A Bayesian regression and multiple changepoint model for systems biology

Frank Dondelinger[1,2,6], Andrej Aderhold[1], Sophie Lèbre[3], Marco Grzegorczyk[4,5], and Dirk Husmeier[1]

[1] Biomathematics and Statistics Scotland, JCMB, Edinburgh, EH9 3JZ, UK.
[2] Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, UK
[3] LSIIT - UMR 7005, Université de Strasbourg, 67412 Illkirch, France.
[4] Department of Statistics, TU Dortmund, 44221 Dortmund, Germany
[5] Department of Mathematics, Carl von Ossietzky University Oldenburg, Germany
[6] Communicating Author. Email: frankd@bioss.ac.uk

Abstract: We propose a Bayesian regression and multiple changepoint model for reverse engineering gene regulatory networks from high-throughput gene expression profiles. We report results from a recently held international gene network reconstruction competition, in which our method was objectively assessed in a blind study. While we did not win the competition, the scores indicate that the proposed method favourably compares with the majority of competing approaches and clearly belongs to the group of highest-ranked performers.

Keywords: Systems biology; gene regulatory network inference; Bayesian multiple changepoint model; RJMCMC; DREAM

1 Introduction

The objective of the highly topical field of systems biology is the reverse engineering of molecular regulatory networks and signalling pathways from high-throughput post-genomic data, and a flurry of activities in the statistics and machine learning communities are currently aimed at solving this problem. A variety of methods from statistics and machine learning have been applied to this end. See e.g. Grzegorczyk et al. (2008) and Cantone et al. (2009) for brief reviews. In the present paper, we propose a Bayesian regression and multiple changepoint model, with Bayesian inference based on reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). We participated in a recently held gene regulatory network prediction competition (DREAM 5), which assures that the comparative evaluation with other methods was done objectively.

2 Model

Multiple changepoints: Let p be the number of target genes, whose expression values $y = \{y_i(t)\}_{1\leq i\leq p, 1\leq t\leq N}$ are measured on N separate chips. $\mathcal{M}_i$ is the set of parents (regulators) associated with target gene i in the gene regulatory network. We model the differences in the regulatory relationships measured by different chips (assumed to be in some natural order, e.g. a time series) with a multiple changepoint process. For each target gene i, an unknown number k_i of changepoints define $k_i + 1$ non-overlapping segments. Segment $h \in \{1, .., k_i+1\}$ starts at changepoint ξ_i^{h-1} and stops before ξ_i^h, so that $\xi_i = (\xi_i^0, ..., \xi_i^{h-1}, \xi_i^h, ..., \xi_i^{k_i+1})$ with $\xi_i^{h-1} < \xi_i^h$. This changepoint process induces a partition of the chip ordering, $y_i^h = (y_i(t))_{\xi_i^{h-1}\leq t<\xi_i^h}$. The network structure $\mathcal{M}_i$ remains the same for each segment h, but the other parameters of the model can vary.
Regression model: For all genes i, the random variable $Y_i(t)$ refers to the expression of gene i on chip t. Within any segment h, the expression of gene i at chip t depends on the gene expression values on chip t of a set R_i of m potential regulator genes (parents), with $i \notin R_i$. We define a regression model by (a) the set of s_i parents denoted by $\mathcal{M}_i = \{j_1, ..., j_{s_i}\} \subseteq R_i$, and (b) a set of parameters $((a_{ij}^h)_{j\in R_i},\ \sigma_i^h)$; $a_{ij}^h \in \mathbb{R}$, $\sigma_i^h > 0$. For all $j \neq 0$, $a_{ij}^h = 0$ if $j \notin \mathcal{M}_i$. For all genes i, for all chips t in segment h ($\xi_i^{h-1} \leq t < \xi_i^h$), the random variable $Y_i(t)$ depends on the m variables $\{Y_j(t)\}_{j\in R_i}$ according to

$$Y_i(t) \ = \ a_{i0}^h + \sum\nolimits_{j\in\mathcal{M}_i} a_{ij}^h \ Y_j(t) \ + \varepsilon_i(t) \tag{1}$$

where the noise $\varepsilon_i(t)$ is assumed to be Gaussian with mean 0 and variance $(\sigma_i^h)^2$, $\varepsilon_i(t) \sim N(0, (\sigma_i^h)^2)$. We define $a_i^h = (a_{ij}^h)_{j\in R_i}$.
Prior: The k_i+1 segments are delimited by k_i changepoints, where k_i is distributed a priori as a truncated Poisson random variable with mean λ and maximum $\overline{k} = N-2$: $P(k_i|\lambda) \propto \frac{\lambda^{k_i}}{k_i!}\mathbb{1}_{\{k_i\leq\overline{k}\}}$. Conditional on k_i changepoints, the changepoint positions vector $\xi_i = (\xi_i^0, \xi_i^1, ..., \xi_i^{k_i+1})$ takes non-overlapping integer values, which we take to be uniformly distributed a priori. For all genes i, the number s_i of parents for node i follows a truncated Poisson distribution with mean Λ and maximum $\overline{s} = 5$: $P(s_i|\Lambda) \propto \frac{\Lambda^{s_i}}{s_i!}\mathbb{1}_{\{s_i\leq\overline{s}\}}$. Conditional on s_i, the prior for the parent set $\mathcal{M}_i$ is a uniform distribution over all parent sets with cardinality s_i: $P(\mathcal{M}_i\,||\mathcal{M}_i| = s_i) = 1/\binom{p}{s_i}$. The overall prior on the network structures is given by marginalization:

$$P(\mathcal{M}_i|\Lambda) = \sum\nolimits_{s_i=1}^{\overline{s}} P(\mathcal{M}_i|s_i)P(s_i|\Lambda) \tag{2}$$

Conditional on the parent set $\mathcal{M}_i$ of size s_i, we assume for the prior distribution $P(a_i^h|\mathcal{M}_i, \sigma_i^h)$ of the s_i+1 regression coefficients for each segment h a zero-mean multivariate Gaussian with covariance matrix $(\sigma_i^h)^2\Sigma_{a_i^h}$, where following Andrieu and Doucet (1999) we set $\Sigma_{a_i^h} = \delta^{-2}D_{a_i^h}^{\dagger}(y)D_{a_i^h}(y)$, and $D_{a_i^h}(y)$ is the $(\xi_i^h - \xi_i^{h-1}) \times (s_i + 1)$ matrix whose first column is a vector of 1 (for the constant in model (1)) and each $(j+1)^{th}$ column contains the observed values $(y_j(t))_{\xi_i^{h-1}-1\leq t<\xi_i^h-1}$ for all regulatory genes j in $\mathcal{M}_i$.

Finally, the conjugate prior for the variance $(\sigma_i^h)^2$ is the inverse gamma distribution, $P((\sigma_i^h)^2) = \mathcal{IG}(\upsilon_0, \gamma_0)$. Following Lèbre et al. (2010), we set the hyperparameters for shape, $\upsilon_0 = 0.5$, and scale, $\gamma_0 = 0.05$, to fixed values that give a vague distribution. The terms λ and Λ can be interpreted as the expected number of changepoints and parents, respectively, and δ^2 is the expected signal-to-noise ratio. These hyperparameters are drawn from vague conjugate hyperpriors, which are in the (inverse) gamma distribution family: $P(\Lambda) = P(\lambda) = \mathcal{G}a(0.5, 1)$ and $P(\delta^2) = \mathcal{IG}(2, 0.2)$.

Posterior: Equation (1) implies that

$$P(y_i^h|\xi_i^{h-1}, \xi_i^h, \mathcal{M}_i, a_i^h, \sigma_i^h) \quad \propto \quad \exp\left(-\frac{(y_i^h - D_{a_i^h}(y)a_i^h)^\dagger \ (y_i^h - D_{a_i^h}(y)a_i^h)}{2(\sigma_i^h)^2}\right) \quad (3)$$

From Bayes theorem, the posterior is given by the following equation:

$$P(k, \xi, \mathcal{M}, a, \sigma, \lambda, \Lambda, \delta^2|y) \ \propto \ P(\delta^2)P(\lambda)P(\Lambda)\prod_{i=1}^{p} P(k_i|\lambda)P(\xi_i|k_i)P(\mathcal{M}_i|\Lambda) \quad (4)$$

$$\prod_{h=1}^{k_i} P([\sigma_i^h]^2)P(a_i^h|\mathcal{M}_i, [\sigma_i^h]^2, \delta^2)P(y_i^h|\xi_i^{h-1}, \xi_i^h, \mathcal{M}_i, a_i^h, [\sigma_i^h]^2)$$

Inference: An attractive feature of the chosen model is that the marginalization over the parameters a and σ in the posterior distribution of (4) is analytically tractable: $P(k,\xi,\mathcal{M},\lambda,\Lambda,\delta^2|y) = \int P(k,\xi,\mathcal{M},a,\sigma,\lambda,\Lambda,\delta^2|y)dad\sigma$ See Andrieu and Doucet (1999), Lèbre et al. (2010) for details and an explicit expression. The number of changepoints and their location, k, ξ, the network structure $\mathcal{M}$ and the hyperparameters λ, Λ and δ^2 can be sampled from the posterior $P(k, \xi, \mathcal{M}, \lambda, \Lambda, \delta^2|y)$ with RJMCMC. A detailed description can be found in Lèbre et al. (2010). The posterior probabilities of the gene interactions submitted to DREAM are obtained from the posterior sample of network structures $\mathcal{M}$ by marginalization.

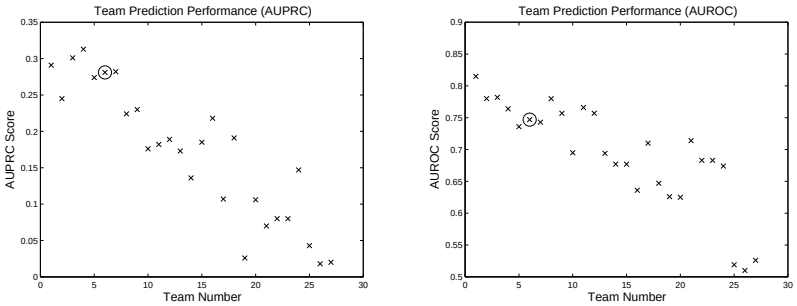

FIGURE 1. Areas under the precision recall (left) and ROC (right) curves obtained on an in silico data set by all teams participating in the DREAM 5 competition. The circles indicate the performance of our proposed method.

3 Simulations and Results

To assess the performance of the proposed method we participated in a competition organised by the DREAM (Dialogue for Reverse Engineering

TABLE 1. This table summarises the information about the DREAM 5 Network Inference Challenge data sets. For each data set, we show which organism it came from, how many genes were measured, how many of those genes were identified as transcription factors (possibly regulatory genes) and how many chips (datapoints) were included.

Data Set	Organism	Genes	Transcription Factors	Chips
1	Synthetic	1643	195	806
2	*S. Aureus*	2810	99	160
3	*E. Coli*	4511	334	805
4	*S. Cerevisiae*	5950	333	536

Assessments and Methods) consortium in autumn of 2010. The goal was to reverse engineer gene regulatory networks from gene expression data sets. Participants were given four microarray compendia and were challenged to infer the structure of the underlying transcriptional regulatory networks. The first compendium was based on an in-silico (i.e. simulated) network, the other three compendia were obtained from microorganisms. Each compendium consisted of hundreds of microarray experiments, which included a wide range of genetic, drug, and environmental perturbations. More information is available in Table 1 and at `http://wiki.c2b2.columbia.edu/dream/index.php/The_DREAM_Project`. Network predictions were evaluated by the organisers on a subset of known interactions for each organism, or on the known network for the in-silico case (which is more objective).

Our method assumes an ordering of the microarray chips. While this condition is naturally met for time course experiments, it does not hold for the varying experimental conditions of the DREAM data. We therefore resorted to the heuristic pre-processing step of mapping the high-dimensional gene expression profiles onto a one-dimensional self-organising map (SOM) initialized by the first principal component. We applied the software package *som* in R with default parameter settings. To reduce the computational complexity of the RJMCMC simulations we applied a pre-filtering step based on TESLA (Ahmed and Xing, 2009), a time-varying network inference method based on L1-regularised linear regression. For each gene we identified a set of 20 potential candidate regulators, based on the 20 regression coefficients with the largest modulus.

We assessed the convergence of our simulations with standard diagnostics based on Gelman-Rubin potential scale reduction factors (PSRF). Owing to unexpected downtime of the computer cluster we were using, only the simulations on the first two data sets showed a sufficient degree of convergence (PSRF≤ 1.2); for the latter data sets we submitted the results from TESLA. The second data set was later removed from the evaluation by the organisers. Figure 1 shows the results for the in silico data set obtained from the rankings of interactions submitted by all participating teams, using two

criteria: the area under the precision-recall curve (AUPRC), and the area under the receiver-operator characteristic (AUROC) curve. As discussed in Davis and Goadrich (2006), AUPRC gives a more faithful indication of the network reconstruction accuracy than AUROC, and it is thus seen that our method clearly lies in the group of the 5 top-ranked models. This suggests that it compares favourably with the majority of existing schemes and provides a useful tool for contemporary research in systems biology.

Acknowledgments: Marco Grzegorczyk is supported by the Graduate School Statistische Modellbildung of the Department of Statistics, TU Dortmund University. Dirk Husmeier is supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD) and under the EU FP7 project TiMet'. Andrej Aderhold's involvement in this project was funded by RERAD. Frank Dondelinger's research is funded by RERAD and the UK Engineering and Physical Sciences Research Council (EPSRC). We are grateful to Tony Travis for granting us user time on the Beowulf cluster at the Rowett Institute in Aberdeen and for providing support with respect to a parallelisation of the processes.

References

Ahmed, A., and Xing, E.P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, **106**:29, 11878-11883.

Andrieu, C., and Doucet, A. (1999). Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Transactions on Signal Processing*, **47**, 2667-2676.

Cantone, I., Marucci, L., Iorio, F., Ricci, M. A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., and Cosma, M. P. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, **4**:130.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proc. of the 23rd Int. Conf. on Machine Learning*

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**:4, 711-732.

Grzegorczyk, M., Husmeier, D. and Werhli, A.V. (2008). Reverse engineering gene regulatory networks with various machine learning methods. In: *Analysis of Microarray Data: A Network-Based Approach*, Wiley Online Library.

Lèbre, S., Becq, J., Devaux, F., Stumpf, M.P.H. and Lelandais, G. (2010). Statistical inference of the time-varying structure of gene regulation networks. *BMC Systems Biology*, **137**, 172-181.