Using Differential Equations to Model Genetic Networks: A Bayesian Approach

Daniel Peavoy

School of Computer Science, University of Birmingham, msc79dxp@cs.bham.ac.uk

The Problem

DNA microarrays enable one to measure the levels of expression for thousands of genes simultaneously. Each microscopic spot corresponds to a single gene in a particular environmental condition (see Figure 1). Imaging technology first detects the level of fluorescence at each spot and converts this to a concentration of mRNA, indicating how active the corresponding gene is. Using this data the aim is to infer the underlying network of interactions that control the behaviour of a cell and ultimately answer important medical questions. For example; how does the process of cell differentiation work to produce the large variety of different tissues in the human body? All cells contain the same genes, but why are only some "switched on"?

There have been several approaches to inferring, or reverse-engineering, genetic networks, with effort from the biology, statistics and machine learning communities, the intersection of which forms part of bioinformatics research. We focus on time series microarray data, where the samples are at different time points. The problem is how to infer the causal relations between genes, as in Figure 4, and the dynamical system controlling gene expression. The problem is difficult because there is usually many more genes than there are time points, the data contains much noise (both biological and measurement) and the reactions between genes are very complicated, highly nonlinear and involve many intermediary factors such as the proteins.

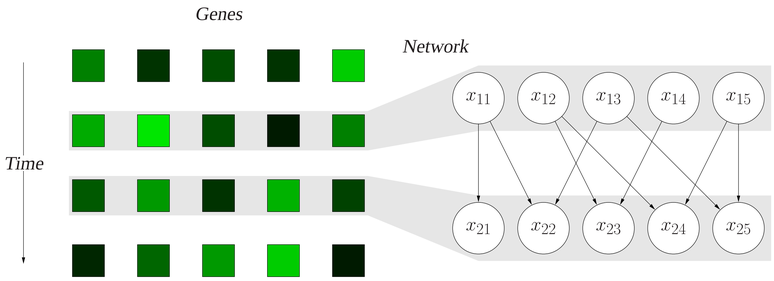

Figure 4: Representation of the gene network inference problem. On the left there are simply expression levels for many genes varying over time with complex noise processes. The right hand side shows the intended result with causal relations between genes.

Bayesian Networks

Bayesian Networks (BNs) are tools designed to express the (conditional) dependence/independence relations between sets of variables. They are used to write a full joint distribution as products of conditionals, which can be interpreted as causal relations [1]. BNs have been applied to the gene network inference problem [1][2] but are limited because they can't express cyclic relations such as in $x_{11} \rightarrow x_{12}$ in Figure 4. Kim et al. [3] used Dynamic Bayesian Networks (DBNs), which are BNs "unfolded" in time to look like Figure 4 and can encorporate cycles. As a DBN, the graph in Figure 4 corresponds to the joint distribution

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1, x_3)p(x_3|x_2, x_4)p(x_4|x_2, x_5)p(x_5|x_3)\,. \tag{5}$$

After specifying the graph one must decide upon the form of the conditional distributions. Rangel et al. [5] used a linear Gaussian model, where the mean is a linear function of the parent value and Husmeier [1] discretised the data and used a multinomial distribution over values such as "not expressed" or "largely expressed".

Both approaches have associated problems. The linear Gaussian approach can not model the complicated nonlinear reactions between genes and, while the multinomial distribution potentially can, it suffers from large information loss upon discretising the data. Imoto et al. [2] address these problems by using b-spline functions for the conditionals. Their model is therefore continuous and nonlinear but there are a huge number of parameters, which do not have any direct meaning; it is a nonparametric approach.

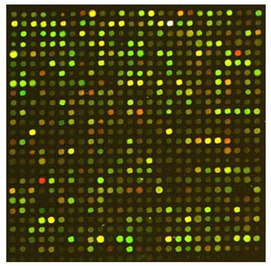

Figure 1: A DNA microarray image displaying the expression levels of many genes. Also known as a gene chip, it potentially permits the inference of the causal networks controlling cell behaviour.

Nonlinear Differential Equations

In our work we tried to use more informative parameters by considering the general form of interations between genes and allowing for a number of proteins as hidden variables. We used the Michaelis-Menton equation, from the biological literature, to model the rate of translation of a protein as a function of gene concentration

$$\frac{d[a]}{dt} = \frac{V_a[A]}{K_a + [A]}\,, \tag{1}$$

where $[a]$ is the protein concentration and $[A]$ the gene concentration. This form implies that the relation is almost linear, except with a smooth upper bound governed by parameters V_a and K_a. We also model the protein to gene dynamics by allowing proteins to firstly react with each other (forming dimers) before serving as transcription factors for the next stage of gene expression. By combining rate equations like (1) one can build more complex dynamics [4]. For example, the equation for gene A whose rate of expression is governed by two proteins a and b, forming an intemediary dimer, is

$$\frac{d[A]}{dt} = \frac{V_{Aab}[a][b]}{K_{Aab}K_{ab} + K_{Aab}[a] + K_{Aab}[b] + [a][b]} - h_A[A]\,. \tag{2}$$

The second term governs the natural decay of gene A. The first term is shown graphically in Figure 2. Notice how there is a maximum rate as expected for a real system with limited reacting sites.

Using more structured nonlinear relations we create a model for the whole network as a system of coupled ordinary differential equations. As with Bayesian networks we first define a graph except with a more restricted structure. The first layer corresponds to genes at time i, then there are a layer of hidden protein variables followed by the genes at time $i+1$. Figure 3 shows a potential graph for two genes. The proteins can combine in a number of ways to form transcription factors. The black node with two plus signs represents two proteins forming a promotory dimer, therefore, including Eq. (2). Linear combinations of terms can govern the rates, as for gene B, and proteins can react with a number of external ligands, e.g q.

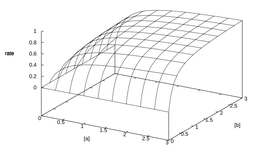

Figure 2: Rate of gene expression as a function of two promotor proteins a and b.

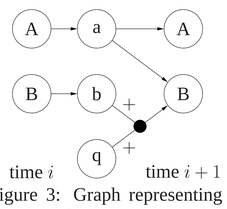

Figure 3: Graph representing a differential equation model for two genes with one external ligand.

Parameter Sampling

In Bayesian statistics one uses prior information as well as knowledge from data. Using a Bayesian approach we must define a space of models and assign a prior probability distribution on this space. This will inevitably be a finite, discrete space. Nonlinear coupled equations can not naturally be decomposed into parts and so we do not define probabilities for edges but consider the model as a whole. The posterior probability of a model $\mathcal{M}$ is given by

$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M})p(\mathcal{M})}{\int_{\mathcal{M}} p(\mathcal{D}|\mathcal{M})p(\mathcal{M})}\,, \tag{3}$$

where the likelihood marginalises over the high dimensional parameter space

$$p(\mathcal{D}|\mathcal{M}) = \int_{\theta} p(\mathcal{D}|\mathcal{M}, \theta)p(\theta)\, d\theta\,. \tag{4}$$

The parameters are all positive and given a gamma prior as in [6]. The likelihood is computed by simulating the equations for the given parameters, sampling values at time intervals and computing a similarity measure to the real data [6].

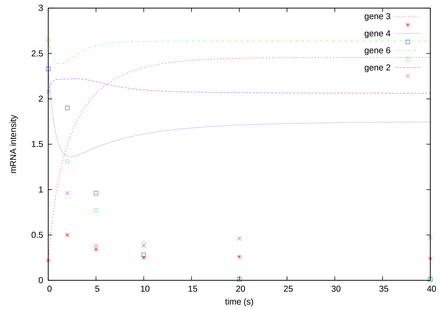

Figure 5: Graph of simulated mRNA expressions from low scoring model. The discrete points represent expressions sampled from an example model of nine genes with zero mean, 0.1 variance Gaussian noise. The continuous lines result from simulating the model with the parameter set of the highest data likelihood.

Model Sampling

Figure 5 shows data points generated using an unknown parameter set for a model of nine genes, eleven hidden proteins and 79 parameters. Starting with an arbitrary model we make small adjustments to propose a new model. They could be changes in edges or an inhibitor becoming a promotor, for example. A model is then accepted or rejected based on their relative marginal likelihoods with respect to the measured data. Models are generated in this way until the Markov Chains (MCs) indicate convergence, then samples of the models are retained. The models gathered form a discrete space with a multinomial probability distribution. This serves as a prior over models. We can then gather more data to reevaluate the models. The continuous lines in Figure 5 represent the simulation of a model with low posterior density. Note how the model gives qualitatively wrong predictions. Figure 6 shows data points superimposed with the simulations from a higher scoring model. This model is able to generate qualitatively correct predictions except for gene 1. This demonstrates that the posterior distribution for parameters of coupled nonlinear equations is very complicated and sensitive to small changes. Ten thousand samples where taken, which may not be sufficient to completely explore the high dimensional space but a larger number means that model evaluation becomes slow. A first improvement to the method of inferring systems of differential equations would be to use a population MC to explore more of the parameter space.

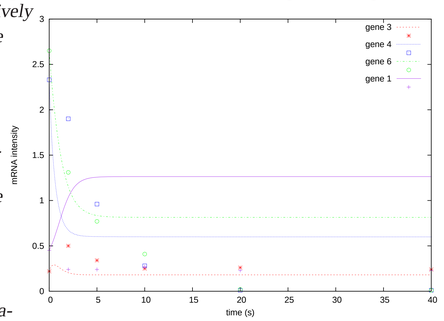

Figure 6: Graph of simulated mRNA expressions from low scoring model on the same data.

References

[1] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. *Using Bayesian Networks to Analyze Expression Data*. *Journal of Computational Biology*, 7:601–620, 2000.

[2] 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*, pages 175–186, 2002.

[3] S. Kim, S. Imoto, and S. Miyano. *Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data*. *Biosystems*, 75:57–65, 2004.

[4] 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.

[5] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D.L. Wild, and F. Falciani. *Modelling T-cell activation using gene expression profiling and state space models*. *Bioinformatics*, 20 (9):1361–1372, 2004.

[6] V. Vyshemirsky and M. Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, December 2007.