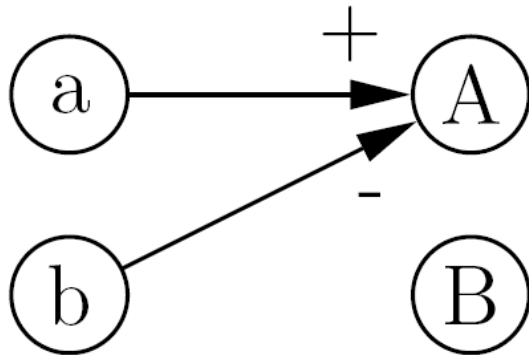
Inhibitory response curve.

where [.] denotes 'concentration of'

V. and K. are parameters (governing boundedness)

- By combining M-M rate equations, one can build more complex dynamics.

E.g.



Promoter a and inhibitor b affecting mRNA A.

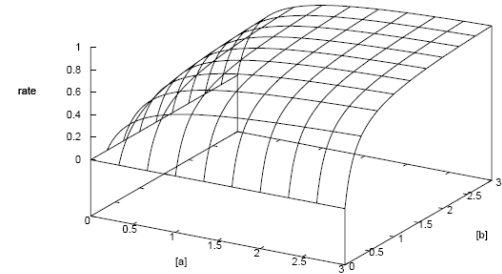
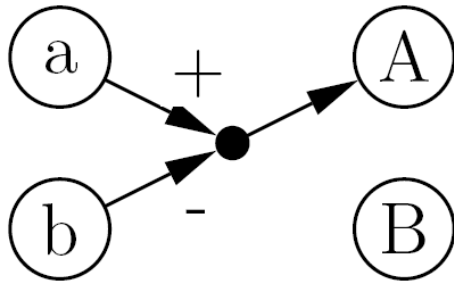
$$\frac{d[A]}{dt} = \underbrace{\frac{V_{Aa}[a]}{K_{Aa} + [a]}}_{\text{promoter's contribution}} + \underbrace{V_{Ab} \left(1 - \frac{[b]}{K_{Ab} + [b]} \right)}_{\text{inhibitor's contribution}}$$

$-h_A[A]$

decay term

- Dimer formation between proteins before acting as transcription factors for the next stage of gene expression

– Promotory dimer:



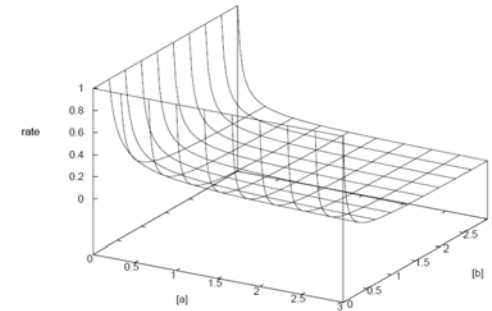
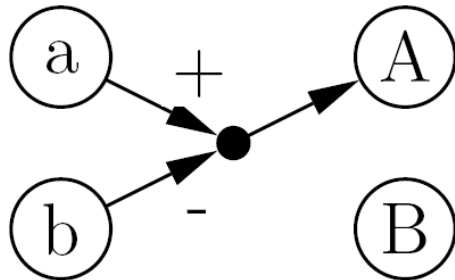
Response curve from promotory dimer ab .

$$[ab] = \frac{V_{ab}[a][b]}{K_{ab} + [a] + [b]} ; \quad \frac{d[A]}{dt} = \frac{V_{Aab}V_{ab} \frac{[a][b]}{K_{ab} + [a] + [b]}}{K_{Aab} + \frac{V_{ab}[a][b]}{K_{ab} + [a] + [b]}}$$

$$= \frac{V_{Aab}[a][b]}{K_{Aab}K_{ab} + K_{Aab}[a] + K_{Aab}[b] + [a][b]}$$

V_{ab} gets absorbed in the other parameters

– Inhibitory dimer:



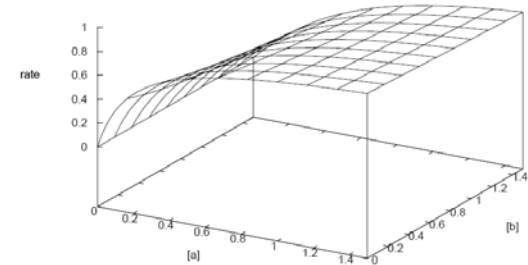
Response curve from inhibitory dimer ab .

$$\frac{d[A]}{dt} = V_{Aab} \left(1 - \frac{\frac{V_{ab}[a][b]}{K_{ab}+[a]+[b]}}{K_{Aab} + \frac{V_{ab}[a][b]}{K_{ab}+[a]+[b]}} \right)$$

$$= V_{Aab} \left(\frac{K_{Aab}K_{ab} + K_{Aab}[a] + K_{Aab}[b]}{K_{Aab}K_{ab} + K_{Aab}[a] + K_{Aab}[b] + [a][b]} \right)$$

V_{ab} gets absorbed in the other parameters

– Promotor inhibited by dimerisation



Response curve from formation of dimer ab inhibiting protein a .

$$\begin{aligned} \frac{d[A]}{dt} &= \frac{V_{Aa}[a] \left(1 - \frac{[b]}{K_{ab} + [a] + [b]}\right)}{K_{Aa} + [a] \left(1 - \frac{[b]}{K_{ab} + [a] + [b]}\right)} \\ &= \frac{V_{Aab}[a](K_{ab} + [a])}{K_{Aab}K_{ab} + K_{Aab}[a] + K_{Aab}[b] + K_{ab}[a] + [a]^2} \end{aligned}$$

Summing up the 'affector' types considered for our inference of model-combination

Num	Affector type
0	No influence
1	Promotor
2	Inhibitor
3	Promotory dimer
4	Inhibitory dimer
5	Promotor inhibited by dimerisation
6	The reverse of 5

- In layer 1 (mRNA \rightarrow protein): 1.
- In layer 2 (protein \rightarrow mRNA): 0-6.

Bayesian framework for model inference

- Conditional data likelihood (given a model and parameters):

$$p(D | M, \theta) = \prod_{i=1}^{\#measurements} \prod_{j=1}^{\#genes} N(D_{ij} | simulation(M, \theta, time_i), \sigma^2)$$

- Parameters $\theta = \{init\ conditions, params\ of\ the\ ODEs\}$
- Generated noisy data from a model with 9 genes & 11 proteins, for validating the proposed inference procedure
- Defined a model space for search/inference, with 9 genes & 15 proteins, and pre-defined that at most 4 proteins are allowed to react with a gene.
- Inference of the model (structure)
 - asserted ‘complexity prior’ on models
 - Metropolis-Hastings sampling to generate new candidate models
- Parameter inference needed to evaluate candidate models acceptance probability
 - Used Gamma(1,3) prior on all parameters
 - Metropolis sampling to obtain parameter posteriors
 - Posterior Harmonic Mean Estimator to approximate the marginal likelihood

Model sampling

- Proposal distribution $p(M2|M1)$ for MH sampling of models = Uniform over models situated in the 'neighbourhood' of $M1$

[Table of possible changes defining the neighbourhood of a model] →

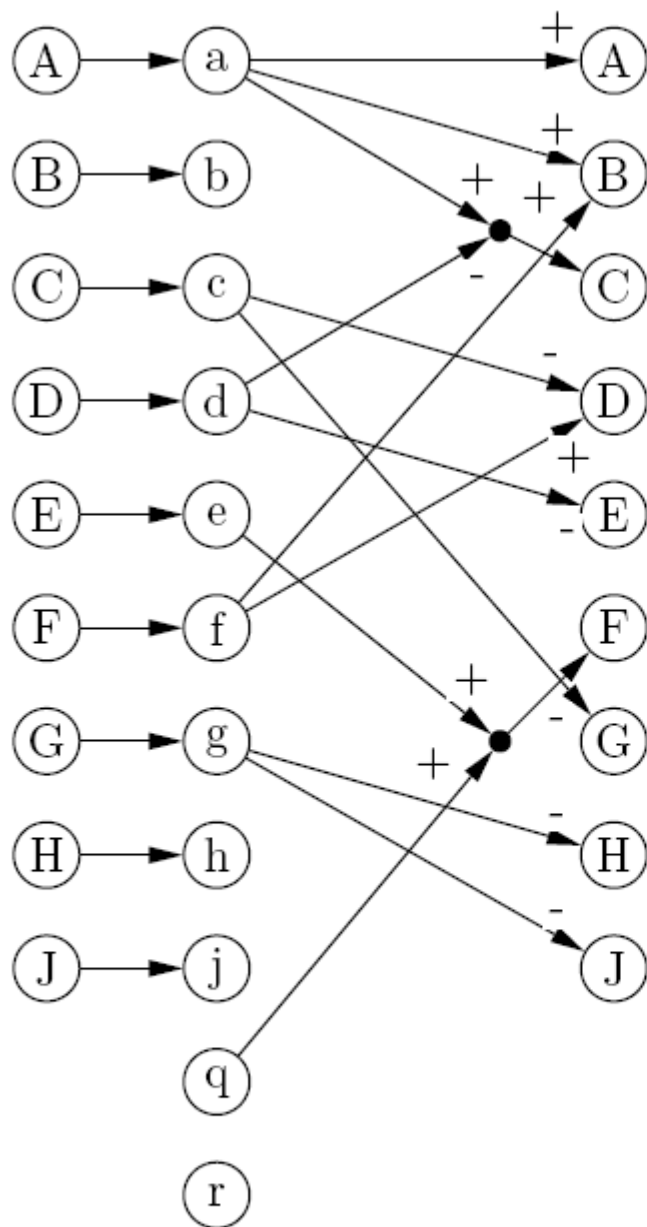
- $M2$ accepted with probability:

$$\alpha(\mathcal{M}_2, \mathcal{M}_1) = \frac{p(\mathcal{M}_2|\mathcal{D}) p(\mathcal{M}_1|\mathcal{M}_2)}{p(\mathcal{M}_1|\mathcal{D}) p(\mathcal{M}_2|\mathcal{M}_1)}$$

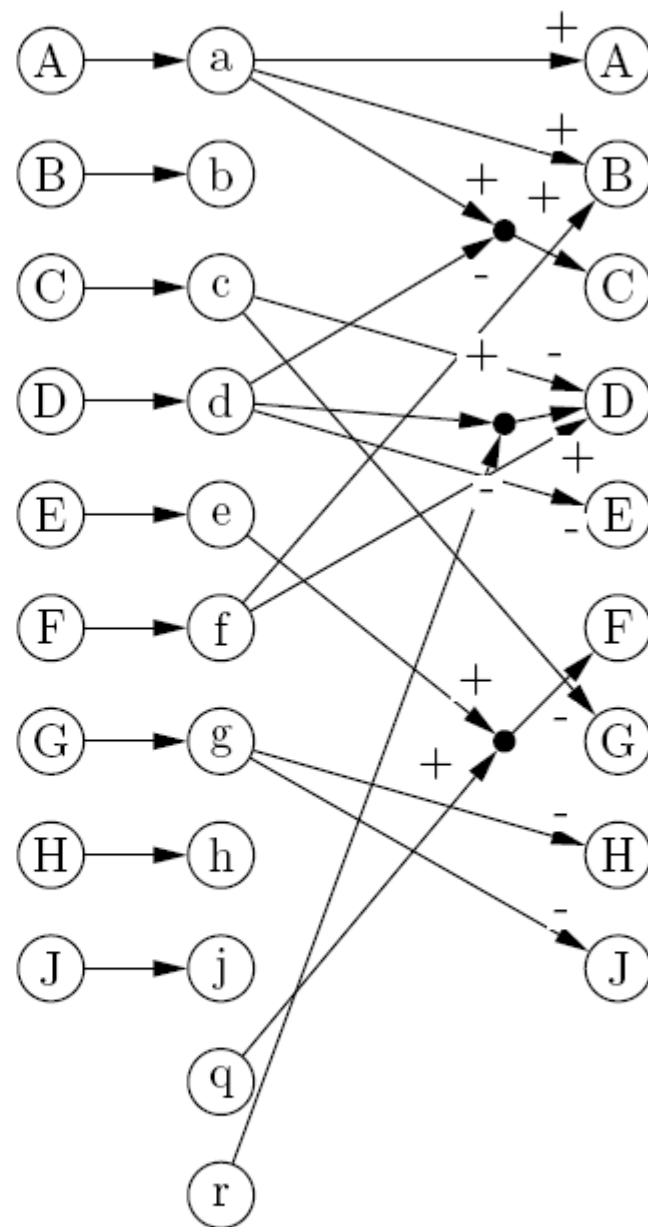
- First 100 models discarded as 'burn-in', next 500 retained.

- [50 such MH samplers were seeded with different initial models each, to cover the model space]

Change of an affector	Description
0→1	
0→2	
⋮	Changes to single edges
2→0	
2→1	
0→3	Single edges changing
⋮	to dimer forming pair
0→6	with a non-influencing
⋮	node (Num = 0)
2→6	
3→4	Dimer forming pair
⋮	changing it's action
5→6	(e.g going from a promotor to inhibitor)
3→3	Dimer node selecting
5→5	new partner node.
3→0	Pairs splitting
4→1	to become independent



(a) Example parent model for a network with nine genes. Ten hidden variables participate although eleven are shown.



(b) Neighbouring network generated randomly showing the addition of an extra ligand and dimer.

Evaluating a model's acceptance probability by parameter inference

$$\alpha(\mathcal{M}_2, \mathcal{M}_1) = \frac{p(\mathcal{M}_2 | \mathcal{D}) p(\mathcal{M}_1 | \mathcal{M}_2)}{p(\mathcal{M}_1 | \mathcal{D}) p(\mathcal{M}_2 | \mathcal{M}_1)}$$

ratio of model posteriors

model priors: $P(M_i) \propto \frac{1}{\#edges(M_i)}$

$$\frac{p(M_2 | D)}{p(M_1 | D)} = \frac{p(M_2)}{p(M_1)} \frac{p(D | M_2)}{p(D | M_1)}$$

Bayes factor (=ratio of marginal likelihoods)

conditional data likelihood:

$$= \prod_{i=1}^{\#measurements} \prod_{j=1}^{\#genes} N(D_{ij} | simulation(M_2, \theta_2, time_i), \sigma^2)$$

$$\sigma^2 = 0.1$$

$$\frac{p(D | M_2)}{p(D | M_1)} = \frac{\int d\theta_2 p(D | M_2, \theta_2) p(\theta_2 | M_2)}{\int d\theta_1 p(D | M_1, \theta_1) p(\theta_1 | M_1)}$$

Parameter priors:
Gamma(1,3)

Marginal likelihoods (analytically intractable integrals) estimated using Posterior Harmonic Mean Estimator (Newton & Raftery).

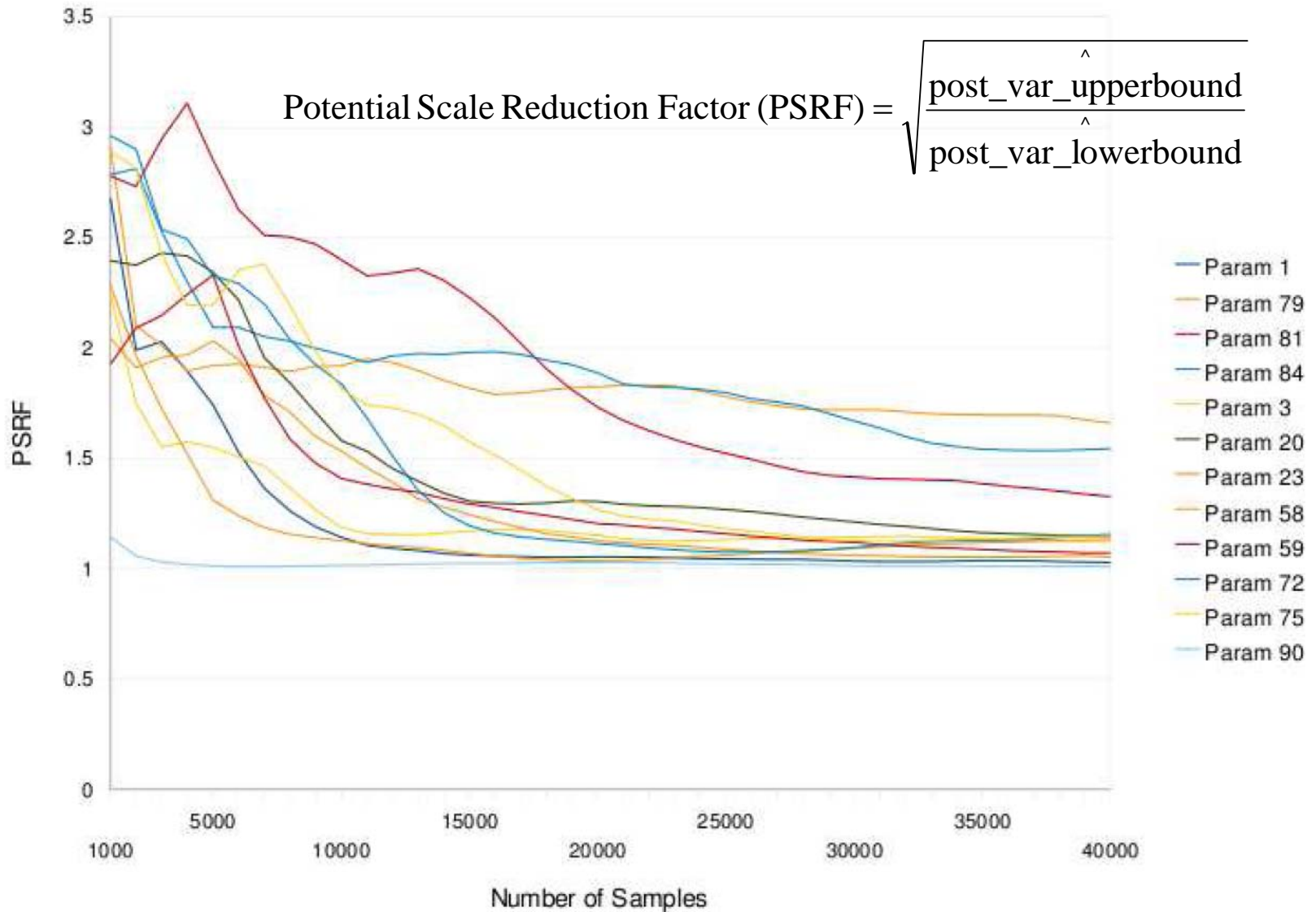
Newton & Raftery, J.Royal Stat. Soc. B, 56(1):3-48, 1994.

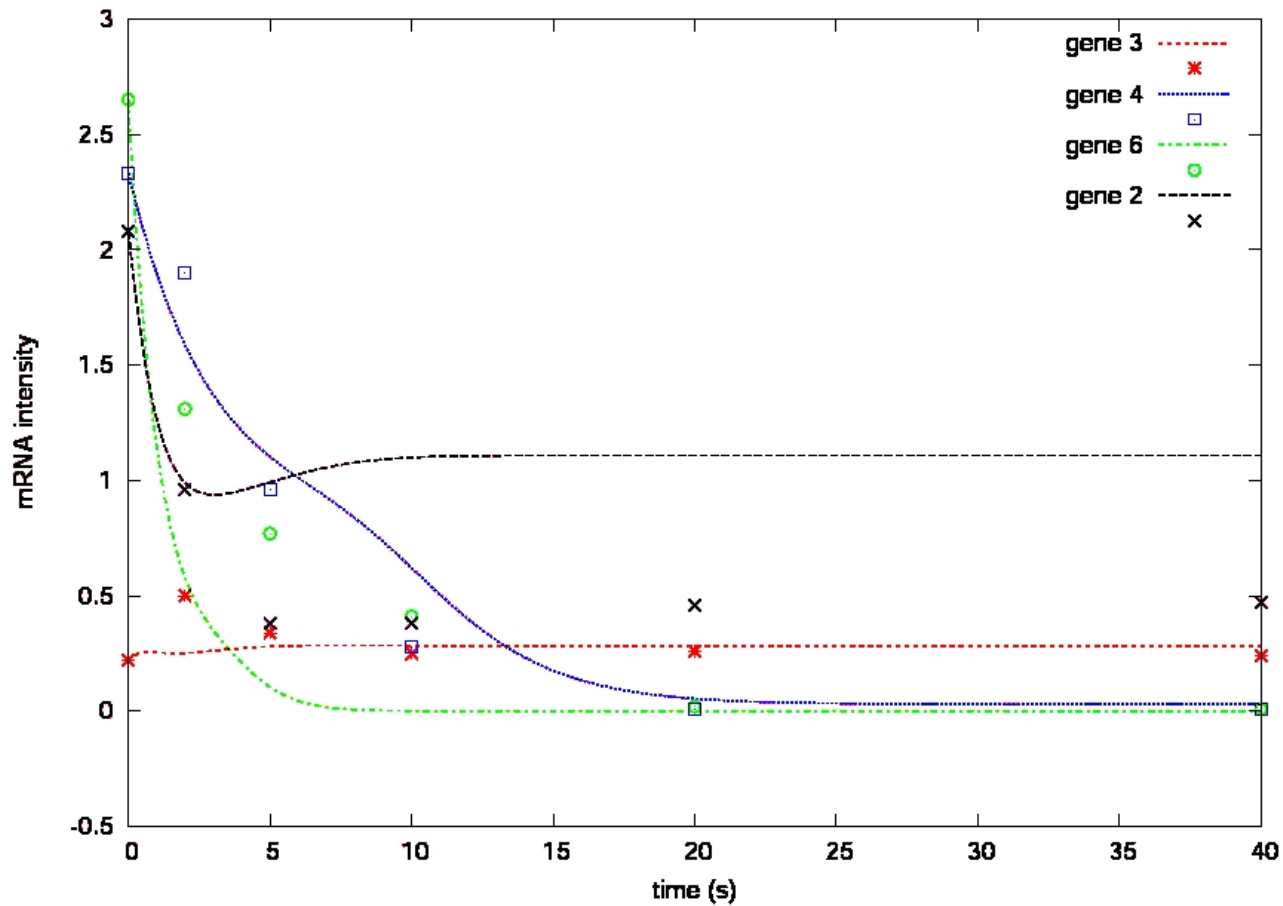
$$p(D|M) \simeq \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{p(D|M, \theta^{(i)})} \right)^{-1};$$

$\theta^{(i)} \sim p(\theta|D, M)$

parameter posteriors: estimates obtained by Metropolis sampling [10 separate Markov Chains, of 40,000 samples each]

Convergence diagnostics





Simulated dynamics from a high scoring model, with its best parameter set found (continuous lines) vs. the noisy data (marker symbols)

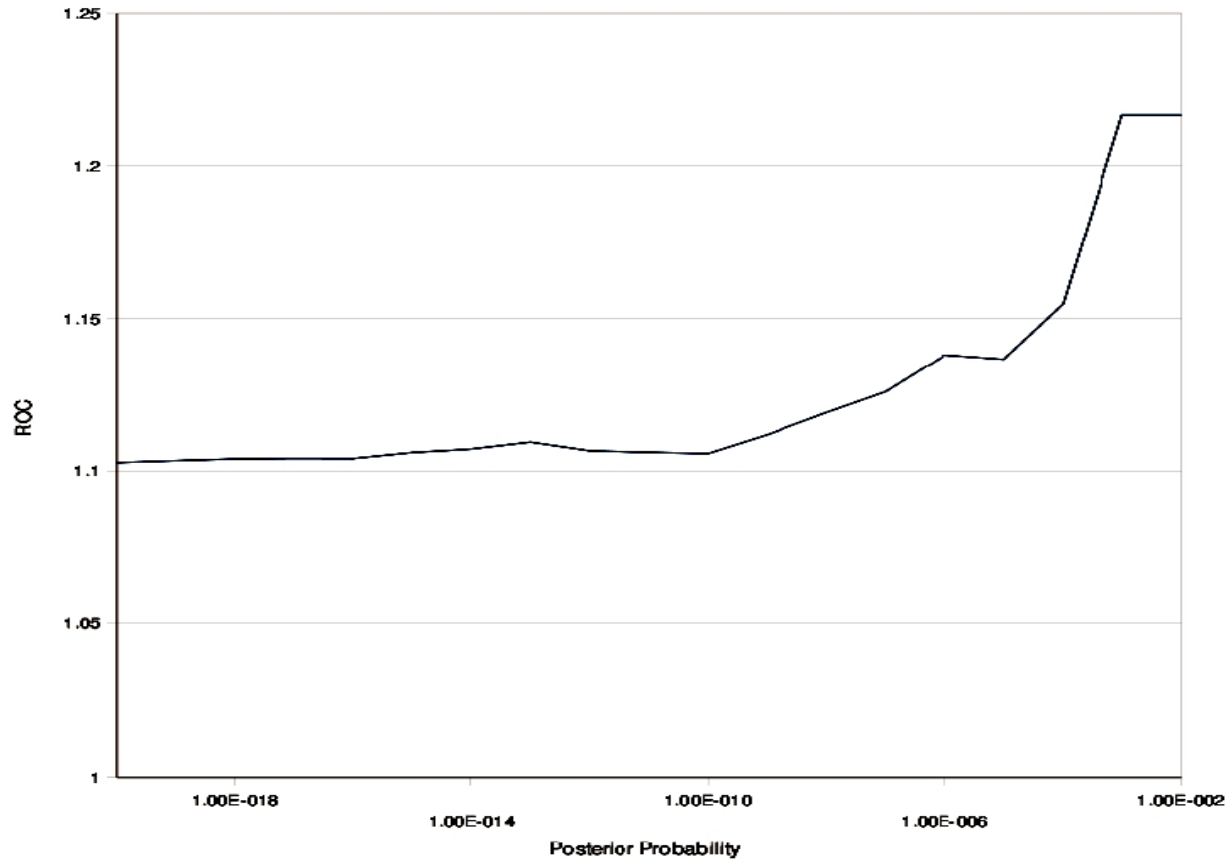
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	100	96	100	100	100	100	97	100	100	100	100	100	100	100	100	Assessment of the posterior distribution of model samples
2	21	96	86	100	97	96	100	100	100	18	18	100	21	100	83	
3	100	100	96	100	21	96	100	100	100	100	100	21	21	100	21	
4	100	86	100	8	21	100	21	86	100	100	100	21	21	21	86	
5	86	96	97	100	100	100	100	100	100	100	86	97	100	97	100	
6	96	98	86	21	100	100	97	100	100	100	99	21	100	86	100	
7	100	100	96	100	100	86	100	97	96	96	100	100	100	83	96	
8	100	100	100	4	100	100	100	100	100	83	21	100	96	96	100	
9	21	21	100	100	100	97	100	100	21	100	100	100	21	97	100	

Table 3: Mean “prediction” of edges expressed as a percentage. The rows are the genes and columns the proteins.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	Assessment of the best model (having highest posterior)
2	0	1	1	1	1	1	1	1	1	0	0	1	0	1	1	
3	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	
4	1	1	1	0	0	1	0	1	1	1	1	0	0	0	1	
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
6	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
8	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	
9	0	0	1	1	1	1	1	1	0	1	1	1	0	1	1	

Table 4: Indication of where edges from the best model agree with the original model. The rows are the genes and columns the proteins.

Structure recovery performance (measured by ROC) vs. increasing posterior probabilities. Higher posterior models correspond to better structures found.



The bottom of the graph (ROC=1) corresponds to a random model.

$$Sens = \frac{TP}{TP + FN}; \quad Spec = \frac{TN}{TN + FP}; \quad ROC = \frac{Sens}{1 - Spec}$$

Conclusions & further work

- The presented approach performs well at learning the network structure from synthetic data
- Model posteriors are in agreement with closeness to the true model structure
- This modelling was targeted towards protein-gene interactions within a cell
- It would be of interest to apply it in other contexts too
- We used some biological prior knowledge in the form of Michaelis-Menten eqs, and the prior on models favouring fewer edges
- The model space is still huge, and inserting more biological knowledge could further refine this and make the approach computationally less demanding
- At present the simulation (50 MH chains making 500 model samples each) took 5 days on a shared cluster

Related references

- H De Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.
- A Gelman and D. B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992.
- S. Imoto, T. Goto, and S. Miyano. Estimation of genetic networks and functional structures between genes by using bayesian networks and nonparametric regression. In *Pacific Symposium on Biocomputing*, pp. 175–186, 2002.
- E. Meir, E.M. Munro, G.M. Odell, and G. Von Dassow. Ingeneue: a versatile tool for reconstituting genetic networks, with examples from the segment polarity network. *Journal of Experimental Zoology*, 294:216–251, 2002.
- Vyshemirski and Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, December 2007.