

Supplementary paper on experimental aspects (E): Implementation details and supplementary figures for the article: Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler

Marco Grzegorzczuk^{1,5,*}, Dirk Husmeier^{2,5,*}, Kieron D. Edwards³, Peter Ghazal^{4,5} and Andrew J. Millar^{1,5}

¹ School of Biological Sciences, The University of Edinburgh, Swann Building, The King's Buildings, Edinburgh EH9 3JR, United Kingdom, ² Biomathematics and Statistics Scotland (BioSS), JCMB, King's Buildings, Edinburgh EH9 3JZ, United Kingdom, ³ Advanced Technologies (Cambridge) Ltd, Cambridge CB4 0WA, United Kingdom, ⁴ Division of Pathway Medicine (DPM), Medical School, The University of Edinburgh, Chancellor's Buildings, Edinburgh EH16 4SB, United Kingdom, and ⁵ Centre for Systems Biology at Edinburgh (CSBE), Darwin Building, King's Buildings, Edinburgh EH9 3JU, United Kingdom

ABSTRACT

Article: In the article we propose a non-linear and non-homogeneous generalization of the classical BGe score for Bayesian networks. The method is based on a mixture model, using latent variables to assign individual measurements to different classes. The practical inference follows the Bayesian paradigm and samples the network structure, the number of classes and the assignment of latent variables from the posterior distribution with MCMC, using the recently proposed allocation sampler as an alternative to RJMCMC.

Supplementary material: Due to space restrictions of the article we provide some additional information as supplementary material. The implementation details of all applied algorithms are given in Section 1. Additional figures and tables are provided in Section 2. The computational complexity of the proposed BGM algorithm is briefly discussed in Section 3. Finally, in Section 4 we provide a theoretical comparison with two related approaches by Lèbre (2008) and Ko *et al.* (2007).

Availability: This supplementary paper on experimental aspects (E) is available from

<http://www.bioass.ac.uk/associates/marco/supplement/E.pdf>

A separate supplementary paper on theoretical aspects (T) providing a more detailed presentation of the mathematical methodology is available from

<http://www.bioass.ac.uk/associates/marco/supplement/T.pdf>

The data sets used in our study are available from

<http://www.bioass.ac.uk/associates/marco/supplement/>

Contact: marco@bioass.ac.uk, dirk@bioass.ac.uk

1 IMPLEMENTATION DETAILS

We implemented structure MCMC according to the presentations given in Madigan and York (1995), and in all experimental

applications we used the following settings: For structure MCMC we set the burn-in length to 1,000,000 and then collected 500 graphs $\{\mathcal{G}_1, \dots, \mathcal{G}_{500}\}$ by sampling every 2000 iterations. For BGM we set the probability for a structure MCMC move to 0.5. And the probabilities of the other four move types, which all leave the graph \mathcal{G} unchanged, are set to 0.125. The maximal number of components \mathcal{K}_{MAX} was set to 10 and we note that this upper limit was never reached during any MCMC simulation. Equal to the structure MCMC setting we set the burn-in length to 1,000,000 and then collected 500 states $\{[\mathcal{G}_1, \mathcal{K}_1, \vec{v}_1], \dots, [\mathcal{G}_{500}, \mathcal{K}_{500}, \vec{v}_{500}]\}$ each consisting of a graph \mathcal{G}_i , a number of mixture components \mathcal{K}_i , and an allocation vector \vec{v}_i .

Following Werhli *et al.* (2006) we restricted the fan-in to 3 and employed the graph prior $P(\mathcal{G})$ given in Eq. (3) of the supplementary paper on theoretical aspects (T) when analysing the synthetic Gaussian data. This guarantees that our results for $\mathcal{K} = 1$ are comparable to those of Werhli *et al.* (2006). But the graph prior employed by Werhli *et al.* (2006) yields an intrinsic penalty for complex networks (see Subsection 1.1 of the supplementary paper on theoretical aspects (T)). Therefore and as we did not have any biological prior knowledge about the interactions in the macrophage and the Arabidopsis domain, the analysis of the gene expression data was performed with a uniform prior over graphs instead, i.e. every graph was set to be equally likely a priori. Furthermore, we decided not to restrict the fan-in for these relatively small domains with $N = 3$ (macrophage) and $N = 9$ (Arabidopsis) nodes only.

For the time series we did not allow for *self-loops*, that is we did not allow that a node can be its own parent node, by restricting the graph's neighbourhoods in Eq. (7) of the supplementary paper on

theoretical aspects (T) correspondingly. We decided to exclude self-loops, as they mainly capture degradation processes which are not of interest for modelling the regulatory interactions between genes.

Finally we note that we always performed two independent structure MCMC runs for each inference model: BDe, BGe, and BGM on every data set. Following Friedman and Koller (2003), we started the structure MCMC simulations with the BDe metric and the BGe metric from the following initialisations: (i) As uninformed initialization the first structure MCMC run was always seeded by an empty graph without any edges. (ii) To obtain an informed initialization we always performed a greedy search algorithm and seeded the second structure MCMC run with the most likely graph outputted by greedy search. We initialised the proposed BGM algorithm with the same graph as the corresponding BGe structure MCMC simulation, and the number of mixture components was set to $\mathcal{K} = 1$ so that all observations were allocated to the same single mixture component at the beginning of the MCMC simulation. We note that BGM inference with the restriction $\mathcal{K} = 1$ is equivalent to structure MCMC inference with the BGe scoring metric. Hence, a greedy search based on BGe can be seen as a greedy search based on the BGM model under the constraint that there is exactly one mixture component, symbolically: $\mathcal{K} = 1$. We note that it may be advisable to initialise the BGM algorithm not only with a graph found by a greedy search algorithm based on BGe but also with an allocation vector outputted by a classification or cluster algorithm. In our experiments we deliberately avoided to employ a more informative initialisation for BGM to demonstrate that BGM succeeds in inferring the true relationships - and especially the mixture components - independently of the initialisation.

Edge posterior probability scatter plots and trace plot diagnostics, e.g. of the number of edges of the sampled graphs or of their logarithmic scores, were used to assess convergence. Except for the synthetic Raf-Mek-Erk pathway data with $m = 480$ data points (where some MCMC simulations did not converge satisfactorily) we could see from the edge posterior probabilities that the total MCMC run-length of 2,000,000 for relatively small domains (between $N = 3$ and $N = 11$ nodes) had led to a satisfactory degree of convergence (Pearson correlation coefficients greater than 0.98) for all three inference models (BGe, BDe, and BGM). Therefore we report only the results of the empty-seeded runs in the article and point out that we had some convergence problems for the Raf-Mek-Erk pathway data with $m = 480$ data points.

The hyperparameters of the BGe and BGM models (see Section 4.1 of the supplementary paper on theoretical aspects (T)) were set as follows: $v = 1$, $\alpha = N + 2$, $\vec{\mu}_0 = (0, \dots, 0)^T$ and $T_0 = 0.5 \cdot I_{N,N}$ where $I_{N,N}$ is the N -by- N identity matrix and N is the number of domain variables. The choices of T_0 and $\vec{\mu}_0$ ensure that we are not explicitly biasing our inference to any particular edge (Friedman *et al.*, 2000). They reflect a prior belief where all N domain variables (genes) are identically and independently standard Gaussian distributed (with mean 0 and variance 1). The effective sample size parameters v and α were set to small values, as this ensures that the weight of the prior distribution (induced by T_0 and $\vec{\mu}_0$) is as uninformative as possible subject to the constraint that the resulting covariance matrix $T_{\mathcal{D}}$ (see Section 4.1 of the supplementary paper on theoretical aspects (T)) is non-singular (Geiger and Heckerman, 1994). The prior parameters for the BDe model were selected as in Giudici and Castelo (2003) to ensure (i) that the prior is uninformative (total prior decision was set to 1) and

(ii) that equal marginal likelihoods are given to equivalent DAGs. See Giudici and Castelo (2003) for further details.

2 SUPPLEMENTARY FIGURES AND TABLES

This second section provides additional figures and tables, which - due to space limitations - could not be included in the main paper. Most of the captions are self-explanatory, but some further explanations are given in the text. Figure 1 shows histograms of the posterior probabilities of the number of MCMC inferred mixture components for the synthetic Raf-Mek-Erk pathway data with $30 \leq m \leq 180$ data points. It can be seen that the proposed BGM model tends to infer the correct number of mixture components for these data sets. There are only 3 out of 16 combinations of \mathcal{K} and m for which an incorrect number of components was inferred, namely: $(\mathcal{K} = 3, m = 60)$, $(\mathcal{K} = 5, m = 60)$, and $(\mathcal{K} = 5, m = 180)$. Especially for the data sets with $\mathcal{K}_{TRUE} = 5$ mixture components it appears that this inaccuracy of the BGM inference is due to the fact that there are only few observations per mixture component, namely $m_i = 12$ for $m = 60$ and $m_i = 36$ for $m = 180$, so that the posterior probability landscape may be relatively flat around the true regulatory relationships. It can be seen from Figure 3 in the main paper that the BGM inference on the number of mixture components becomes more accurate when more data points ($m = 480$) are available.

The time series of the analysed Interferon regulatory factors (Irf1, Irf2, and Irf3) and scatter plots of the three Irf genes are shown in Figures 2 and 3. In both plots symbols indicate to which mixture component the observations were allocated. Concrete allocations were obtained by imposing thresholds on the connectivity matrices, whereby for each condition (CMV, IFN $_{\gamma}$, and CMV+IFN $_{\gamma}$) the threshold was selected such that an allocation consistent with the trends indicated by the corresponding heat matrix (shown in Figure 5 of the main paper) was obtained. From the time series and the scatter plots it appears that the inferred mixture components differ with respect to the marginal distributions of the three Irf genes; especially in Figure 3 most of the observations allocated to the same component tend to appear as clusters of points in the scatter plots.

The directed edge posterior probability estimates for the Interferon regulatory factor domain derived from BDe, BGe and BGM inference are given in Table 1. A concrete network prediction can be obtained from the estimates in Table 1 by imposing a threshold and extracting those edges only whose posterior probability estimate exceeds the predefined threshold. The AUROC scores resulting from the posterior probability estimates in Table 1 - under the assumption that the true regulatory relationships are as follows: $Irf2 \leftrightarrow Irf1 \rightarrow Irf3$ (Darnell *et al.* (1994) and Raza *et al.* (2008)) - are shown in Figure 6 panel (d) of the main paper.

The time series of the nine circadian genes in *Arabidopsis thaliana* are shown in Figure 4. Obviously all these genes have a strong 24hr circadian rhythm, and interestingly it can also be seen that the light:dark entrainment shifts the gene expression profiles. For most of the circadian genes the dashed line (T_{28} corresponding to 14h:14h entrainment) seems to be shifted by approximately 2 hours compared to the solid line (T_{20} corresponding to 10h:10h entrainment). This is in agreement with the BGM inference result where heat maps (see panels (c) and (d) in Figure 7 of the main paper) also indicate a time shift. Although the time lags differ (4-6

hours instead of 2 hours) it seems that the general trend, i.e. a time shift, has been captured by the proposed BGM model.

The directed edge posterior probability estimates for the circadian genes in *Arabidopsis thaliana* are given in Table 3 (T_{20}) and Table 4 (T_{28}). As explained above, concrete network predictions can be obtained from these estimates by imposing an arbitrary threshold on these posterior probability estimates. Furthermore, to illustrate graphically that the light:dark entrainment has an effect on the regulatory relationships, an edge posterior probability estimates scatter plot T_{20} versus T_{28} is given in Figure 5. Interestingly, it appears that the edge posterior probabilities are slightly different but do not completely differ; especially the edges with the highest posterior probability (around 1) are almost the same for both time series T_{20} and T_{28} . The Pearson correlation coefficient is equal to 0.84. Scatter plots of the directed edge posterior probabilities obtained by BGM inference versus BGe inference are shown in Figure 6. It can be seen from the two panels that the posterior probabilities are correlated and do not differ drastically. The Pearson correlations are equal to 0.94 (T_{20}) and 0.93 (T_{28}).

3 COMPUTATIONAL COMPLEXITY AND PERFORMANCE OF THE BGM ALGORITHM

The computational complexity of the proposed BGM algorithm depends on the number of network nodes N and the number of observations m . The computational complexity related to N is the same as for standard Bayesian network inference based on either the BGe or the BDe scoring metric. As the number of domain nodes N increases, convergence and mixing of the MCMC simulations become poorer, and the posterior distributions become more diffuse. To deal with the diffuse posteriors, the analysis of networks should focus on conserved subnetworks and network features, as discussed in Friedman *et al.* (2000). To improve mixing and convergence of the MCMC simulations, improved and alternative proposal scheme have been introduced; see Friedman and Koller (2003) and Grzegorzczuk and Husmeier (2008). These aspects have already been investigated in the literature before, and we therefore do not revisit them.

The additional complexity of the proposed BGM algorithm is also related to the data set size m , as each new data point is associated with a separate allocation variable, that is a new component of the allocation vector \vec{V} . To investigate how well our model scales up as m increases, we have also run simulations on larger synthetic Gaussian data sets with $m = 480$ data points, and we found that the computational costs do not increase substantially.

The BGM inference results suggest that the number of components in the heterogeneous data can be learned more accurately than with the smaller data set (see Figure 3 in the main paper and Figure 1 in this supplementary paper); however, the network reconstruction accuracy appears to slightly deteriorate (see Figure 1 and Figure 2 in the main paper). This finding might be counter-intuitive, as a larger data set contains more information and should therefore lead to a better performance. However, our finding is consistent with the fact that increased data set sizes lead to likelihood landscapes that are more rugged and, hence, result in increased mixing and convergence problems; see Figure 7 in Grzegorzczuk and Husmeier (2008). When learning conventional Bayesian networks based on the BGe and BDe scoring metrics this problem can be addressed, e.g. by improving the MCMC proposal moves, as reported in Grzegorzczuk

and Husmeier (2008). Unfortunately, this approach is not applicable to the proposed BGM model, as the reassignment of allocation variables requires a computationally expensive re-computation of the scores on which the proposal distributions depend. We therefore have to resort to classical structure MCMC Madigan and York (1995), which scales up less favourably to larger systems; see the discussions in Grzegorzczuk and Husmeier (2008). This problem can in principle be alleviated by the development of improved MCMC sampling schemes – akin to the improvement of MCMC schemes for conventional Bayesian networks (Grzegorzczuk and Husmeier, 2008) – but the practical implementation needs to be left for future research.

4 GENERAL DISCUSSIONS AND RELATED WORK

Bayesian networks provide an abstract and simplified representation of regulatory networks and signalling pathways, which is certainly not appropriate when trying to resolve the detailed structure of a specific pathway. There is a clear trade-off between model complexity and inference accuracy/computational complexity. Bayesian networks based on the BDe and BGe scoring metric are of a simple form, but allow the marginal likelihood to be computed analytically. More complex models along the line we discuss below sacrifice inference accuracy and resort to measures that are only reliable in the limit of very large data sets, like the Laplace approximation or, worse, the Bayesian information criterion BIC (Schwarz, 1978). Computing marginal likelihoods for even more accurate models based on differential equations have been attempted, but the computational costs are so high that this approach is restricted to model selection from a very small set of candidate pathways (Vyshemirsky and Girolami, 2008). We therefore hold the view that simpler models, like Bayesian networks using BGe (Geiger and Heckerman, 1994), still play an important role in systems biology.

In principle, one could obtain a model that is more flexible than the proposed BGM method by selecting the components and allocations for each domain variable separately. E.g. Ko *et al.* (2007) apply a mixture of Bayesian networks model to infer gene regulatory networks from expression data. In fact, the model of Ko *et al.* (2007) is more flexible than our BGM model, with node-specific Gaussian mixture models and, hence, node-specific breakpoints. However, the inference procedure is less sound in that the marginal likelihood is intractable. The authors resort to the Bayesian information criterion BIC for model selection, which is only a good approximation to the marginal likelihood in the limit of very large data sets, instead of a proper Bayesian network scoring metric based on the marginal likelihood, such as BGe or BDe. BIC is known to be a crude approximation to the proper BGe score, which in many practical applications is strongly over-regularized, especially when the data are sparse. Additionally, instead of sampling the network structure, the number of components, and the allocation of the observations from the joint posterior distribution with Markov chain Monte Carlo (MCMC), as in our work, the approach proposed in Ko *et al.* (2007) is based on a heuristic optimisation scheme that fails to take the intrinsic inference uncertainty into account.

Finally, we note that it has recently come to our attention that closely related work has been carried out in Lèbre (2008). The main differences are as follows. The present work has been motivated by the attempt to find a non-linear generalization of the BGe

	BDe				BGe				BGM		
	CMV				CMV				CMV		
	Irf ₁	Irf ₂	Irf ₃		Irf ₁	Irf ₂	Irf ₃		Irf ₁	Irf ₂	Irf ₃
Irf ₁	—	0.89	0.04	Irf ₁	—	1.00	0.83	Irf ₁	—	1.00	0.84
Irf ₂	0.63	—	0.91	Irf ₂	0.91	—	0.83	Irf ₂	0.86	—	0.40
Irf ₃	0.33	0.18	—	Irf ₃	0.98	0.51	—	Irf ₃	0.86	0.29	—
	IFN _γ				IFN _γ				IFN _γ		
	Irf ₁	Irf ₂	Irf ₃		Irf ₁	Irf ₂	Irf ₃		Irf ₁	Irf ₂	Irf ₃
Irf ₁	—	0.18	0.67	Irf ₁	—	0.75	0.79	Irf ₁	—	0.94	0.79
Irf ₂	0.05	—	0.03	Irf ₂	0.34	—	0.80	Irf ₂	0.77	—	0.37
Irf ₃	0.73	0.02	—	Irf ₃	0.67	0.44	—	Irf ₃	0.75	0.30	—
	CMV+IFN _γ				CMV+IFN _γ				CMV+IFN _γ		
	Irf ₁	Irf ₂	Irf ₃		Irf ₁	Irf ₂	Irf ₃		Irf ₁	Irf ₂	Irf ₃
Irf ₁	—	0.02	0.39	Irf ₁	—	0.77	0.80	Irf ₁	—	0.80	0.80
Irf ₂	0.01	—	0.02	Irf ₂	0.34	—	0.37	Irf ₂	0.44	—	0.37
Irf ₃	0.01	0.90	—	Irf ₃	0.66	0.34	—	Irf ₃	0.68	0.33	—

Table 1. Macrophage data: Inferred posterior probabilities of directed edges for each combination of experimental condition (CMV, IFN_γ, and CMV+IFN_γ) and BN inference procedure (BDe, BGe, and BGM). In each of the nine subtables the (i,j)-th cell contains the marginal posterior probability for an edge from Irf_i to Irf_j ($i, j = 1, \dots, 3$).

data	CMV	IFN _γ	CMV+IFN _γ
BDe	0.67	0.78	0.00
BGe	1.00	0.22	0.56
BGM	1.00	0.78	0.67

Table 2. Macrophage data: AUROC values. For each of the three macrophage data sets the table shows the BDe, BGe and BGM AUROC values computed from the directed edge relation features. The highest AUROC values for each data set are set in bold.

genes	LHY	CCA1	TOC1	ELF4	ELF3	GI	PRR9	PRR5	PRR3
LHY	—	1.00	0.53	0.37	0.43	0.35	0.19	0.15	0.35
CCA1	0.94	—	0.48	0.36	0.51	0.40	0.32	0.13	0.40
TOC1	0.08	0.15	—	0.28	0.47	0.09	0.28	0.15	0.33
ELF4	0.16	0.13	0.18	—	0.25	0.04	0.94	0.19	0.23
ELF3	0.09	0.15	0.08	0.13	—	0.04	0.53	0.15	0.15
GI	0.99	0.99	0.88	0.48	0.27	—	0.33	0.97	0.98
PRR9	0.49	0.26	0.20	0.43	0.26	1.00	—	0.90	0.19
PRR5	0.07	0.09	0.42	0.63	0.22	0.99	0.14	—	0.18
PRR3	0.11	0.15	0.11	0.14	0.24	0.06	0.17	0.16	—

Table 3. Arabidopsis thaliana T₂₀ data: Inferred posterior probabilities of directed edges. The estimates were obtained with BGM inference for time series T₂₀ (10h:10h light:dark entrainment). The (i,j)-th cell contains the marginal posterior probability of an edge from the gene in the i-th row to the gene in the j-th column.

genes	LHY	CCA1	TOC1	ELF4	ELF3	GI	PRR9	PRR5	PRR3
LHY	—	1.00	0.65	0.71	0.39	0.13	0.44	0.23	0.51
CCA1	0.92	—	0.40	0.39	0.61	0.16	0.35	0.51	0.26
TOC1	0.12	0.06	—	0.24	0.40	0.10	0.60	0.18	0.28
ELF4	0.09	0.11	0.14	—	0.23	0.05	0.44	0.08	0.08
ELF3	0.10	0.08	0.10	0.17	—	0.55	0.53	0.07	0.10
GI	1.00	1.00	0.75	0.63	0.30	—	0.16	0.89	0.92
PRR9	0.20	0.42	0.12	0.15	0.24	0.99	—	0.90	0.11
PRR5	0.18	0.13	0.62	0.37	0.24	0.92	0.21	—	0.65
PRR3	0.31	0.12	0.12	0.17	0.25	0.04	0.13	0.09	—

Table 4. Arabidopsis thaliana T₂₈ data: Inferred posterior probabilities of directed edges. The estimates were obtained with BGM inference for time series T₂₈ (14h:14h light:dark entrainment). The (i,j)-th cell contains the marginal posterior probability of an edge from the gene in the i-th row to the gene in the j-th column.

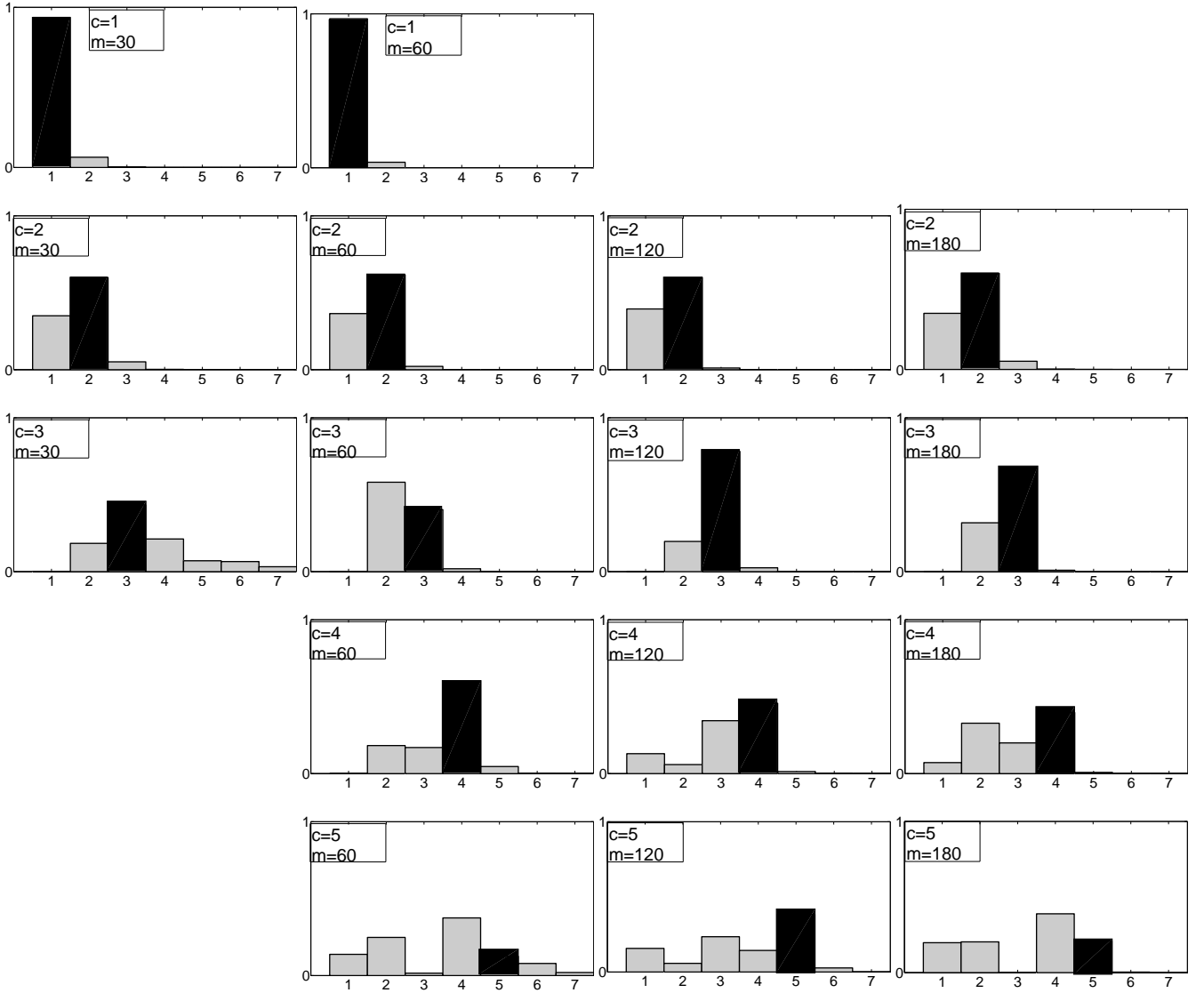


Fig. 1. Synthetic Gaussian data: Histograms of the number of inferred mixture components. For each considered combination of true components ($1 \leq \mathcal{K}_{TRUE} \leq 5$) and sample size m a histogram of the number of BGM-inferred components is shown. In each histogram the vertical axes represent posterior probabilities estimated with MCMC whereby the BGM MCMC trajectories have been merged across the 5 independent replications. From the histograms it can be seen that the posterior distribution of the number of mixture components \mathcal{K} inferred with BGM tends to peak at the correct number (indicated by black bars) for $\mathcal{K} \leq 4$. Only for the combination $\mathcal{K} = 3$ and $m = 60$ the posterior distribution of the number of inferred components wrongly peaks at $\mathcal{K} = 2$. For $\mathcal{K}_{TRUE} = 5$ (last row) the posterior distribution of the number of inferred components becomes flat and does not peak at the correct number of components for $m = 60$ and $m = 180$.

model, using a mixture distribution and the allocation sampler. The regionality, that is, the segmentation of the time series into consecutive segments has come out of the inference automatically, that is, it is purely data-driven. The breakpoint model applied in Lèbre (2008) imposes this structure onto the model a priori. While this is a useful assumption in most cases, it is more restricted in terms of modelling non-linear distributions. Also, if the regionality assumption is valid, it is straightforward to include it as prior knowledge in our model via a Markovian dependence between the latent variables. In fact, this approach could be regarded as a generalization of the breakpoint model, as discussed in Lehrach

(2007). The second difference is that Lèbre (2008) allows the model to learn different graphs between different breakpoints, while in our approach the graph is constrained to remain unchanged. While this makes the approach of Lèbre (2008) more flexible, it implies that there is no sharing of information between different breakpoints. To rephrase this: while the method of Lèbre (2008) infers the breakpoint structure from the whole data set, it infers a graph associated with a breakpoint only from the subset of the data assigned to the respective segment. Note that time series available for contemporary microarray studies are usually limited to a few dozen time points. Further decreasing the effective sample

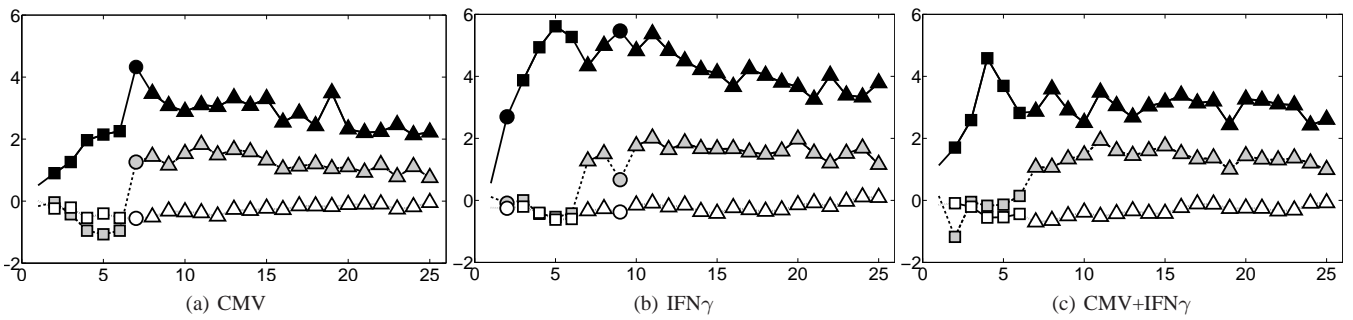


Fig. 2. Macrophage data: Gene expression time series of the Interferon regulatory factors. Black symbols: Irf1; grey symbols: Irf2; and white symbols: Irf3. Concrete allocations were obtained by imposing thresholds on the connectivity matrices, whereby for each condition the threshold was selected such that an allocation consistent with the trends indicated by the corresponding heat matrix shown in Figure 5 of the main paper was obtained. The different symbols (triangles, circles, squares) along the time series indicate which observations are then assigned to the same mixture component by the proposed inference scheme.

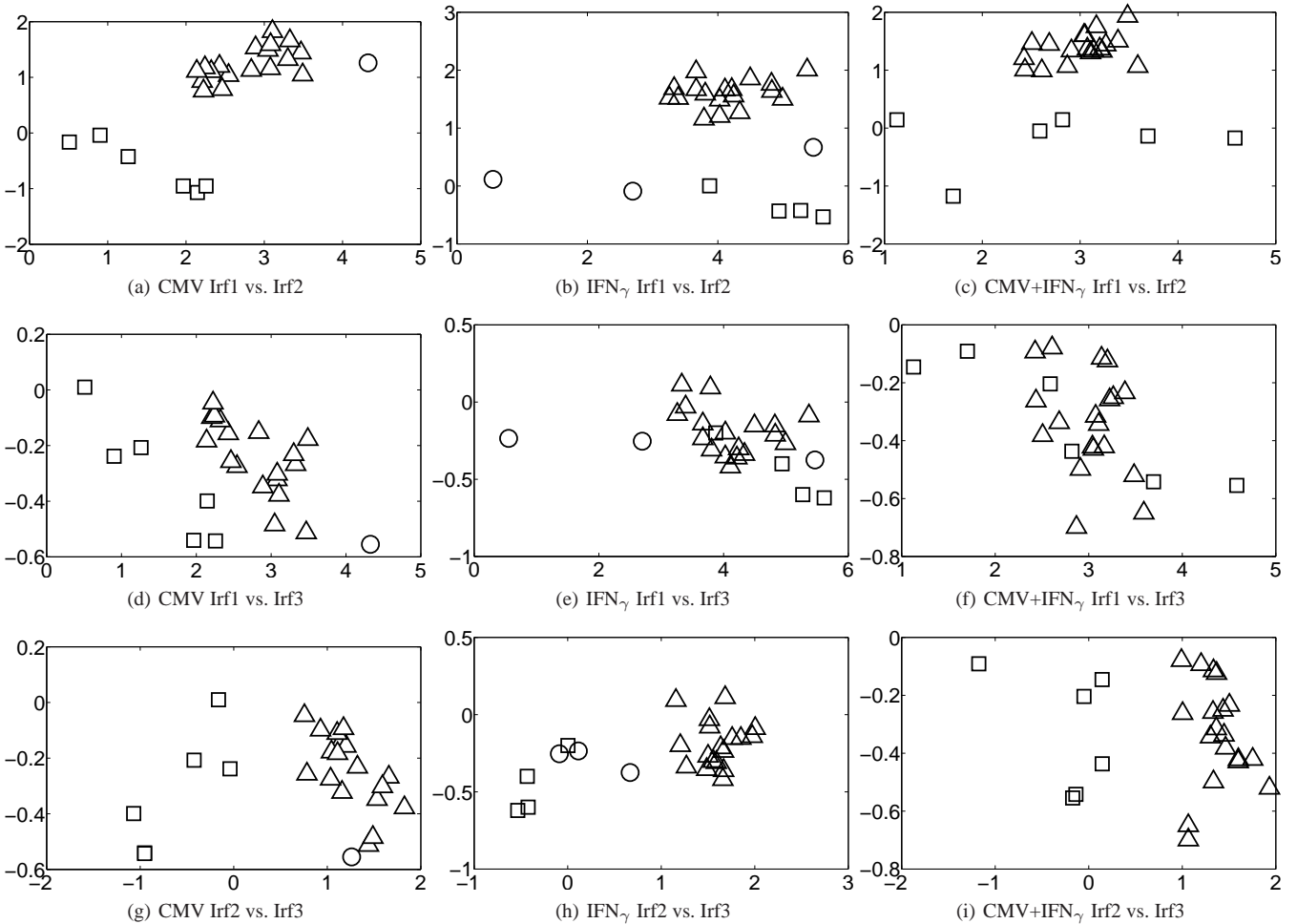


Fig. 3. Macrophage data: Scatter plots for the macrophage data. The figure shows scatter plots of the collected Irf gene expression data. For each condition (CMV, IFN_γ and CMV+IFN_γ.) there is a column with three panels showing the scatter plots for the three Irf gene pairs (Irf1 vs. Irf2, Irf1 vs. Irf3, and Irf2 vs. Irf3). The symbols (rectangles, triangles, and circles) indicate to which component the data points are allocated according to Figure 2. See caption of Figure 2 for more details.

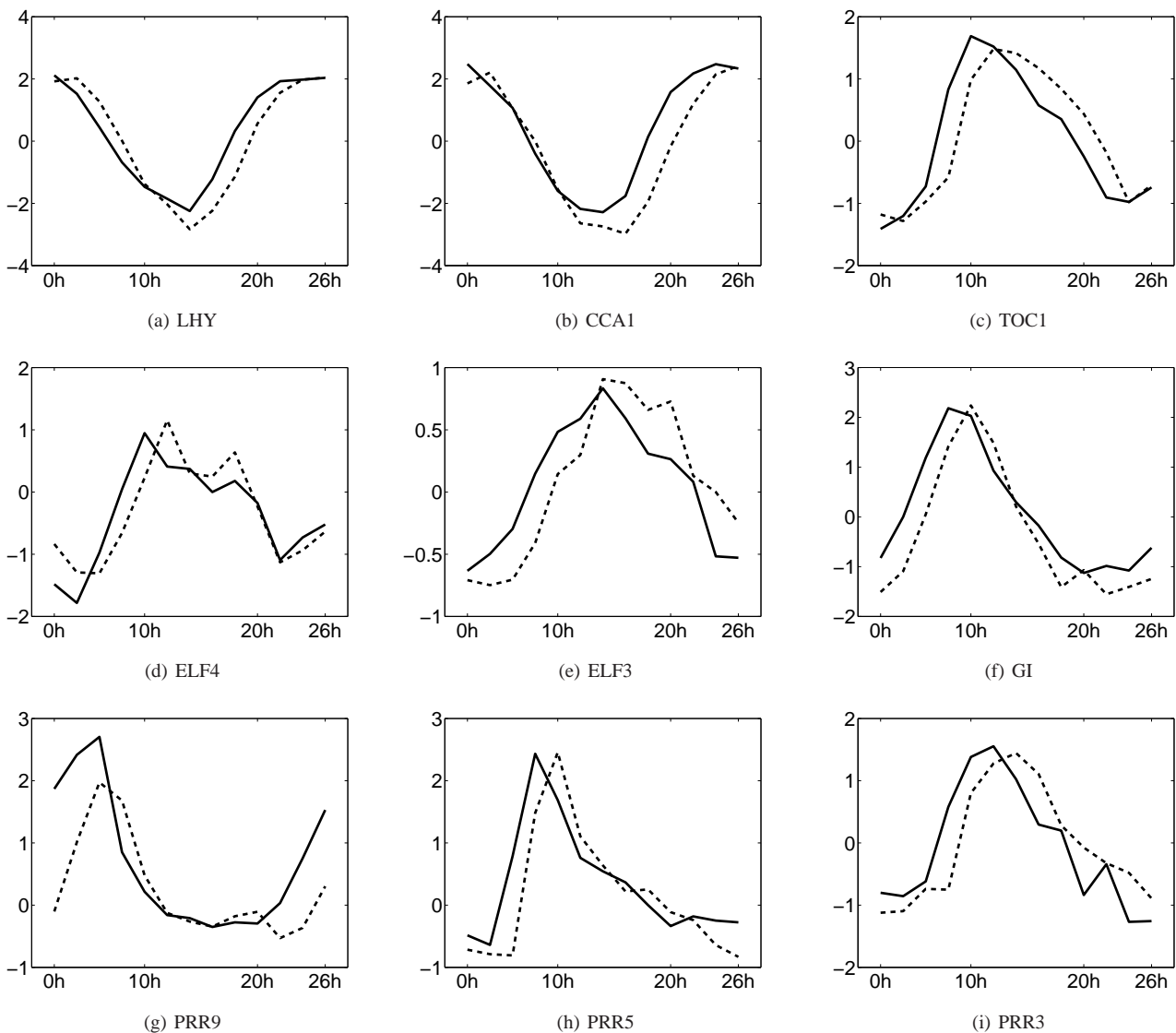


Fig. 4. Gene expression time series of nine circadian genes in *Arabidopsis thaliana*. For each of the selected nine circadian clock-regulated genes there is a plot of two time series. The solid lines refer to the measurements of time series T_{28} (14:14 light:dark entrainment) and the dashed lines refer to the measurements of series T_{20} (10:10 light:dark entrainment). It can be clearly seen that varying the entrainment lead to a phase shift of the gene expression profiles. For most of the circadian genes the dashed line (T_{28}) seems to be shifted by 2h compared to the solid line (T_{20}).

set size will inevitably increase the vagueness of the posterior distribution. By allowing for certain information sharing between the segments, our approach alleviates this problem. In other words, by assuming that the graph remains unchanged, and only allowing the distributions of the parameters associated with the interactions to vary between segments, the inference uncertainty is considerably reduced.

REFERENCES

- Darnell, J., Kerr, I. and Stark, G. (1994) Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science*, **264**, 1415–1421.
- Friedman, N. and Koller, D. (2003) Being Bayesian about network structure. *Machine Learning*, **50**, 95–126.
- Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.
- Geiger, D. and Heckerman, D. (1994) Learning Gaussian networks. *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 235–243.
- Giudici, P. and Castelo, R. (2003) Improving Markov chain Monte Carlo model search for data mining. *Machine Learning*, **50**, 127–158.
- Grzegorzczak, M. and Husmeier, D. (2008) Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, **71**, 265–305.
- Ko, Y., Zhai, C. and Rodriguez-Zas, S. (2007) Inference of gene pathways using Gaussian mixture models. In *BIBM International Conference on Bioinformatics and Biomedicine*, pp. 362–367. Fremont, CA.
- Lèbre, S. (2008) *Analyse de processus stochastiques pour la génomique : étude du modèle MTD et inférence de réseaux bayésiens dynamiques*. Ph.D. thesis, Université

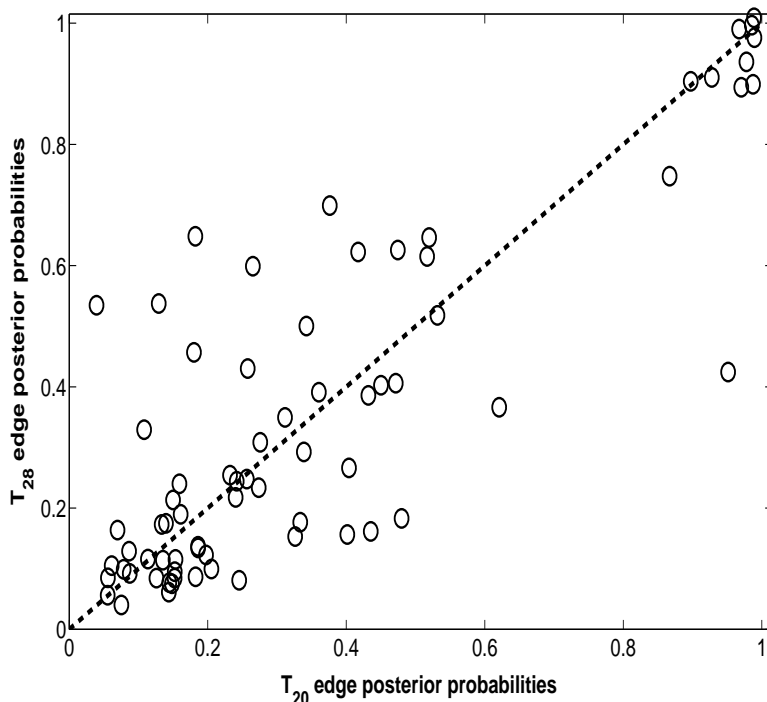


Fig. 5. Arabidopsis thaliana data. Scatter plot of edge posterior probabilities: T_{20} (horizontal axis) versus T_{28} (vertical axis). The Pearson correlation coefficient is equal to 0.84. The coordinates of all points were randomly slightly perturbed to visualize clusters of points.

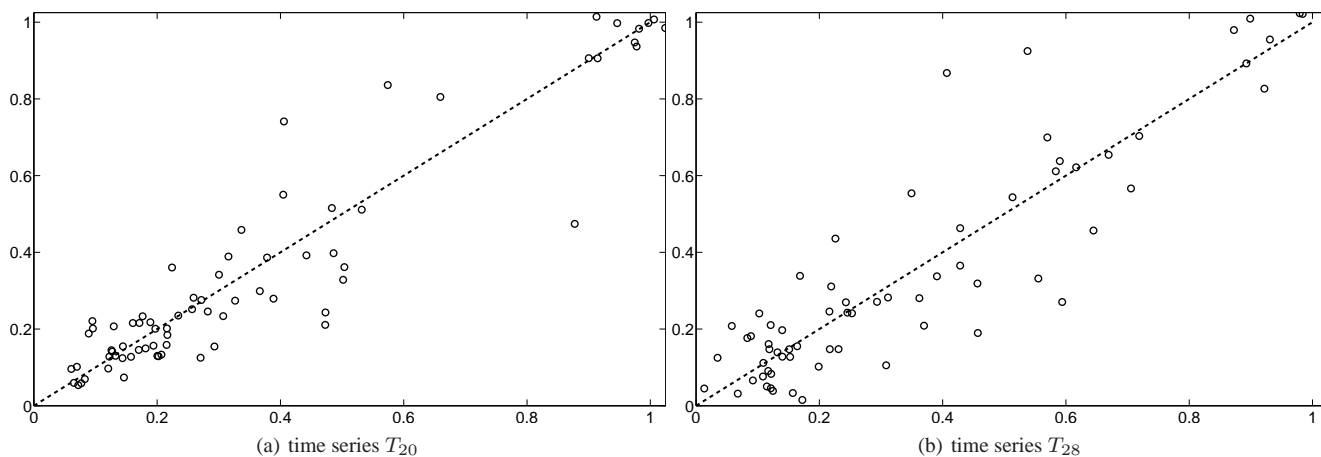


Fig. 6. Arabidopsis thaliana data. Edge posterior probabilities BGe versus BGM. For both time series T_{20} (panel (a)) and T_{28} (panel (b)) the edge posterior probabilities of BGM (horizontal axis) have been plotted against the edge posterior probabilities of BGe (vertical axis). The Pearson correlation coefficients are equal to 0.94 (T_{20}) and 0.93 (T_{28}). The coordinates of all points were randomly slightly perturbed to visualize clusters of points.

d'Evry-Val-d'Essonne.
 Lehrach, W. (2007) *Bayesian machine learning methods for predicting protein-peptide interactions and detecting mosaic structures in DNA sequence alignments*. Ph.D. thesis, University of Edinburgh.
 Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215–232.
 Raza, S., Robertson, K., Lacaze, P., Page, D., Enright, A., Ghazal, P. and Freeman, T. (2008) A logic based diagram of signalling pathways central to macrophage

activation. *BMC Systems Biology*, **2**. Article 36.
 Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
 Vyshemirsky, V. and Girolami, M. (2008) Bayesian ranking of biochemical system models. *Bioinformatics*, **24**, 833–839.
 Werhli, A. V., Grzegorzcyk, M. and Husmeier, D. (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, **22**, 2523–2531.