

BAYESIAN INTEGRATION OF BIOLOGICAL PRIOR KNOWLEDGE INTO THE RECONSTRUCTION OF GENE REGULATORY NETWORKS WITH BAYESIAN NETWORKS

Dirk Husmeier* and Adriano V. Werhli

*Biomathematics and Statistics Scotland,
Edinburgh, United Kingdom*

*Email: dirk@bioss.sari.ac.uk

There have been various attempts to improve the reconstruction of gene regulatory networks from microarray data by the systematic integration of biological prior knowledge. Our approach is based on pioneering work by Imoto et al.¹¹, where the prior knowledge is expressed in terms of energy functions, from which a prior distribution over network structures is obtained in the form of a Gibbs distribution. The hyperparameters of this distribution represent the weights associated with the prior knowledge relative to the data. To complement the work of Imoto et al.¹¹, we have derived and tested an MCMC scheme for sampling networks and hyperparameters simultaneously from the posterior distribution. We have assessed the viability of this approach by reconstructing the RAF pathway from cytometry protein concentrations and prior knowledge from KEGG.

1. INTRODUCTION

Bayesian networks have received increasing attention from the computational biology community as models of gene regulatory networks, following up on pioneering work by Friedman et al.⁴ and Hartemink et al.⁶. Several tutorials on Bayesian networks have been published^{8, 10, 16}. We therefore only qualitatively recapitulate some aspects that are of relevance for the present study, and refer the reader to the above tutorials for a thorough and more rigorous introduction.

The structure of a Bayesian network is defined by a directed acyclic graph (DAG) indicating how different variables of interest, represented by nodes, “interact”. The word “interact” has a causal connotation, which is ultimately of interest to the biologist, but has to be taken with caution in this context, as explained shortly. The edges of a Bayesian network are associated with conditional probabilities, defined by a functional family and their parameters. The interacting entities are associated with random variables, which represent some measured entities of interest, like relative gene expression levels or protein concentrations. We denote the set of all the measurements of all the random variables as the data, represented by the letter D . As a consequence of the acyclicity of the network structure, the joint probability of all the random variables can be factorized into a product of lower-complexity con-

ditional probabilities according to conditional independence relations defined by the graph structure G . Under certain regularity conditions, the parameters associated with these conditional probabilities can be integrated out analytically. This allows us to compute the marginal likelihood or evidence $P(D|G)$, which captures how well the network structure G explains the data D . In the present study we computed $P(D|G)$ under the assumption of a linear Gaussian distribution. The resulting score was derived by Geiger and Heckerman⁵ and is referred to as the BGe score.

We are interested in learning a network of causal relations between interacting nodes. While such a causal network forms a valid Bayesian network, the inverse relation does not hold: when we have learned a Bayesian network from the data, the resulting graph does not necessarily represent the correct causal graph. One reason for this discrepancy is the existence of unobserved nodes. When we find a probabilistic dependence between two nodes, we cannot necessarily conclude that there exists a causal interaction between them, as this dependence could have been brought about by a common yet unobserved regulator. However, even under the assumption of complete observation the inference of causal interaction networks is impeded by symmetries within so-called equivalence classes, which consist of networks that yield the same evidence scores $P(D|G)$.

*Corresponding author.

A simple example are two conditionally dependent nodes, say A and B , where the two networks related to the two possible directions of the edge, $A \rightarrow B$ and $A \leftarrow B$, are equivalent.

There are two ways to break the symmetries of the equivalence classes. One approach is to use active interventions, like gene knockouts and over-expressions. When knocking out gene A affects gene B , while knocking out gene B does not affect gene A , then $A \rightarrow B$ will tend to have a higher evidence than $A \leftarrow B$. For more details, see Refs. 23, 24. An alternative way to break the symmetries, investigated in this paper, is to use prior information. When genes A and B are conditionally dependent, and we have prior knowledge that A is a transcription factor that regulates genes in the functional category that B belongs to, then we will presumably favour $A \rightarrow B$ over $A \leftarrow B$. To formalize this notion, we score networks by the posterior probability

$$P(G|D) \propto P(D|G)P(G) \quad (1)$$

where $P(D|G)$ is the evidence, and $P(G)$ is the prior distribution over network structures; the latter distribution captures the biological knowledge that we have prior to measuring the data D . While different graphs might have identical scores in light of the data, $P(D|G)$, symmetries can be broken by the inclusion of prior knowledge, $P(G)$, and these two sources of information are systematically integrated into the posterior distribution $P(G|D)$. Our ultimate objective, hence, is to find the network structure G that maximizes $P(G|D)$. Unfortunately, the number of structures increases super-exponentially with the number of nodes. Also, in systems biology, where we aim to learn complex interaction patterns involving many components, the amount of information from the data and the prior is usually not sufficient to render the distribution $P(G|D)$ sharply peaked at a single graph. Instead, the distribution is usually diffusely spread over a large set of networks. Summarizing this distribution by a single network would not be appropriate. Instead, we aim to sample network structures from the posterior distribution $P(G|D)$ so as to obtain a typical collection of high-scoring networks and, thereby, capture intrinsic inference uncertainty. Direct sampling from this distribution is usually intractable, though. Hence, we resort to a Markov chain Monte Carlo (MCMC) scheme¹⁷, which under fairly general regu-

larity conditions is theoretically guaranteed to converge to the posterior distribution of equation (1)⁷. Given a network structure G_{old} , a new network structure G_{new} is proposed from the proposal distribution $Q(G_{\text{new}}|G_{\text{old}})$, which is then accepted according to the standard Metropolis-Hastings scheme⁷ with the following acceptance probability:

$$A = \min \left\{ \frac{P(D|G_{\text{new}})P(G_{\text{new}})}{P(D|G_{\text{old}})P(G_{\text{old}})} \times \frac{Q(G_{\text{old}}|G_{\text{new}})}{Q(G_{\text{new}}|G_{\text{old}})}, 1 \right\} \quad (2)$$

The functional form of the proposal distribution $Q(G_{\text{new}}|G_{\text{old}})$ depends on the chosen type of proposal moves. In the present paper, we consider three edge-based proposal operations: creating, deleting, or inverting an edge. The computation of the Hastings factor $Q(G_{\text{old}}|G_{\text{new}})/Q(G_{\text{new}}|G_{\text{old}})$ is, for instance, discussed in Ref. 10.

2. METHODOLOGY

2.1. Biological prior knowledge

To integrate biological prior knowledge into the inference of gene regulatory networks, we define a function that measures the agreement between a given network G and our biological prior knowledge. Following an approach first proposed by Imoto et al.¹¹ and subsequently applied in Refs. 12, 18, 21, 22, we call this measure the energy E , borrowing the name from statistical physics. We split E into two components. One of the components, E_0 , is associated with the absence of edges. The other component, E_1 , is associated with the presence of edges. A network G is represented by a binary adjacency matrix, where each entry G_{ij} can be either 0 or 1. A zero entry, $G_{ij} = 0$, indicates the absence of an edge between node _{i} and node _{j} . Conversely if $G_{ij} = 1$ there is a directed edge from node _{i} to node _{j} . We define the biological prior knowledge matrix B to be a matrix in which the entries $B_{ij} \in [0, 1]$ represent our knowledge about interactions between nodes as follows: If entry $B_{ij} = 0.5$, we do not have any prior knowledge about the presence or absence of the directed edge between node _{i} and node _{j} . If $0 \leq B_{ij} < 0.5$ we have prior evidence that the directed edge between node _{i} and node _{j} is absent. The evidence is stronger as B_{ij} is closer to 0. If $0.5 < B_{ij} \leq 1$ we have prior evidence

that the directed edge pointing from node i to node j is present. The evidence is stronger as B_{ij} is closer to 1. Having defined how to represent a network G and the biological prior knowledge B , we now define the energies associated with the presence and absence of edges as follows:

$$E_0(G) = \sum_{\substack{i,j=1 \\ B_{i,j} < 0.5}}^n |B_{i,j} - G_{i,j}| \quad (3)$$

$$E_1(G) = \sum_{\substack{i,j=1 \\ B_{i,j} > 0.5}}^n |B_{i,j} - G_{i,j}| \quad (4)$$

where n is the total number of nodes.

To integrate the prior knowledge expressed by Equations (3) and (4) into the inference procedure, we follow Imoto et al.¹¹ and define the prior distribution over network structures G to take the form of a Gibbs distribution:

$$P(G|\beta_0, \beta_1) = \frac{e^{-\{\beta_0 E_0(G) + \beta_1 E_1(G)\}}}{Z(\beta_0, \beta_1)} \quad (5)$$

where the partition function is defined as:

$$Z(\beta_0, \beta_1) = \sum_{G \in \mathcal{G}} e^{-\{\beta_0 E_0(G) + \beta_1 E_1(G)\}} \quad (6)$$

Unfortunately, the number of graphs increases super-exponentially with the number of nodes, rendering the computation of Z not viable for large networks. To proceed, we define:

$$E_0(G) = \sum_n \mathcal{E}_0(n, \pi_n[G]) \quad (7)$$

$$E_1(G) = \sum_n \mathcal{E}_1(n, \pi_n[G]) \quad (8)$$

where $\pi_n[G]$ is the set of parents of node n in the graph G and we have defined:

$$\mathcal{E}_0(n, \pi_n) = \sum_{\substack{i \in \pi_n \\ B_{in} < 0.5}} (1 - B_{in}) + \sum_{\substack{i \notin \pi_n \\ B_{in} < 0.5}} B_{in} \quad (9)$$

$$\mathcal{E}_1(n, \pi_n) = \sum_{\substack{i \in \pi_n \\ B_{in} > 0.5}} (1 - B_{in}) + \sum_{\substack{i \notin \pi_n \\ B_{in} > 0.5}} B_{in} \quad (10)$$

Akin to the ideal gas approximation in statistical physics, we now approximate the partition function of the whole network by a product of single-node partition functions:

$$Z \approx \prod_n \sum_{\pi_n} e^{-\{\beta_0 \mathcal{E}_0(n, \pi_n) + \beta_1 \mathcal{E}_1(n, \pi_n)\}} \quad (11)$$

Here, the summation in the last equation extends over all parent configurations π_n of node n , which in the case of a fan-in restriction is subject to constraints on their cardinality. Note that the essence of equation (11) is a dramatic reduction in the computational complexity. Rather than summing over the whole space of network structures, whose cardinality increases super-exponentially with the number of nodes N , we only need to sum over all parent configurations of each node; the complexity of this operation is polynomial in N . However, we have ignored interactions between the nodes; modifications of a parent configuration π_n may lead to a directed cyclic structure, which is invalid and should be excluded from the summation in equation 11. The detection of directed cycles is a global operation. This destroys the modularity inherent in equation 11, and leads to a considerable explosion of the computational complexity. Note, however, that equation 11 still provides an upper bound on the true partition function. When densely connected graphs are ruled out by a fan-in restriction, as commonly done, the number of cyclic terms that need to be excluded from equation 11 can be assumed to be relatively small. We can then expect the bound to be rather tight, and use it to approximate the true partition function. In all our simulations we assumed a fan-in restriction of three, as has widely been applied by different authors^{3, 4, 9}.

2.2. MCMC sampling scheme

Having defined the prior probability distribution over network structures, our next objective is to extend the MCMC scheme of equation 2 to sample both the network structure and the hyperparameters from the posterior distribution.

Starting from a definition of the prior distributions on the hyperparameters β_0 and β_1 , $P(\beta_0)$ and $P(\beta_1)$, our aim is to sample the network structure G and the hyperparameters β_0 and β_1 from the posterior distribution $P(G, \beta_0, \beta_1|D)$. To this end, we propose a new network structure G_{new} from the proposal distribution $Q(G_{\text{new}}|G_{\text{old}})$ and, additionally, new hyperparameters from the proposal distributions $R(\beta_{0\text{new}}|\beta_{0\text{old}})$ and $R(\beta_{1\text{new}}|\beta_{1\text{old}})$. We then accept this move according to the standard Metropolis-Hastings update rule⁷ with the following acceptance

probability:

$$\begin{aligned}
A &= \min \left\{ \frac{P(D, G_{\text{new}}, \beta_{0_{\text{new}}}, \beta_{1_{\text{new}}})}{P(D, G_{\text{old}}, \beta_{0_{\text{old}}}, \beta_{1_{\text{old}}})} \right. \\
&\quad \times \frac{Q(G_{\text{old}}|G_{\text{new}})R(\beta_{0_{\text{old}}}| \beta_{0_{\text{new}}})}{Q(G_{\text{new}}|G_{\text{old}})R(\beta_{0_{\text{new}}}| \beta_{0_{\text{old}}})} \\
&\quad \times \left. \frac{R(\beta_{1_{\text{old}}}| \beta_{1_{\text{new}}})}{R(\beta_{1_{\text{new}}}| \beta_{1_{\text{old}}})}, 1 \right\} \\
&= \min \left\{ \frac{P(D|G_{\text{new}})P(G_{\text{new}}|\beta_{0_{\text{new}}}, \beta_{1_{\text{new}}})}{P(D|G_{\text{old}})P(G_{\text{old}}|\beta_{0_{\text{old}}}, \beta_{1_{\text{old}}})} \right. \\
&\quad \times \frac{P(\beta_{0_{\text{new}}})P(\beta_{1_{\text{new}}})Q(G_{\text{old}}|G_{\text{new}})}{P(\beta_{0_{\text{old}}})P(\beta_{1_{\text{old}}})Q(G_{\text{new}}|G_{\text{old}})} \\
&\quad \times \left. \frac{R(\beta_{0_{\text{old}}}| \beta_{0_{\text{new}}})R(\beta_{1_{\text{old}}}| \beta_{1_{\text{new}}})}{R(\beta_{0_{\text{new}}}| \beta_{0_{\text{old}}})R(\beta_{1_{\text{new}}}| \beta_{1_{\text{old}}})}, 1 \right\} \tag{12}
\end{aligned}$$

To increase the acceptance probability and, hence, mixing and convergence of the Markov chain, it is advisable to break the move up into three submoves:

- Sample a new network structure G_{new} from the proposal distribution $Q(G_{\text{new}}|G_{\text{old}})$ for fixed hyperparameters β_0 and β_1 .
- Sample a new hyperparameter $\beta_{0_{\text{new}}}$ from the proposal distribution $R(\beta_{0_{\text{new}}}| \beta_{0_{\text{old}}})$ for fixed hyperparameter β_1 and fixed network structure G .
- Sample a new hyperparameter $\beta_{1_{\text{new}}}$ from the proposal distribution $R(\beta_{1_{\text{new}}}| \beta_{1_{\text{old}}})$ for fixed hyperparameter β_0 and fixed network structure G .

Assuming uniform prior distributions $P(\beta_0)$ and $P(\beta_1)$ as well as symmetric proposal distributions $R(\beta_{0_{\text{new}}}| \beta_{0_{\text{old}}})$ and $R(\beta_{1_{\text{new}}}| \beta_{1_{\text{old}}})$, the corresponding acceptance probabilities are given by the following expressions:

$$\begin{aligned}
A(G_{\text{new}}|G_{\text{old}}) &= \min \left\{ \frac{P(D|G_{\text{new}})}{P(D|G_{\text{old}})} \right. \\
&\quad \times \frac{P(G_{\text{new}}|\beta_0, \beta_1)}{P(G_{\text{old}}|\beta_0, \beta_1)} \\
&\quad \times \left. \frac{Q(G_{\text{old}}|G_{\text{new}})}{Q(G_{\text{new}}|G_{\text{old}})}, 1 \right\} \tag{13}
\end{aligned}$$

$$A(\beta_{1_{\text{new}}}| \beta_{1_{\text{old}}}) = \min \left\{ \frac{P(G|\beta_{1_{\text{new}}}, \beta_2)}{P(G|\beta_{1_{\text{old}}}, \beta_2)}, 1 \right\} \tag{14}$$

$$A(\beta_{2_{\text{new}}}| \beta_{2_{\text{old}}}) = \min \left\{ \frac{P(G|\beta_1, \beta_{2_{\text{new}}})}{P(G|\beta_1, \beta_{2_{\text{old}}})}, 1 \right\} \tag{15}$$

The two submoves are iterated until some convergence criterion is satisfied, discarding an initial burn-in phase before sampling configurations. In our simulations, we chose the prior distribution of each hyperparameter $P(\beta_i)$, $i \in \{0, 1\}$, to be the uniform distribution over the interval $[0, \text{MAX}]$, with $\text{MAX} = 30$. The proposal distribution of the hyperparameters $R(\beta_{i_{\text{new}}}| \beta_{i_{\text{old}}})$ was chosen to be a uniform distribution over a moving interval of length $L = 6 \ll \text{MAX}$, centred on the current value of the respective hyperparameter and subject to the constraint $\beta_{i_{\text{new}}} \in [0, \text{MAX}]$. Note that L only affects the convergence and mixing of the Markov chain – that is, the computational efficiency – and could, in principle, be adjusted during the burn-in phase. To test for convergence of the MCMC simulations, various methods have been developed¹. In our work, we applied the scheme used in Ref. 23: each MCMC run was repeated from independent initializations, and consistency in the marginal posterior probabilities of the edges was taken as indication of sufficient convergence, leading to a typical trajectory length of 5×10^5 steps, of which the first half was discarded as the burn-in phase.

3. DATA

3.1. Cytometry data

Sachs et al.¹⁹ have applied intracellular multicolour flow cytometry experiments to quantitatively measure protein concentrations related to the RAF pathway. RAF is a critical signalling protein involved in regulating cellular proliferation in human immune system cells. The deregulation of the RAF pathway can lead to carcinogenesis, and this pathway has therefore been extensively studied in the literature^{2, 19}; see Figure 1 for a representation of the currently accepted gold standard network. In our experiments we used 5 data sets with 100 measurements each, obtained by randomly sampling subsets from the original observational (i.e. unintervened) data of Sachs et al.¹⁹. This subsampling was motivated by the fact that we wanted to investigate the learning performance on sample sizes typical of current microarray experiments, which do not provide the abundance of experimental conditions that one gets from cytometry experiments. Details about how we standardized the data can be found in Ref. 23.

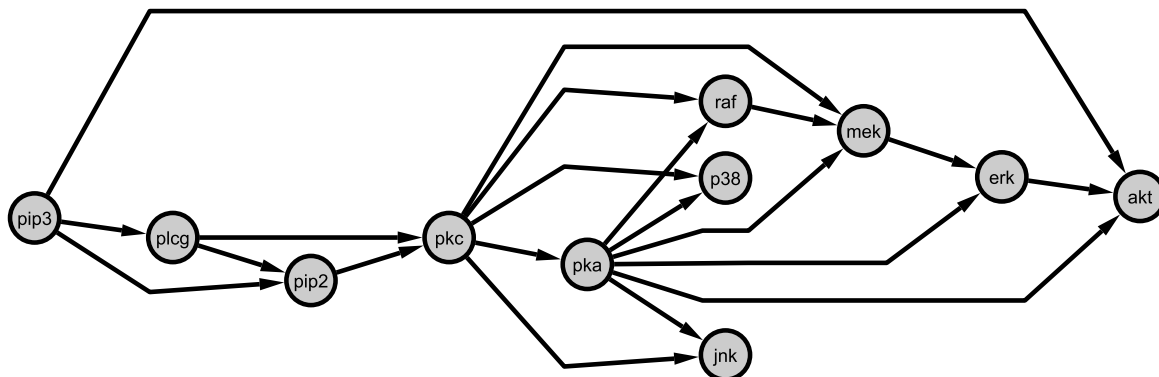


Fig. 1. **RAF signalling pathway.** The graph shows the currently accepted RAF signalling network, taken from Ref. 19. Nodes represent proteins, edges represent interactions, and arrows indicate the direction of signal transduction.

3.2. Synthetic data

A realistic simulation of data typical of signals measured in molecular biology is based on treating the interactions in the network as enzyme-substrate reactions in organic chemistry. From chemical kinetics it is known that the concentrations of the molecules involved in these reactions can be described by a system of ordinary differential equations (ODEs)²⁵. Assuming equilibrium and adopting a steady-state approximation, it is possible to derive a set of closed-form equations that describe the product concentrations as nonlinear (sigmoidal) functions of combinations of substrates. However, instead of solving the steady-state approximation to ODEs explicitly we approximate the solution with a qualitatively equivalent combination of multiplications and sums of sigmoidal transfer functions. The resulting sigma-pi formalism has been implemented in the software package Netbuilder^{26, 27}, which we have used for simulating the data from the RAF signalling pathway, displayed in Figure 1. We used the same amount of data as for the flow cytometry experiments and created 5 simulated data sets with 100 measurements each. To model the stochastic influences, all nodes were subjected to additive Gaussian noise with zero mean and standard deviation equal to 0.1. More details about the generation of these data can be found in Ref. 23.

3.3. Biological prior knowledge

We extracted biological prior knowledge from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database^{13–15}. KEGG pathways represent the current knowledge of the molecular interaction and reaction networks related to metabolism, other cellular processes, and human diseases. As KEGG contains different pathways for different diseases, molecular interactions and types of metabolism, it is possible to find the same pair of genes^a in more than one pathway. We therefore extracted all pathways from KEGG that contained at least one pair of the 11 proteins/phospholipids included in the RAF pathway. We found 20 pathways that satisfied this condition. From these pathways, we computed the prior knowledge matrix, introduced in Section 2.1, as follows. Define by M_{ij} the total number of times a pair of genes i and j appears in a pathway, and by m_{ij} the number of times the genes are connected by a (directed) edge in the KEGG pathway. The elements B_{ij} of the prior knowledge matrix are then defined by

$$B_{ij} = \frac{m_{ij}}{M_{ij}} \quad (16)$$

If a pair of genes is not found in any of the KEGG pathways, we set the respective prior association to $B_{ij} = 0.5$, implying that we have no information about this relationship.

^aWe use the term “gene” generically for all interacting nodes in the network. This may include proteins encoded by the respective genes.

4. SIMULATIONS

4.1. Motivation

As described in Section 3.1, the RAF pathway has been extensively studied in the literature. We therefore have a sufficiently reliable gold standard network for evaluating the results of our inference procedure, as depicted in Figure 1. Additionally, recent work by Sachs et al.¹⁹ provides us with an abundance of protein concentration data from cytometry experiments, and the authors have also demonstrated the viability of learning the regulatory network from these data with Bayesian networks. However, the abundance of cytometry data substantially exceeds that of currently available gene expression data from microarrays. We therefore pursued the approach taken in Ref. 23 and downsampled the data to a sample size representative of current microarray experiments (100 exemplars).

Although the RAF pathway has been extensively studied, we have to appreciate that the published gold standard network only reflects the current state of our knowledge and does not necessarily represent the true biological network. As we will discuss in the final two sections, there are, in fact, indications that the currently accepted gold standard network is incomplete and possibly partially wrong. In order to evaluate the performance of the proposed Bayesian inference scheme on data for which we know the true network, we tested it independently on data generated from the gold standard network with the Net-builder simulator, as described in Section 3.2. Hence, we repeated all evaluations twice: on real cytometry protein concentrations, and on data synthetically generated from the published gold standard network.

As described in Section 3.1, the objective of our study is to assess the viability of the proposed Bayesian inference scheme and to estimate by how much the network reconstruction results improve as a consequence of combining the data with prior knowledge from the KEGG pathway database. To this end, we have compared the results obtained with the methodology described in Section 2 with our earlier results from Werhli et al.²³, where we had evaluated the performance of Bayesian networks (BNs) and Graphical Gaussian models (GGMs, applied as

described in Ref 20) without the inclusion of prior knowledge.

4.2. Reconstructing the regulatory network

While the true network is a directed graph, our reconstruction methods may lead to undirected, directed, or partially directed graphs^b. To assess the performance of these methods, we applied two different criteria. The first approach, referred to as the *undirected graph evaluation* (UGE), discards the information about the edge directions altogether. To this end, the original and learned networks are replaced by their skeletons, where the skeleton is defined as the network in which two nodes are connected by an undirected edge whenever they are connected by any type of edge. The second approach, referred to as the *directed graph evaluation* (DGE), compares the predicted network with the original directed graph. A predicted undirected edge is interpreted as the superposition of two directed edges, pointing in opposite directions. The application of any of the machine learning methods considered in our study leads to a matrix of scores associated with the edges in a network. For BNs sampled from the posterior distribution with MCMC, these scores are the marginal posterior probabilities of the edges. For GGMs, these are partial correlation coefficients. Both scores define a ranking of the edges. This ranking defines a receiver operator characteristics (ROC) curve, where the relative number of true positive (TP) edges is plotted against the relative number of false positive (FP) edges. The results are shown in Figure 2.

5. RESULTS AND DISCUSSION

Figure 2 shows the ROC curves for four different network reconstruction methods: using the prior knowledge from KEGG only, according to (16); learning Bayesian networks and graphical Gaussian models from the protein concentration data alone; and the proposed Bayesian inference scheme for integrating prior knowledge and data. The figure also distinguishes between learning the skeleton of the graph only (UGE: undirected graph evaluation) and considering the direction of the edges also (DGE: di-

^bGGMs are undirected graphs. While BNs are, in principle, directed graphs, partially directed graphs may result as a consequence of equivalence classes, which were briefly discussed in Section 1.

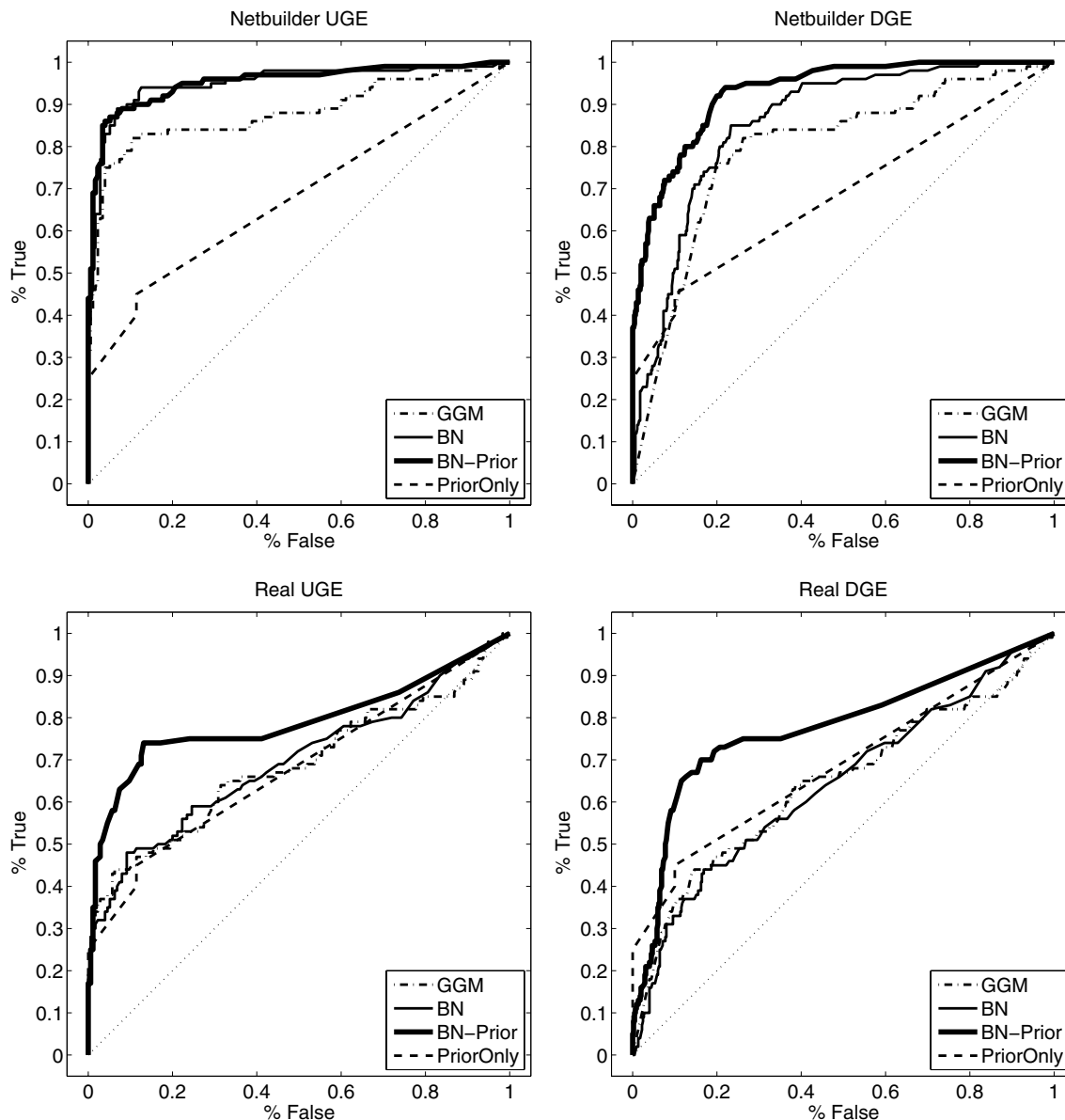


Fig. 2. **Reconstruction of the RAF signalling pathway.** The figure evaluates the accuracy of inferring the RAF signalling network from cytometry data (bottom row) and from simulated Netbuilder data (top row), each combined with prior information from KEGG. This evaluation was carried out twice: with and without taking the edge direction into account (UGE: undirected graph evaluation, left column; DGE: directed graph evaluation, right column). Four machine learning methods were compared: Bayesian Networks without prior knowledge (BNs), Graphical Gaussian Models without prior knowledge (GGMs), Bayesian Networks with prior knowledge from KEGG (BN-Prior), and prior knowledge from KEGG only (PriorOnly). In the latter case, the elements of the prior knowledge matrix (introduced in Section 2.1) were computed from equation (16). The ROC curves presented are the mean ROC curves obtained by averaging the results over five different data sets.

rected graph evaluation). Recall that larger areas under the ROC curves indicate a better prediction performance overall, although the slope on the left is also of interest, as we are usually interested in keeping the number of false positives bounded at low values. The figure suggests that the system-

atic integration of prior knowledge with the proposed Bayesian inference scheme leads, overall, to a considerable improvement in the prediction performance over the three alternative schemes that are based on either the data or the prior knowledge from KEGG alone. There are various interesting trends

synthetic data; see Figure 2, top right. This result is consistent with what has been discussed in the Introduction section: when learning Bayesian networks from non-dynamical non-interventional data (as considered here) without prior knowledge, there is inherent uncertainty about the direction of edges owing to intrinsic symmetries within network equivalence classes; see Section 1. These symmetries are broken by the inclusion of prior knowledge; hence the improvement in the prediction performance. This improvement is also observed on the real cytoflow data (Figure 2, bottom right), but to a lesser extent. Although the area under the ROC curve related to the Bayesian integration scheme exceeds that of all other ROC curves, the prediction based on prior knowledge alone shows a steeper slope in the very left region of the false-positive axis. This implies that for very high values of the threshold on the edge scores, a network learned from prior knowledge alone is more accurate than a network learned with any of the three methods that make use of the data. While the resulting network itself would not be particularly interesting – it would only contain a very small number (3 or 4) of the highest scoring edges – this observation is interesting nevertheless, and can be explained as follows. The discrepancy between the UGE and DGE scores indicates that the Bayesian network learns the skeleton of the graph more accurately than the direction of the interactions, with some of the edge directions systematically inverted. A possible explanation are errors in the gold standard network. The recent literature describes evidence for a negative feedback loop between RAF and ERK via MEK. Active RAF phosphorylates and activates MEK, which, in turn, activates ERK. This corresponds to the directed regulatory path shown in Figure 1. However, through a negative feedback mechanism involving ERK, RAF is phosphorylated on inhibitory sites, generating an inactive, desensitized RAF. Details can be found in Ref. 2. This feedback loop is not included in the gold-standard network reported by Sachs et al.¹⁹, shown in Figure 1. Such as yet unaccounted feedback loops could explain systematic deviations between the predicted and the gold standard network, not only because the structure of a Bayesian network is constrained to be acyclic, but also because we ultimately don't have a reliable gold standard to assess the quality of the predictions. This example points to a fundamental problem inherent in any evaluation based

solely on real biological data, and illustrates clearly the advantage of our combined evaluation based on both laboratory and simulated data.

It is obviously of interest to test how well the inference of the hyperparameters β_0 and β_1 works, especially as this inference depends on the partition function Z of equation (6), which can only be computed approximately; see (11). To this end, we repeated the MCMC simulations for a large set of fixed values of β_0 and β_1 , selected from the grid $[0, 20] \times [0, 20]$. For each pair of fixed values (β_0, β_1) , we sampled BNs from the posterior distribution with MCMC, and evaluated the network reconstruction accuracy using the evaluation criteria described in Section 4.2. We compared these results with the proposed Bayesian inference scheme, where both hyperparameters and networks are simultaneously sampled from the posterior distribution with the MCMC scheme discussed in Section 2.2. The results are shown in Figure 3. The grey shading of the contour plots indicates the network reconstruction accuracy in terms of the directed (DGE: left panels) and undirected (UGE: right panels) graph evaluation, obtained from the synthetic (top panels) and real cytometry data (bottom panels). The black dots show the hyperparameter values sampled with the MCMC simulations. While the distribution of β_0 , the hyperparameter associated with the non-edges, is fairly peaked, the distribution of β_1 , the hyperparameter associated with the edges, is rather diffuse. This diffusion is particularly noticeable on the synthetic data. However, even on the real cytometry data, the distribution of β_1 has a long tail, with values being sampled across the whole permissible spectrum. An inspection of the prior knowledge matrix B extracted from KEGG according to (16) reveals that the prior knowledge associated with the energy function E_1 – equation (4) – accounts for only 25% of the true edges in the gold standard network of Figure 1, while the prior knowledge associated with the energy function E_0 – equation (3) – accounts for 92% of the non-edges. Consequently, it appears that E_0 captures more relevant information for network reconstruction than E_1 , which is reflected by the tighter distribution of the respective hyperparameter. The location of the sampled values of the hyperparameters β_0 and β_1 falls into the region of high network reconstruction scores. This suggests that the proposed Bayesian sampling scheme suc-

- formatics*. Advanced Information and Knowledge Processing. Springer, New York, 2005.
11. S. Imoto, T. Higuchi, T. Goto, S. Kuhara, and S. Miyano. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proceedings IEEE Computer Society Bioinformatics Conference*, (CSB'03):104–113, 2003.
 12. S. Imoto, T. Higuchi, T. Goto, and S. Miyano. Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks. *Statistical Methodology*, 3(1):1–16, January 2006.
 13. M. Kanehisa. A database for post-genome analysis. *Trends Genet*, 13:375–376, 1997.
 14. M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28:27–30, 2000.
 15. M. Kanehisa, S. Goto, M. Hattori, K. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34:D354–357, 2006.
 16. P. J. Krause. Learning probabilistic networks. *Knowledge Engineering Review*, 13:321–351, 1998.
 17. D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.
 18. N. Nariai, S. Kim, S. Imoto, and S. Miyano. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pacific Symposium on Biocomputing*, 9:336–347, 2004.
 19. K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, Vol 308, Issue 5721, 523-529 , 22 April 2005, 308(5721):523–529, 2005.
 20. J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4, 2005. Article 32.
 21. Y. Tamada, H. Bannai, S. Imoto, T. Katayama, M. Kanehisa, and S. Miyano. Utilizing evolutionary information and gene expression data for estimating gene networks with Bayesian network models. *Journal of Bioinformatics and Computational Biology*, 3(6):1295–1313, June 2005.
 22. Y. Tamada, S. Kim, H. Bannai, S. Imoto, K. Tashiro, S. Kuhara, and S. Miyano. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, 19:ii227–ii236, June 2003.
 23. A. V. Werhli, M. Grzegorzcyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, 22(20):2523–2531, 2006.
 24. L. Wernisch and I. Pournara. Reconstruction of gene networks using bayesian learning and manipulation experiments. *Bioinformatics*, 20:2934–2942, 2004.
 25. C.-R. Yang, B. E. Shapiro, E. D. Mjolsness, and G. W. Hatfield. An enzyme mechanism language for the mathematical modeling of metabolic pathways. *Bioinformatics*, 21(6):774–780, 2005.
 26. C. H. Yuh, H. Bolouri, and E. H. Davidson. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279:1896–1902, March 1998.
 27. C. H. Yuh, H. Bolouri, and E. H. Davidson. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development*, 128:617–629, 2001.