Systems biology

Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler

Marco Grzegorczyk$^{1,2,*}$, Dirk Husmeier$^{2,3,*}$, Kieron D. Edwards$^4$, Peter Ghazal$^{2,5}$ and Andrew J. Millar$^{1,2}$

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

Received on May 17, 2008; revised on July 9, 2008; accepted on July 14, 2008

ABSTRACT

Method: The objective of the present article is to propose and evaluate a probabilistic approach based on Bayesian networks for modelling non-homogeneous and non-linear gene regulatory processes. 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 Markov Chain Monte Carlo (MCMC), using the recently proposed allocation sampler as an alternative to RJMCMC.

Results: We have evaluated the method using three criteria: network reconstruction, statistical significance and biological plausibility. In terms of network reconstruction, we found improved results both for a synthetic network of known structure and for a small real regulatory network derived from the literature. We have assessed the statistical significance of the improvement on gene expression time series for two different systems (viral challenge of macrophages, and circadian rhythms in plants), where the proposed new scheme tends to outperform the classical BGe score. Regarding biological plausibility, we found that the inference results obtained with the proposed method were in excellent agreement with biological findings, predicting dichotomies that one would expect to find in the studied systems.

Availability: Two supplementary papers on theoretical (T) and experimental (E) aspects and the datasets used in our study are available from http://www.bioss.ac.uk/associates/marco/supplement/

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

1 INTRODUCTION

The ultimate objective of systems biology is the elucidation of the regulatory networks and signalling pathways of the cell. The ideal approach would be the deduction of a detailed mathematical description of the entire system in terms of a set of coupled non-linear differential equations. As high-throughput measurements at the cell level are inherently stochastic and most kinetic rate constants cannot be measured directly, the parameters of the system would have to be estimated from the data. Unfortunately, multiple parameter sets of non-linear systems of differential equations can offer equally plausible solutions, and standard optimization techniques in high-dimensional multimodal parameter spaces are not robust and do not provide a reliable indication of the confidence intervals. Most importantly, model selection would be impeded by the fact that more complex pathway models would always provide a better explanation of the data than less complex ones, rendering this approach intrinsically susceptible to over-fitting.

To assist the elucidation of regulatory network structures, probabilistic machine learning methods based on Bayesian networks can be employed, as proposed in the seminal paper by Friedman et al. (2000). In a nutshell, the idea is to simplify the mathematical description of the biological system by replacing coupled differential equations by simple conditional probability distributions of a standard form such that the unknown parameters can be integrated out analytically. This results in a scoring function (the ‘marginal likelihood’) of closed form that depends only on the structure of the regulatory network and avoids the over-fitting problem referred to above. Novel fast Markov Chain Monte Carlo (MCMC) algorithms, like Grzegorczyk and Husmeier (2008), can be applied to systematically search the space of network structures for those that are most consistent with the data. To obtain the closed form expression of the marginal likelihood referred to above, two probabilistic models with their respective conjugate prior distributions have been employed in the past: the multinomial distribution with the Dirichlet prior, leading to the so-called BDe score (Cooper and Herskovits, 1992), and the linear Gaussian distribution with the normal-Wishart prior, leading to the BGe score (Geiger and Heckerman, 1994). These approaches are restricted in that they either require the data to be discretized (BDe) or can only capture linear regulatory relationships (BGe). A non-linear non-discretized model based on heteroscedastic regression has been proposed by Imoto et al. (2003). However, this approach no longer allows the marginal likelihood to be obtained in closed form and requires a restrictive approximation (the Laplace approximation).
to be adopted. Another non-linear model based on node-specific Gaussian mixture models has been proposed in Ko et al. (2007). Again, the marginal likelihood is intractable. The authors resort to the Bayesian information criterion (BIC) of Schwarz (1978) for model selection, which is only a good approximation to the marginal likelihood in the limit of very large datasets. In the present article we propose a non-linear generalization of the BGe score, which is motivated by the fact that any probability distribution can, in principle, be approximated arbitrarily closely by a mixture model. Our method is based on recent work by Nobile and Fearnside (2007), who proposed the allocation sampler as an alternative to the computationally expensive approach of reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). We will describe the method in Section 2. We have evaluated our approach on a set of synthetic and real-world datasets described in Section 3 according to criteria outlined in Section 4. The results are presented in Section 5 and discussed in Section 6. A concluding summary of the proposed method and the results can be found in Section 7.

2 METHOD

We focus on the most important methodological aspects. A more detailed representation can be found in Supplementary Material T.

2.1 Bayesian network methodology

Static Bayesian networks (BNs) are interpretable and flexible models for representing probabilistic relationships between interacting variables. At a qualitative level, the graph of a BN describes the relationships between the domain variables in the form of conditional independence relations. At a quantitative level, local relationships between variables can be computed. We focus on the most important methodological aspects. A more detailed representation can be found in Supplementary Material T.

Equation (4) is replaced by:

\[ A(G_{i+1}|G_i) = \min \left\{ 1, \frac{P(D|G_{i+1})P(G_{i+1})}{P(D|G_i)P(G_i)} \right\} \]

otherwise the chain is left unchanged, symbolically \( G_{i+1} = G_i \). The Markov chain \( G_i \) converges to the posterior distribution \( P(D|G) \) (Madigan and York, 1995). If a fan-in restriction is imposed on the cardinality of the parent sets, all graphs possessing a node with more than fan-in parent nodes have to be excluded from the graph neighbourhoods. Structure MCMC generates a graph sample \( \{G_1, \ldots, G_T\} \), from which posterior probabilities of edges can be computed. We focus on undirected edges for independent data and directed edges for time-dependent data. There is an undirected edge between \( X_i \) and \( X_j \) (i < j) in \( G \), if \( G \) possesses either the edge \( X_i \rightarrow X_j \) or the edge \( X_j \rightarrow X_i \). Likewise, there is a directed edge from \( X_i \) to \( X_j \) (i ≠ j) in \( G \), if \( G \) possesses the edge \( X_j \rightarrow X_i \). An estimator for the posterior probabilities of an edge is given by the fraction of graphs in the sample that contain the edge of interest. When the true graph for the domain is known, the concept of receiver operator characteristic (ROC) curves and area under receiver operator characteristic (AUROC) values can be used to evaluate the global network reconstruction accuracy of BN inference (see e.g. Husmeier (2003) for details). An alternative and more intuitive criteria is given by (TP/FP = 5) counts: for each MCMC output a threshold \( \psi \) is imposed on the inferred edge posterior probabilities such that five false positive (FP) edges are extracted and the corresponding number of true positive (TP) edges, symbolically (TP/FP = 5), exceeding the threshold \( \psi \), is counted (Werhli et al., 2006).

2.2 Gaussian mixture Bayesian network model

We assume that we have either \( m \) independent and identically distributed (iid) observations (BNs) or \( m+1 \) time-dependent observations with a homogeneous first-order Markovian dependence structure (DBNs) for the variables \( X_1, \ldots, X_N \). This gives a dataset matrix of size \( N \times m \), where \( D_j \) (j = 1, ..., m) is the j-th observation of the N nodes. The allocation vector \( \tilde{V} \) of size \( m \) defines an allocation of the \( m \) observations to \( K \) mixture components: \( \tilde{V}(j) = k \) means that the j-th observation is allocated to the k-th component. \( D^{(V,k)} \) denotes the data subset consisting of all observations allocated to the k-th component by \( \tilde{V} \) (1 ≤ k ≤ K). The joint posterior probability of a graph \( G \), an allocation vector \( \tilde{V} \), and \( K \) mixture components can be factorized as follows:

\[ P(G, \tilde{V}, K|D) = P(G, \tilde{V}|K)P(K|D) \propto P(K)P(\tilde{V}|K)P(G|D) \]

where \( P(K) \) is the prior distribution of the number of mixture components. The new observation belongs to the k-th component if \( \tilde{V}(j) = k \). The allocation vector \( \tilde{V} \) can be estimated by maximum likelihood.
Absorption moves are of the same functional form:

\[ P(D|g, V, K) = \prod_{k=1}^{K} P(D|V^{(k)}|g) \]  

(7)

In Equation (7) the likelihood terms \( P(D|V^{(k)}|g) \) for the data subsets \( D^{V^{(k)}} \) given the same graph \( g \) can be computed independently with the BGe scoring metric of Geiger and Heckerman (1994), as derived and discussed in Supplementary Materials T. If no observation is allocated to the \( k \)-th component \( (D^{V^{(k)}} = \emptyset) \), then \( P(D|V^{(k)}|g) \) is equal to 1. As we do not have any prior knowledge about the graph topology we assume a uniform prior distribution on graphs for the real gene expression data, \( P(g) = \text{const.} \). For the synthetic Raf-Mek-Erk network data we employ a more restrictive graph prior (see Supplementary Materials T and E). For the prior on \( K \), \( P(K) \), we take a truncated Poisson distribution with parameter \( \lambda = 1 \) restricted to \( 1 \leq K \leq K_{\text{MAX}} \). This prior is known to be suitable for finite mixture models (Nobile, 2005). We further assume that the probability distribution of the allocation vector \( V \) conditional on \( K \) is given by:

\[ P(V|K, \tilde{p}) = \prod_{k=1}^{K} p_{k}^{n_{k}} (1 - p_{k})^{(a_k)} \]  

(8)

where \( \tilde{p} = (p_1, \ldots, p_K)^{T} \) and \( \sum_{k=1}^{K} p_{k} = 1 \) are the non-negative mixture weights, and \( n_k \) is the number of observations allocated to the \( k \)-th component by \( V \). The prior on the mixture weights \( \tilde{p} = (p_1, \ldots, p_K)^{T} \) is chosen to be a Dirichlet distribution, \( P(\tilde{p}) = \text{Dir}(\alpha_1, \ldots, \alpha_K) \), with hyperparameters \( \tilde{a} = (\alpha_1, \ldots, \alpha_K)^{T} \). This prior is conjugate, and the marginal probability of \( V \) conditional on \( K \) is thus given by

\[ P(V|K) = \int P(V|K, \tilde{p}) P(\tilde{p}) d\tilde{p} = \text{Dir}(n_1 + \alpha_1, \ldots, n_K + \alpha_K) \]  

(9)

2.3 Gaussian mixture allocation MCMC inference

The new Gaussian mixture allocation MCMC sampling scheme (BGM) generates a sample from the joint posterior distribution \( P(g, K, V|D) \) given in Equation (6) and comprises six different types of moves in the state-space \( [g, K, V] \). The first move type is a structure MCMC single edge operation on the graph \( g \) while the number of components \( K \) and the allocation vector \( V \) are left unchanged. According to Equation (4), a new graph \( g' \) is proposed, and the new state \( [g', K, V] \) is accepted or rejected according to Equation (5) where the likelihood terms \( P(D|g') \) in Equation (5) have to be replaced by the \( P(D|g, K) \) terms given in Equation (7). The five other move types are adapted from Nobile and Fearnside (2007) and operate on \( V \) or on \( K \) and \( V \). If there are \( K > 2 \) mixture components, then moves of the type M1 and M2 can be used to re-allocate some observations from one component \( k_1 \) to another \( k_2 \). That is, a new allocation vector \( V' \) is proposed while \( g \) and \( K \) are left unchanged. The Ejection move type proposes an increase in the number of mixture components by one. It randomly selects two mixture components and simultaneously tries to re-allocate some observations to fill the new component. More precisely, it randomly selects a mixture component and tries to re-allocate some of its observations to the newly proposed component \( K+1 \), while \( g \) is left unchanged. The Absorption move is complementary to the Ejection move and decreases the number of mixture components by one. It randomly selects two mixture components and deletes one of them after having reallocated all of its observations to the other component. The acceptance probabilities for M1, M2, Ejection and Absorption moves are of the same functional form:

\[ A = \begin{bmatrix} 1 & P(V'|K) & P(D|g, V', K) & Q(V'|V) & P(V') \end{bmatrix} \begin{bmatrix} P(V|K) & P(D|g, V, K) & Q(V|V') & P(V) \end{bmatrix}^{-1} \]  

(10)

where the likelihood terms have been specified in Equation (7), the proposal probabilities \( Q(\cdot|\cdot) \) depend on the move type (M1, M2, Ejection or Absorption), and \( K^* = K \) for M1 and M2 moves \( K^* = K + 1 \) for Ejection moves, and \( K^* = K - 1 \) for Absorption moves. See Supplementary Material T and Nobile and Fearnside (2007) for details. Finally, the sixth move type uses Gibbs sampling to re-allocate a single observation by sampling its new allocation from the corresponding univariate conditional distribution, while leaving \( K \) and the other components of \( V \) unchanged.

3 DATA

We have evaluated the proposed method on synthetic data generated from a widely studied protein signalling network and on gene expression time series from two different biological systems. For details of the simulation studies see Supplementary Material E.

3.1 Synthetic data

For a comparative evaluation study, Werhli et al. (2006) generated synthetic datasets for the Raf-Mek-Erk signalling pathway presented in Sachs et al. (2005), which consists of 11 nodes representing phosphorylated proteins and 20 directed edges. As Werhli et al. (2006) assigned Gaussian regulatory mechanisms with varying (randomly sampled) parameters, we can generate Gaussian mixture network data as follows: for obtaining data with \( K = 1, \ldots, 5 \) components we randomly selected \( K \) of the original datasets, sampled the same number of observations \( m/K \) from each and merged these observations to a single dataset of size \( m \). For each of 16 combinations of \( m = 30, 60, 120, 180 \) and \( K \) we generated five datasets by applying this procedure. Additionally, for each \( K = 1, \ldots, 5 \) we generated five further datasets along this line Werhli et al. (2006) with \( m = 480 \) observations each.

3.2 Bone marrow-derived macrophages

Interferons (IFNs) play a pivotal role in the innate and adaptive mammalian immune response against infection, and central research efforts, therefore, aim to elucidate their regulatory interactions (Honda et al., 2006). For the present study, we have applied our method to gene expression time series from bone marrow-derived macrophages, which were sampled at 24 × 30 min time intervals. The macrophages were subjected to three external conditions: (1) infection with Cytomegalovirus (CMV), (2) treatment with Interferon Gamma (IFNγ) and (3) infection with Cytomegalovirus after pretreatment with IFNγ (CMV+IFNγ). To obtain the gene expression profiles, samples derived from the macrophages were hybridized to Agilent mouse genome arrays. Samples were co-hybridized with a pooled common control RNA. Expression levels were obtained in the form of log2 scale signal intensity ratios between the sample and the pooled control RNA. Differential dye-label incorporation between the two samples on each array was corrected by applying a within-array, non-linear, loess normalization to the ratios. Global non-biological variations between ratio distributions were corrected by applying median absolute-deviation between-array normalization. We focus on time series of the Interferon regulatory factors (Irfs) 1, 2 and 3 (which we write as Irf1, Irf2 and Irf3, respectively), as a gold standard network for the interactions between these factors can be derived from the literature (Darnell et al., 1994; Raza et al., 2008): Irf2 −→ Irf1 −→ Irf3. The Irfs are the key regulators in the response of the macrophage cell to pathogens. They mediate the cellular signalling that leads to a transcriptional response to the initial binding events on the surface of the cell.

3.3 Circadian regulation in Arabidopsis thaliana

We have also applied our method to two gene expression time series from A.thaliana cells, which were sampled at 13 × 2 h time intervals with Affymetrix microarray chips, and robust multi-array (RMA) normalized. The expressions were measured twice independently.
under experimentally generated constant light condition, but differed with respect to the prehistories. In the first experiment, \( T_{20} \), the plant was entrained in a 10h:10h light/dark cycle, while the plant in the second experiment, \( T_{28} \), was entrained in 14h:14h light/dark cycle. Our analysis focuses on nine genes, namely LHY,CCA1,TOC1,ELF4,ELF3,GI,PRR9,PRR5 and PRR3, which are known to be involved in circadian regulation (Mas, 2008; Salome and McClung, 2004).

4 EVALUATION

To evaluate the proposed method (BGM), we have compared it with Bayesian learning of homogeneous BNs using the standard BGe score, as described in Geiger and Heckerman (1994). We have applied a 3-fold evaluation procedure. First, we use static synthetic network data to show that if the data are, in fact, of a heterogeneous nature, BGM achieves improved network reconstruction results. Second, using gene expression time series from bone marrow-derived macrophages, we focus on a small subsystem of the IFN pathway whose biology is well understood, and we demonstrate that BGM leads to a better pathway reconstruction. Third, we consider a larger set of circadian genes involved in circadian regulation (Mas, 2008; Salome and McClung, 2008). ELF4, ELF3, GI, PRR9, PRR5 and PRR3, which are known to be involved in circadian regulation (Mas, 2008; Salome and McClung, 2004).

As before, \( K \) denotes the number of mixture components, \( \tilde{V} \) denotes the allocation vector, \( G \) denotes the graph, and let \( \tilde{q} \) denote the vector of parameters associated with \( G \). We get the following expression for the predictive distribution:

\[
P(\tilde{D}|D,H_1) = \sum_{K,V,G} P(\tilde{D}|K,\tilde{V},G,q,H_1)P(K,\tilde{V},G,q|D,H_1)d\tilde{q}
\]

A possible approach is to approximately sample \([K,\tilde{V},G]\) and \(\tilde{q}\) from the posterior distribution \(P(K,\tilde{V},G,q|D,H_1)\) with MCMC and to approximate the integral in Equation (13) by a sum over this sample. A better method is to use the expansion \(P(K,\tilde{V},G,q|D,H_1) = P(\tilde{q}|K,\tilde{V},G,D,H_1)P(K,\tilde{V},G,D,H_1)\) and draw on the fact that

\[
\Psi(K,\tilde{V},G,\tilde{D}) = \int P(\tilde{D}|K,\tilde{V},G,q,H_1)P(\tilde{q}|K,\tilde{V},G,D,H_1)d\tilde{q}
\]

which in practice is computed from a sample \([\tilde{K},\tilde{V}_1,\tilde{G}_1],...,[\tilde{K}_T,\tilde{V}_T,\tilde{G}_T]\) approximately drawn from the posterior distribution \(P(K,\tilde{V},G|D,H_1)\) with MCMC:

\[
P(\tilde{D}|D,H_1) = \frac{1}{T}\sum_{i=1}^{T} \Psi(K,\tilde{V}_i,\tilde{G}_i,\tilde{D})
\]

The computation of \(\Psi(K,\tilde{V},G,\tilde{D})\) in (14) requires only a minor modification of the standard BGe score discussed in Geiger and Heckerman (1994). The vector \(\tilde{V}\) acts as a filter dividing the data into different categories, for which separate BGe scores are computed. For instance, if we have 2 states, 10 time points and \(\tilde{V}=[1111122222]\), then separate BGe scores are computed for the first five and the last five time points. The computation of the BGe score is modified by the fact that the prior distribution \(P(\tilde{q}|G,H_1)\) is replaced by the posterior distribution \(P(\tilde{q}|K,\tilde{V},G,D,H_1)\). This results in a straightforward modification of the score as follows: in Equation (13) of Geiger and Heckerman (1994), those training data that correspond to the corresponding state \(k, [D_i \in D|\tilde{V}(i)=k]\), are included in the conditioning part of the distribution, and the sufficient statistics are adjusted accordingly. We note that BDe and BGM cannot be compared in terms of predictive distributions, as the required data discretization (BDe) is not part of the BN model. That is, while BGM and BGe model the same datasets \(D\) and \(\tilde{D}\), BDe is based on their discretized counterparts, resulting from some (heuristic) pre-processing.

5 RESULTS

The mean AUROC values and the mean (TP|FP=5) counts for assessing the reconstruction of the Raf-Mek-Erk pathway from the synthetic data, described in Section 3.1, are represented as histograms in Figures 1 and 2. It can be seen that BGM performs significantly better than BDe and BGe for almost all combinations of \(K\) and \(m\). Only if there is either one single component or a small sample size \((m=30)\), there is no (significant) difference between BGM and BGe. In particular for \(K=1\), BGM assigns all observations to one single component, and so does not differ.
Fig. 1. Raf-Mek-Erk network reconstruction accuracy for synthetic data. Histograms of the network reconstruction accuracy for different combinations of \( \mathcal{K}_{\text{TRUE}} (\mathcal{K}_{\text{TRUE}} = 1, \ldots, 5) \) and sample size \( m \) (\( m = 30, 60, 120, 180 \)) assessed in terms of mean AUROC values (panels (a)-(e)) and (TP|FP=5) counts (panels (f)-(j)) derived from undirected edges. White bars refer to BDe, grey bars refer to BGe, and black bars refer to BGM. The SDs are indicated by vertical black lines.

Fig. 2. Raf-Mek-Erk network reconstruction accuracy for synthetic data with \( m = 480 \). Histograms of the network reconstruction accuracy for \( \mathcal{K}_{\text{TRUE}} = 1, \ldots, 5 \) assessed in terms of mean AUROC values (a) and (TP|FP=5) counts (b) derived from undirected edges. White bars refer to BDe, grey bars refer to BGe and black bars refer to BGM. The SDs are indicated by vertical black lines.

from BGe. Figure 3 reveals that BGM infers for each number of components \( \mathcal{K}_{\text{TRUE}} \) the correct number of components for the synthetic data with \( m = 480 \) observations. Histograms of the numbers of inferred components for the synthetic data with fewer data points \( m \) are provided in Figure 1 in Supplementary Material E.

A comparison of Figures 1 and 2 reveals that the reconstruction accuracy is slightly worse for the datasets with \( m = 480 \) observations. This finding might appear counterintuitive, as larger datasets contain more information and should therefore lead to better performances. However, our finding is consistent with the fact that increased dataset sizes lead to likelihood landscapes that are more rugged and, hence, result in increased mixing and convergence problems. This shortcoming of the structure MCMC sampler by Madigan and York (1995) has already been reported (e.g. see Grzegorczyk and Husmeier, 2008).

For the macrophage gene expression time series, BGM infers \( \mathcal{K} = 2 \) components for the conditions CMV and IFN-\( \gamma \), while for the third condition (CMV+IFN-\( \gamma \)) most of the sampled states consist of \( \mathcal{K} = 1 \) component only, as shown in Figure 4. The fraction of sampled states for which two observations \( i \) and \( j \) are allocated to the same component \( k \) (\( 1 \leq k \leq \mathcal{K} \)) can be used as a connectivity measure \( C(i,j) \). Figure 5 displays the resulting connectivity matrices graphically as heat matrices. From the heat matrices the same systematic trend can be observed for the three conditions. The first part (observations no. 2–6) and the last part of the three time series (observations no. 8–25) are allocated to different components. For condition CMV (IFN-\( \gamma \)) the allocation of observation no. 7 (no. 9) is not fixed, that is, the allocation changes during the MCMC simulation. For CMV+IFN-\( \gamma \), whose number of components peaks at \( \mathcal{K} = 1 \) (Fig. 4), the separation between the two parts is less pronounced, though consistent with the other results. To understand whether BGM also leads to a better network reconstruction accuracy, we compare the mean posterior probabilities of the true and false edges of BGM in Figure 6 with those obtained from BGe and BDe. For the IFN-\( \gamma \) condition (Fig. 6b) it becomes obvious that BGM has performed substantially better than BGe and BDe. For the other two conditions the difference between the posterior means for the true and the false edges is also best for BGM, but the difference is less pronounced [BDe: 0.24, BGe: 0.24, BGM: 0.39 (CMV) and BDe: −0.42, BGe: 0.09, BGM: 0.14 (CMV+IFN-\( \gamma \)]. Since it appears that the three conditions do not lead to systematic deviations between the expression profiles of Ifr1, Ifr2 and Ifr3, we treat the three experiments as independent replications and compute predictive probabilities, as discussed in Section 4. The predictive probabilities for BGM are much higher than those of BGe (Table 2). This finding provides further evidence that BGM does not overfit the data but outputs results that can be confirmed by independent replications. The BGM/BGe Bayes factors are: 36.45 (CMV), 2.73 (IFN-\( \gamma \)) and 0.71 (CMV+IFN-\( \gamma \)). This finding is consistent with Figure 4, where the peaks for CMV and IFN-\( \gamma \) are at 2, while CMV+IFN-\( \gamma \) peaks at 1. Gene expression time series plots and scatter plots for the three Ifr factors can be found in Supplementary Material E.

For both A.thaliana gene expression time series (see Supplementary Material E) the number of components inferred with BGM peaks at 2 (Fig. 7a and b). The heat matrices shown in Figure 7a and b appear to be of a similar structure, but subject to a translation along the main diagonal. More precisely, it appears that the transition from the first to the second component is shifted by 2–3 time points (4–6 h). Compared with BGe the Bayes factors are in favour of BGM: 5.66 (\( T_{20} \)) and 9.41 (\( T_{28} \)). The predictive probabilities are given in Table 1 and confirm the improved generalization performance. Further plots for the Arabidopsis data
Fig. 3. Histograms of the numbers of BGM components for synthetic Gaussian data with $m = 480$ observations. The posterior probabilities (vertical axis) of the number of components $K$ (horizontal axis) have been estimated from the MCMC trajectories. For $K_{\text{TRUE}} = 1, \ldots, 5$ the MCMC trajectories for the five datasets have been merged.

Fig. 4. Histograms of the numbers of BGM components for macrophage gene expression time series. For each experimental condition the posterior probability (vertical axis) of the number of components $K$ (horizontal axis) have been estimated from the MCMC trajectories.

Fig. 5. Graphical presentation of the temporal connectivity structure for the macrophage gene expression data. The figure shows heatmap representations that indicate the estimated posterior probability of two time points being assigned to the same state (component). The probabilities are represented by a grey shading, where white corresponds to a probability of 1, and black corresponds to a probability of 0. The numbers on the axes represent the time points of the time course experiment. The analysis was repeated for all three experimental conditions CMV, IFN$\gamma$ and CMV+IFN$\gamma$, as explained in the text.

are provided in Supplementary Material E, and the inference results are discussed in more detail in Section 6.

6 DISCUSSION

The results for the synthetic data generated from the Raf-Mek-Erk pathway of Sachs et al. (2005) show that the proposed BGM scheme consistently outperforms the conventional BGe and BDe metrics in terms of global network reconstruction accuracy (Figs 1 and 2). This confirms that BGM is superior when the data stem from a mixture distribution, and that the proposed sampling scheme (allocation MCMC) renders the inference, which is more complex than for the conventional case, practically viable. Furthermore, histograms of the number of inferred mixture components (Fig. 3) reveal that BGM succeeds in inferring the correct number of components. To assess whether BGM achieves any improvement for real biological applications, we applied it to gene expression data obtained from two different platforms (Agilent and Affymetrix) for two different systems: macrophages challenged with viral infection, and circadian rhythms in plants.

For macrophages challenged with CMV or pretreated with IFN$\gamma$, BGM tends to infer a two-stage process (Fig. 4). This two-stage process reflects a state change in the host macrophage brought about by infection (CMV) or immune activation (IFN$\gamma$), and can be found in all three experimental conditions (Fig. 5). Interestingly, though, this state change is less pronounced in the combined condition CMV+IFN$\gamma$ (Fig. 4c), where the Bayes factor does not support the more flexible heterogeneous model (see the previous section).
Fig. 6. Reconstructing the regulatory network of the Irf5s. (a–c): Mean posterior probabilities (vertical axis) of true and false edges in the Irf regulatory network, inferred with BDe, BGe and BGM (horizontal axis) from the macrophage gene expression time series. According to the biological literature the true edges are: Irf1 → Irf2, Irf2 → Irf1 and Irf3 → Irf1, while the edges Irf1 → Irf3, Irf2 → Irf3 and Irf3 → Irf2 are spurious. In (d) an AUROC histogram plot is given. For each of the three conditions the histogram shows bars of the BDe (white), BGe (grey) and BGM (black) AUROC values. It can be seen that BGM is never inferior to BDe or BGe in terms of AUROC scores, but BGM outperforms (i) BDe for conditions CMV and CMV + IFNγ and (ii) BGe for conditions IFNγ and CMV + IFNγ.

Table 1. Logarithmic predictive probabilities for the A.thaliana data: $\log_e(P(D|D,H_0))$ (BGe) and $\log_e(P(D|D,H_1))$ (BGM)

<table>
<thead>
<tr>
<th>$D$</th>
<th>$H_i$</th>
<th>$\tilde{D}=T_{20}$</th>
<th>$\tilde{D}=T_{28}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{20}$</td>
<td>BGe</td>
<td>–</td>
<td>–64.29 (+0.29)</td>
</tr>
<tr>
<td></td>
<td>BGM</td>
<td>–</td>
<td>–53.69 (+0.42)</td>
</tr>
<tr>
<td>$T_{28}$</td>
<td>BGe</td>
<td>–63.93 (+0.22)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>BGM</td>
<td>–54.78 (+0.63)</td>
<td>–</td>
</tr>
</tbody>
</table>

The SDs of the logarithmic probabilities are given in brackets.

Table 2. Logarithmic predictive probabilities for the macrophage data: $\log_e(P(D|D,H_0))$ (BGe) and $\log_e(P(D|D,H_1))$ (BGM)

<table>
<thead>
<tr>
<th>$D=\mathcal{D}_{\text{TRAIN}}$</th>
<th>Model</th>
<th>CMV</th>
<th>IFNγ</th>
<th>CMV and IFNγ</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGe</td>
<td>–</td>
<td>–76.01 (+0.07)</td>
<td>–45.26 (+0.03)</td>
<td></td>
</tr>
<tr>
<td>BGM</td>
<td>–</td>
<td>–63.63 (+0.02)</td>
<td>–33.80 (+0.38)</td>
<td></td>
</tr>
<tr>
<td>IFNγ</td>
<td>BGe</td>
<td>–56.78 (+0.05)</td>
<td>–57.30 (+0.05)</td>
<td></td>
</tr>
<tr>
<td>BGM</td>
<td>–39.62 (+0.02)</td>
<td>–42.69 (+0.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMV + IFNγ</td>
<td>BGe</td>
<td>–37.76 (+0.08)</td>
<td>–69.19 (+0.06)</td>
<td></td>
</tr>
<tr>
<td>BGM</td>
<td>–21.67 (+0.33)</td>
<td>–53.26 (+0.51)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The SDs of the logarithmic probabilities are given in brackets.

This observation is consistent with the known biological responses of macrophages to simultaneous infection by virus (mCMV) and immune (IFNγ) activation. It suggests that upon dual challenge with both an infection and immune activation (CMV + IFNγ) signalling leads to a pronounced singular response. This is in agreement with observations of cooperation between viral and immune signalling in effective vigorous anti-viral state within the host macrophage, as discussed in Benedict et al. (2001).

For the A.thaliana gene expression time series, BGM also infers a two-stage process (Fig. 7). In this application, the two stages are most likely related to the diurnal nature of the dark/light cycle. We have applied our method to two sets of plant samples, which were subjected to different prehistories, related to different lengths of the artificial, experimentally controlled light/dark cycle. Although the two-stage nature of the process is preserved, the state co-allocation posterior probabilities, shown in the heatmap of Figure 7, points to a phase shift of about 4-6 h as a consequence of the increased day length. This phase shift is biologically plausible and indeed expected. It can be explained by the early phase of entrainment that is required to elicit a phase delay that matches the 24-h period of the wild-type plants to the longer light/dark cycle ($T_{28}$), compared to the later phase of entrainment required to elicit a phase advance to match the shorter light/dark cycle ($T_{20}$) (Johnson et al., 2003).

We anticipate that a non-linear and non-homogeneous generalization of (Bayesian) networks will have broader general utility for reconstructing regulatory networks in systems biology. In this regard there is increasing interest in the development of new statistical methods, as exemplified by the recent and related work of Lèbre (2008). Our article complements this work and constitutes a natural generalization of the BGe score of (Geiger and Heckerman, 1994) by applying the ideas of mixture models and allocation sampling presented in Nobile and Fearnside (2007). This is an advantage over the work of Ko et al. (2007). While the latter model is more flexible owing to the fact that different nodes can have different breakpoints, it leaves the computation of the marginal likelihood intractable. The authors resort to BIC (Schwarz, 1978) as a crude approximation to the marginal likelihood. However, this approximation is only valid in the limit of very large datasets, and BIC is known to be over-regularized in many practical applications. For a more detailed theoretical comparison between BGM and the approaches of Lèbre (2008), and Ko et al. (2007) see Supplementary Material E. The evaluation of the proposed BGM approach on synthetic benchmark data and the novel application to two real biological scenarios provide an encouraging demonstration of the viability of the proposed BGM method.

7 CONCLUSION

We have proposed a non-linear and non-homogeneous generalization of the BGe score for Bayesian networks (BGM). BGM 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 graph, the number of classes and the assignment of latent variables from the posterior distribution with MCMC, using the allocation sampler of Nobile and Fearnside (2007) as an alternative to RJMCMC (Green, 1995). We have evaluated BGM using three criteria: network reconstruction, statistical significance and agreement with
intrinsic biological features. In terms of network reconstruction, we found improved results both for a synthetic network of known structure (Figs 1 and 2) and for a small real regulatory network derived from the literature (Fig. 6). For assessing the statistical significance of the improvement, we computed two scores: Bayes factors and predictive distributions. We applied these scores to gene expression time series obtained on different platforms (Agilent and Affymetrix) for two different systems (viral challenge of Arabidopsis thaliana and interferon-β establishing host–virus détente. Immunity, 20, 741–744.

In Figures 5. A comparison between the two panels reveals a phase shift of about 2–3 time points (4–6 h.) between the different entrainments $T_{28}$. This result provides excellent agreement with intrinsic dichotomies that we expect to find in these systems, related to the dichotomy between the healthy and diseased state of the cell, and the diurnal contrast between light and darkness.

ACKNOWLEDGEMENTS

We are grateful to T. Forster, S. Watterson, K. Robertson, and P. Dickinson for discussions and assistance with the handling and interpretation of the data, and to D. Nutter for technical support with computational issues.

Funding: M.G. is supported by the Biotechnology and Biological Sciences Research Council (BBSRC) and by the Engineering and Physical Sciences Research Council (EPSRC). D.H. is supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD). Experimental work on Arabidopsis was supported by BBSRC grant G19886 to A.J.M. The Centre for Systems Biology at Edinburgh is a Centre for Integrative Systems Biology (CiSB) funded by BBSRC and EPSRC, reference BB/D019621/1.

Conflict of Interest: none declared.

REFERENCES


