

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

Daniel Peavoy and Ata Kaban

School of Computer Science, The University of Birmingham (A.Kaban@cs.bham.ac.uk)

Keywords: gene networks, differential equations, Metropolis-Hastings, Bayesian model inference and model averaging

Introduction. Reverse engineering of genetic regulatory networks (GRNs) from the expression levels of thousands of genes over time is a central problem in systems biology. A number of approaches have been developed, with effort from the biology, statistics and machine learning communities. The difficulty lies with the very large number of variables and unknowns involved. Gene interactions exhibit complex networks of intra-cellular regulation and inter cellular communication, with highly non-linear dynamics. All modelling approaches necessarily approximate the true complexity to some extent, in order to proceed with limited computational resources. However, to gain understanding of the mechanisms of a system (rather than aiming at black-box prediction only), a more detailed and more realistic description of the true dynamics is required.

Here we develop a method of approximating the gene interaction dynamics with a system of coupled, non-linear differential equations. Differential equation models were used before as a means of detailed modelling of non-linear gene interaction dynamics (De Jong, '02, Meir et al, '02, Vyshemirski and Girolami, '07, and others). These existing approaches and their implementation are able to automatically search for a good set of parameters to reproduce desired behaviour. However, they have no capability of generating the gene network *model* itself that could reproduce the data --- instead, the user must design the model first. The Bayesian model selection of (Vyshemirsky and Girolami, '07) uses Bayes factors to rank four different models in terms of their ability to generate the data. However, those four candidate models were designed manually. The next challenge, therefore, is to devise a method that allows a more general system of differential equations to be learnt simply from the given microarray data. This is what we address in this work.

Approach. The proposed method is based on using biochemical kinetics as an ansatz for the relations between genes. As it is easier to specify how a gene affects the rate of expression of another, it is natural to use differential equations, which are then solved (simulated) for a given time. We make use of software developed in (Meier et al. '02) for genetic network simulation, called Ingeneue. It is designed to model complex patterns of gene and protein interactions within and between cells. The software constructs a system of differential equations based on the nodes and interactions that the user has specified. For example, the basic process where a product AS is formed from components A and S (and where A is in excess and there are limiting amounts of S) is given by the Michaelis-Menten equation.

$$d[AS]dt = V_{AS}[A] / (K_A + [A]) \quad (1)$$

Here, $[A]$ denotes the concentration of A. This form implies that the reaction is approximately linear for low concentrations of A but gives a natural upper bound to the rate of production of AS, due to the limiting amount of S. The parameter V_{AS} controls the maximum rate and K_A , the concentration of A at half the maximum rate. Eqn. (1) could be used to model the reaction between a protein and a DNA binding site, for example, forming a transcription factor, which promotes or inhibits the expression of a nearby gene. We use (1) as a building block for the purposes of this work.

Many Bayesian sampling techniques are available for model inference. For speed and ease of implementation we have chosen to use the Metropolis-Hastings (MH) estimator to compute the marginal likelihood for each model. We address the problem of large variance by designing models so that we can include prior information, which, in practice restricts the sampled space. The models are designed to have positive priors and, since the proposed differential equations are intended to model smooth dynamics, the parameter values are likely to be small. To account for this, we placed Gamma priors on all parameters (mean 3; shape 1). The samples are generated using the Metropolis algorithm with normal proposal distribution of variance 0.1.

The models we use are constructed from components of the Ingeneue program called 'affectors'. In order to infer the model structure, we use complexity-based prior distributions assigning probabilities that these affectors be included in a model. Given a model $M1$, another is proposed $M2$ and the MH acceptance ratio is computed. The MH acceptance probability (Hastings, '70), for sampling models, is:

$$\alpha(M2, M1) = [p(M2|D)p(M1|M2)] / [p(M1|D)p(M2|M1)] \quad (2)$$

The new model is then accepted with probability α . The accepted models can be further assessed on unseen data until the resulting posterior multinomial distribution becomes more sharply "peaked" at a fewer number of models.

To sample models, one must have a well-defined model space and a proposal distribution to generate previously unseen models. One can represent a system of differential equations in a graphical form: The first layer represents the observed

mRNA intensities at time i . The middle layer consists of 'hidden nodes' such as proteins. The edges between these two layers represent the translation process. We have also allowed for external ligands, which are proteins not originating from that network. The final layer includes the observed mRNA expressions at time $i+1$, and the edges to these are the defining part of our model that we wish to infer -- they represent the way in which proteins react with each other, forming intermediary dimers and catalysing other reactions before serving as transcription factors, influencing the expression levels at time $i + 1$. The transcription process is included in the equations corresponding to this layer.

Results. To assess our methodology, we generated and used noisy data from a model with 11 proteins. For the main computation, i.e. to infer the model structure, 50 Metropolis-Hastings (MH) samplers were seeded with different initial models to cover the large space of models. For each model, the marginal likelihood was estimated with 5000 samples. The prior used was intended to limit model complexity (and hence overfitting of data) being equal to the inverse of the number of edges in the model. Models were generated, then accepted or rejected according to (2). The first 100 models in each chain were rejected and 500 retained. The parameters of the differential equations were also inferred by Markov chain sampling (10 separate Markov chains, of 40000 length each), and we used the method of (Gelman and Rubin, '92) to assess how many iterations before the MC converges to the posterior distribution. This whole simulation was run on a cluster using Sun Grid Engine.

Illustration 1 shows the simulated dynamics of the best parameter set found versus the noisy data. A good fit can be observed. More interesting is to assess the performance of inferring the correct model structure. The Receiver Operating Characteristic (ROC) is an appropriate measure that we employ for this purpose. Illustration 2 shows both the ROC values and the model posteriors plotted against each other. We see the ROC value increases almost monotonically as higher probability models are included in the calculation. This confirms our priors are reasonable and we are able to correctly identify better and better model structures by means of models that have higher posterior, in a fully automated manner. Future work is intended to include more biological domain knowledge, which is expected to achieve further improvements and also computational gains.

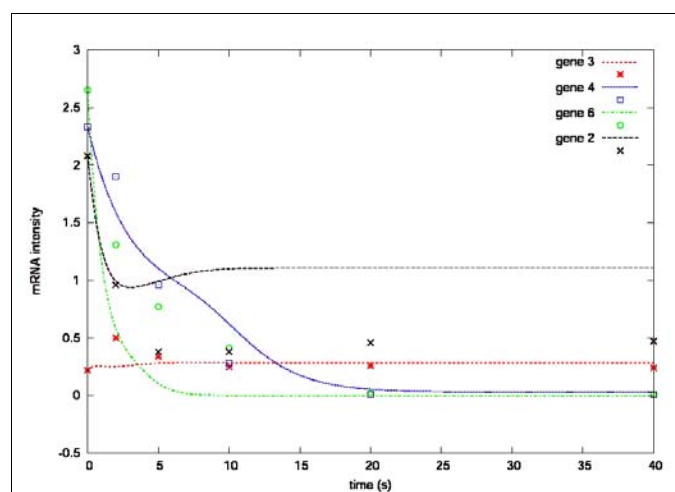


Illustration 1: The continuous lines represent the simulated dynamics of the best parameter set found. The discrete points are the noisy data.

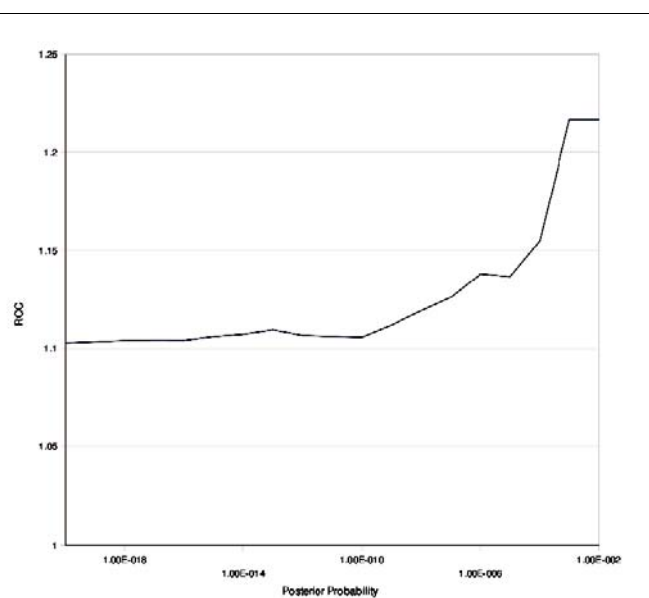


Illustration 2: ROC values plotted against increasing posterior probabilities (logarithmic scale). The bottom of the graph would correspond to a random model.

Acknowledgements. DP was funded by an EPSRC CTA studentship for MSc in Natural Computation. AK has been supported by an MRC Discipline Hopping Award (G0701858).

References

- H De Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.
- W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57 (1):97–109, 1970.
- A Gelman and D. B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992.
- 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.
- Vysheirski and Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, December 2007.