Model based identification of transcription factor activity from microarray data

Simon Rogers[1], Raya Khanin[2], and Mark Girolami[1]

[1] Bioinformatics Research Centre
Department of Computing Science
University of Glasgow
Glasgow, UK
[2] Department of Statistics
University of Glasgow
Glasgow, UK
{srogers,girolami}@dcs.gla.ac.uk, raya@stats.gla.ac.uk
http://www.dcs.gla.ac.uk/∼srogers/tfa/

1 Abstract

With the increase in volume of gene expression data available from high throughput microarray experiments, much research interest has been directed at building mathematical models of the process of gene regulation. Such models have primarily been used for the so called reverse engineering of regulatory networks; inferring possible regulatory interactions directly from microarray data, for example [1–4]. By using microarray data, all of these techniques make the implicit assumption that there is a direct relationship between the level of mRNA of genes coding for transcription factors (TFs) and the mRNA levels of their gene-targets. Whilst for some TF-gene pairs, this is likely to be a reasonable assumption, there are many examples of regulatory interactions where it is not due to modifications of the TF after translation. Such modifications cannot be measured on the microarray leading to minimal correlation between the expression levels of the TF gene and it's targets. It is obvious therefore that any models of regulation encoding a direct relationship between the mRNA levels of the two genes will be highly inaccurate over a wide range of interactions and conditions.

A particularly important example of such phenomena being observed in practice is the Hypoxia Inducible Factor-1 (HIF-1) gene investigated in [5]. HIF-1 is a TF that stimulates tumour growth and metastases. [5] found that although the HIF-1α protein was over-expressed in the majority of patients, no amplification of the HIF-1α gene were detected. Hence, some other process must be responsible. A second example, and one that we will study here can be found in the cell-cycle regulation of fission yeast. The SEP gene regulates several targets whos expression has been seen to vary periodically during the cell-cycle [6]. However, the expression of SEP does not vary periodically (see figure 1) suggesting some external influence on regulation.

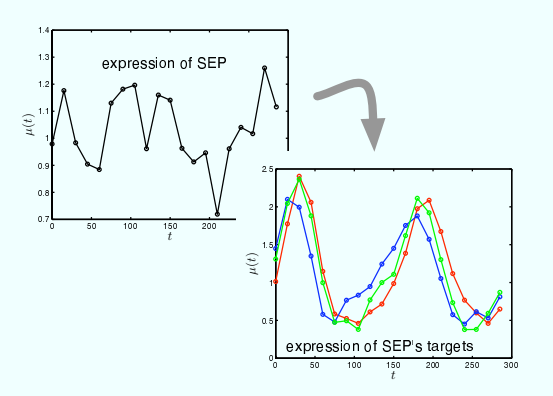


Fig. 1. Non-periodic expression of SEP and periodic expression of it's gene targets

Several methods have been proposed to overcome this problem, the majority of which limit themselves to linear (or log-linear) models of transcription [7, 8], More recently, [9] proposed a model based on a non-linear model of transcription (particularly, Michaelis-Menten kinetics). We extend on this work here.

The model of [9] assumed that the rate of mRNA production of the target genes for a particular TF followed the well known Michaelis-Menten (MM) kinetic model, given below for the case of activation

$$\dot{\mu}_{gi} = \alpha_g + \beta_g \frac{\eta_i}{\eta_i + \gamma_g} - \delta_g \mu_{gi} \tag{1}$$

where μ_{gi} is the expression of gene g at time i, η_i is the TFA at time i and $\theta_g = \{\alpha_g, \beta_g, \gamma_g, \delta_g\}$ is a set of gene specific kinetic parameters. The MM kinetic model is given by the second term on the right hand side and is parameterised by a gain term, β_g and a half-saturation constant γ_g. In addition to this, there are two additional terms corresponding to a basal level of production, α_g and a linear decay term with parameter δ_g . Given the values of the parameters and the TFA, integration of this differential equation provides an expression profile of gene g. Using a log-normal likelihood function, the probability of an expression dataset comprising G genes measured at T time-points with R replicates all regulated by a common transcription factor is given by

$$p(\mathbf{E}|\boldsymbol{\theta}, \boldsymbol{\eta}, \sigma^2) = \prod_{i=0}^{T} \prod_{r=1}^{R} \prod_{g=1}^{G} \frac{1}{\sqrt{2\pi}\sigma e_{gir}} \exp\left\{ -\frac{1}{2\sigma^2} (\log e_{gir} - m_{gi})^2 \right\} \tag{2}$$

where the location parameter m_{gi} is calculated by equating the μ_{gi} with the expected value of the log-normal distribution and the noise level σ^2 is treated as a parameter to be inferred. [9] perform inference of both the kinetic parameters and TFA by maximising this likelihood. The results presented are promising

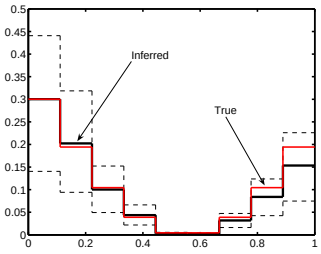
but there are certain drawbacks of the maximum likelihood approach. Namely, such approaches provide only point estimates of the quantities of interest and traditional methods of calculating confidence intervals are non-trivial due to the non-standard form of the model.

In this work, we make use of the Metropolis algorithm to perform fully Bayesian inference by sampling from the full posterior over TFA and parameter values. In addition to the benefit of obtaining full posterior distributions, a full Bayesian analysis provides further benefits. Any biological knowledge can be included through a suitable choice of prior distributions and the fact that it is not necessary to differentiate the likelihood function (as is required in the maximum likelihood solution) makes the model far more conducive to extension.

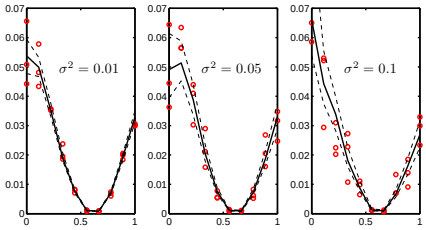
An example of the output of the sampler can be seen in figure 2. Here synthetic microarray datasets have been produced with $T = 10$ discrete timepoints, 3 replicates, $G = 10$ genes all activated by a particular TF and three different levels of noise variance, $\sigma^2 = \{0.01, 0.05, 0.1\}$. Figure 2(a) shows the true and inferred η profiles over time for the highest noise level, $\sigma^2 = 0.1$. The dotted lines indicate the 5th and 95th percentiles. Figure 2(b) shows the inferred expression profile for one of the genes across the three different levels of noise variance and figure 2(c) shows the inferred posterior for σ^2 for each of the three true values. It is obvious from the example that the method is able to re-create the true TFA profile with the percentiles providing useful information regarding confidence in values.

Figure 3 shows an example with real data from the cell-cycle regulation of fission yeast (data from [6]). We have already seen how the SEP TF has periodically expressed targets but is not periodically expressed itself. Figure 3(a) shows the inferred TFA which is periodic, as would be expected from the periodic expression of the target genes. An example of the data and the expression profile fitted by the model can be seen in figure 3(b). Finally, as a comparison, we have repeated this experiment but with eta fixed at the value of SEP's expression as if such a model will fit the data well, there is no benefit in inferring the TFA. The expression of the same gene shown in figure 3(b) under this simpler model can be seen in figure 3(c). It is obvious that the model is unable to convincingly fit the data. The improvement in likelihood when the TFA is inferred is significant with respect to the additional number of parameters (likelihood ratio test at 1%, Bayesian model comparison via marginal likelihoods is currently under investigation).

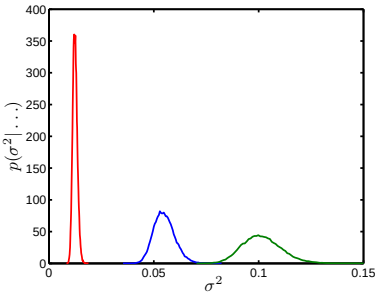
As well as providing confidence in inference and allowing the incorporation of prior knowledge, the Bayesian framework is far easier to extend than the maximum likelihood technique. For example, if we make the assumption that the kinetic parameters are constant across different conditions we are able to combine separate microarray datasets, a key problem in microarray analysis (see for example, [10]), where each dataset has its own TFA and noise parameter.



(a) True and inferred η profile with 5th and 95th percentiles



(b) Data and inferred expression for one gene at each noise level



(c) Inferred posterior over σ^2 (treated as a parameter to be inferred) for true $\sigma^2 = \{0.01, 0.05, 0.1\}$.

Fig. 2. Synthetic data example

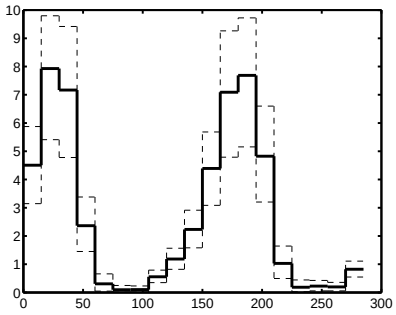
Such a technique is useful when one has only limited data for the organism of interest whilst other larger datasets are available. The Bayesian inference is particularly useful in this area as it is possible to see how the addition of more data provides tighter posterior distributions and hence higher confidence in the TFA and parameter values.

1.1 Acknowledgments

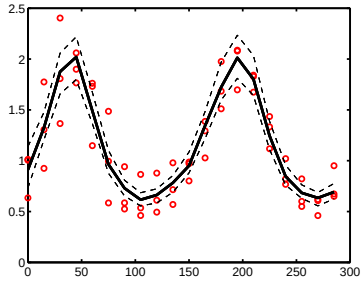
SR and MG are supported by EPSRC grant EP/C010620/1. RK is supported by a RCUK fellowship.

References

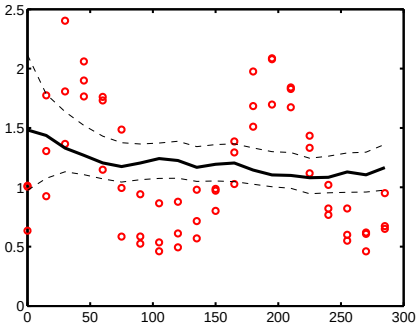
1. Rogers, S., Girolami, M.: A bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics **21**(14) (2005) 3131–3137
2. Yeung, M.K.S., Tegner, J., Collins, J.J.: Reverse engineering gene networks using singular value decomposition and robust regression. PNAS **99**(9) (2002) 6163–6168



(a) Inferred η profile



(b) Expression profile for individual gene when TFA is inferred.



(c) Expression profile for same gene when TFA is fixed at expression of SEP.

Fig. 3. Fission yeast example

3. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. Bioinformatics **19**(17) (2003) 2271–2282
4. Rice, J.J., Tu, Y., Stolovitzky, G.: Reconstructing biological networks using conditional correlation analysis. Bioinformatics **21**(6) (2005) 765–773
5. Vleugel, M., et al.: No amplifications of hypoxia-inducible factor-1a gene in invasive breast cancer: A tissue microarray study. Cellular Oncology **26**(5-6) (2004) 347–351
6. Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C., Burns, G., Hayles, J., Brazma, A., Bahler, J.: Periodic gene expression program of the fission yeast cell cycle. Nature genetics **36**(8) (2004) 809–817
7. Boulesteix, A.L., Strimmer, K.: Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. Theoretical Biology and Medical Modelling **2**(23) (2005)
8. Beal, M.J., Falciani, F., Ghahramani, Z., Rangel, C., Wild, D.L.: A bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioin-

formatics **21**(3) (2005) 349–356
9. Khanin, R., Vinciotti, V., Mersinias, M., Smith, C., Wit, E.: Statistical reconstruction of transcription factor activity using michaelis-menten kinetics. Biostatistics, submitted (2006)
10. Gilks, W.R., Tom, B.D.M., Brazma, A.: Fusing microarray experiments with multivariate regression. Bioinformatics **21**(supplement 2) (2005) ii137–143