Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 15

Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge

Adriano V. Werhli* Dirk Husmeier†

*Biomathematics & Statistics Scotland (BioSS) and Edinburgh University, adriano@bioss.ac.uk

†Biomathematics & Statistics Scotland (BioSS), dirk@bioss.ac.uk

Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge*

Adriano V. Werhli and Dirk Husmeier

Abstract

There have been various attempts to reconstruct gene regulatory networks from microarray expression data in the past. However, owing to the limited amount of independent experimental conditions and noise inherent in the measurements, the results have been rather modest so far. For this reason it seems advisable to include biological prior knowledge, related, for instance, to transcription factor binding locations in promoter regions or partially known signalling pathways from the literature. In the present paper, we consider a Bayesian approach to systematically integrate expression data with multiple sources of prior knowledge. Each source is encoded via a separate energy function, from which a prior distribution over network structures in the form of a Gibbs distribution is constructed. The hyperparameters associated with the different sources of prior knowledge, which measure the influence of the respective prior relative to the data, are sampled from the posterior distribution with MCMC. We have evaluated the proposed scheme on the yeast cell cycle and the Raf signalling pathway. Our findings quantify to what extent the inclusion of independent prior knowledge improves the network reconstruction accuracy, and the values of the hyperparameters inferred with the proposed scheme were found to be close to optimal with respect to minimizing the reconstruction error.

KEYWORDS: gene regulatory networks, Bayesian networks, Bayesian inference, Markov chain Monte Carlo, microarrays, gene expression data, immunoprecipitation experiments, KEGG pathways

---

*Adriano Werhli is supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Dirk Husmeier is supported by the Scottish Executive Environmental and Rural Affairs Department (SEERAD). We are grateful to Peter Ghazal for stimulating discussions on the biological aspects of signal transduction and regulatory networks. We would also like to thank Chris Theobald and Chris Glasbey for helpful comments on the manuscript.

1 Introduction

An important and challenging problem in systems biology is the inference of gene regulatory networks from high-throughput microarray expression data. Various machine learning and statistical methods have been applied to this end, like Bayesian Networks (BNs) (Friedman *et al.*, 2000), Relevance Networks (Butte and Kohane, 2003) and Graphical Gaussian Models (Schäfer and Strimmer, 2005). An intrinsic difficulty with these approaches is that complex interactions involving many genes have to be inferred from sparse and noisy data. This leads to a poor reconstruction accuracy and suggests that the inclusion of complementary information is indispensable (Husmeier, 2003). A promising approach in this direction has been proposed by Imoto *et al.* (2003). The authors formulate the learning scheme in a Bayesian framework. This scheme allows the systematic integration of gene expression data with biological knowledge from other types of postgenomic data or the literature via a prior distribution over network structures. The hyperparameters of this distribution are inferred together with the network structure in a maximum a posteriori sense by maximizing the joint posterior distribution with a heuristic greedy optimization algorithm. As prior knowledge, the authors extracted protein-DNA interactions from the Yeast Proteome Database. The framework has subsequently been applied to a variety of different sources of biological prior knowledge, where gene regulatory networks were inferred from a combination of gene expression data with transcription factor binding motifs in promoter sequences (Tamada *et al.*, 2003), protein-protein interactions (Nariai *et al.*, 2004), evolutionary information (Tamada *et al.*, 2005), and pathways from the KEGG database (Imoto *et al.*, 2006). The objective of the present paper is to complement this work in various respects.

First, we adopt a sampling-based approach to Bayesian inference as opposed to the optimization schemes applied in the work cited above. The latter aims to find the network structure and the hyperparameters that maximize the joint posterior distribution. This approach is appropriate for posterior distributions that are sharply peaked. However, when gene expression data are sparse and noisy and the prior knowledge is susceptible to intrinsic uncertainty as well, this condition is unlikely to be met. In that case, it is more appropriate to follow Madigan and York (1995), Giudici and Castelo (2003) and Friedman and Koller (2003) and sample network structures from the posterior distribution with Markov chain Monte Carlo (MCMC). We pursue the same approach, and additionally sample the hyperparameters associated with the prior distribution from the joint posterior distribution with MCMC.

Second, we aim to obtain a deeper understanding of the proposed mod-

elling and inference scheme. The prior distribution proposed in Imoto *et al.* (2003) takes the form of a Gibbs distribution, in which the prior knowledge is encoded via an energy function, and an inverse temperature hyperparameter determines the weight that is assigned to it. In our study, we have designed a scenario in which the energy takes on a particular form such that computing the marginal posterior distribution over the hyperparameter becomes analytically tractable. This closed-form expression is compared with MCMC simulations on simulated and real-world data for the more general scenario in which the marginal posterior distribution is intractable, elucidating various aspects of the modelling approach.

Third, we extend the approach of Imoto *et al.* (2003) to include more than one energy function. This approach allows the simultaneous inclusion of different sources of prior knowledge, like promoter motifs and KEGG pathways, each modelled by a separate energy. Each energy function is associated with its own hyperparameter. All hyperparameters are sampled from the posterior distribution with MCMC. In this way, the relative weights related to the different sources of prior knowledge are consistently inferred within the Bayesian context, automatically trading off their relative influences in light of the data.

Fourth, we provide a set of independent evaluations of the viability of the Bayesian inference scheme on various synthetic and real-world data, thereby complementing the results of the studies referred to above. In particular, we apply the proposed method to the integration of two independent sources of transcription factor binding locations from immunoprecipitation experiments with microarray gene expression data from the yeast cell cycle, and the integration of KEGG pathways with cytometry experiments for determining protein interactions related to the Raf signalling pathway.

We have organized our paper as follows. In Section 2 we briefly review the methodology of Bayesian networks and present the proposed Bayesian approach to integrating biological prior knowledge into the inference scheme. In Section 3 we investigate the behaviour of the proposed inference scheme on an idealized population of network structures, for which a closed-form expression of the relevant posterior distribution can be obtained. Section 4 presents the synthetic and real data sets that we used for evaluating the performance of the proposed method. Finally, we present our results in Section 5, followed by a concluding discussion in Section 6.

2 Methodology

2.1 Bayesian networks (BNs)

Bayesian networks (BNs) have been introduced to the problem of reconstructing gene regulatory networks from expression data by Friedman *et al.* (2000) and Hartemink *et al.* (2001). In the present section, we present a brief review of the methodological aspects that are relevant to the work presented in our paper. A more comprehensive overview can be obtained from one of the many tutorials that have been written on this subject, like Heckerman (1999) or Husmeier *et al.* (2005).

BNs are directed graphical models for representing probabilistic independence relations between multiple interacting entities. Formally, a BN is defined by a graphical structure G, a family of (conditional) probability distributions F, and their parameters q, which together specify a joint distribution over a set of random variables of interest. The graphical structure G of a BN consists of a set of nodes and a set of directed edges. The nodes represent random variables, while the edges indicate conditional dependence relations. When we have a directed edge from node A to node B, then A is called the parent of B, and B is called the child of A. The structure G of a BN has to be a directed acyclic graph (DAG), that is, a network without any directed cycles. This structure defines a unique rule for expanding the joint probability in terms of simpler conditional probabilities. Let $X_1, X_2, ..., X_N$ be a set of random variables represented by the nodes $i \in \{1, ..., N\}$ in the graph, define $\pi_i[G]$ to be the parents of node i in graph G, and let $X_{\pi_i[G]}$ represent the set of random variables associated with $\pi_i[G]$. Then

$$P(X_1, ..., X_N) = \prod_{i=1}^{N} P(X_i | X_{\pi_i[G]}) \tag{1}$$

When adopting a score-based approach to inference, our objective is to sample model structures G from the posterior distribution

$$P(G|D) \propto P(D|G)P(G) \tag{2}$$

where D is the data, and $P(G)$ is the prior distribution over network structures. The computation of the marginal likelihood $P(D|G)$ requires a marginalization over the parameters q:

$$P(D|G) = \int P(D|q, G)P(q|G)dq \tag{3}$$

in which $P(D|q,G)$ is the likelihood, and $P(q|G)$ is the prior distribution of the parameters. If certain regulatory conditions, discussed in Heckerman (1999), are satisfied and the data are complete, the integral in Equation 3 is analytically tractable. Two function families F that satisfy these conditions are the *multinomial* distribution with a Dirichlet prior (Heckerman *et al.*, 1995) and the *linear Gaussian* distribution with a normal-Wishart prior (Geiger and Heckerman, 1994). The resulting scores $P(D|G)$ are usually referred to as the BDe (discretized data, multinomial distribution) and the BGe (continuous data, linear Gaussian distribution) score. A nonlinear continuous distribution based on heteroscedastic regression has also been proposed (Imoto *et al.*, 2003), although this approach only allows an approximate solution to the integral in Equation 3, based on the Laplace method. Direct sampling from the posterior distribution (Equation 2) is usually intractable, though. Hence, a Markov chain Monte Carlo (MCMC) scheme is adopted (Madigan and York, 1995), which under fairly general regularity conditions is theoretically guaranteed to converge to the posterior distribution of Equation 2 (Hastings, 1970). 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 (Hastings, 1970) scheme with the following acceptance probability:

$$A = \min\left\{\frac{P(D|G_{\text{new}})P(G_{\text{new}})Q(G_{\text{old}}|G_{\text{new}})}{P(D|G_{\text{old}})P(G_{\text{old}})Q(G_{\text{new}}|G_{\text{old}})}, 1\right\} \tag{4}$$

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 Husmeier *et al.* (2005). For dynamic Bayesian networks (discussed in the next subsection) proposal moves are symmetric: $Q(G_{\text{new}}|G_{\text{old}}) = Q(G_{\text{old}}|G_{\text{new}})$. Hence, the proposal probabilities cancel out.

One of the limitations of the approach presented here is the fact that several networks with the same skeleton but different edge directions can have the same marginal likelihood $P(D|G)$, which implies that we cannot distinguish between them on the basis of the data. This equivalence, which is intrinsic to static Bayesian networks (Chickering, 1995), loses information about some edge directions and thus about possible causal interactions between the genes. Moreover, the directed acyclic nature of Bayesian networks renders the modelling of recurrent structures with feedback loops impossible. Both shortcomings can be overcome when time series data are available, which can be analyzed with dynamic Bayesian networks.

2.2 Dynamic Bayesian networks (DBNs)

Consider the left structure in Figure 1, where two genes interact with each other via feedback loops. Note that this structure is not a valid Bayesian network as it violates the acyclicity constraint. When we unfold the network in the left panel of Figure 1 in time, as represented in the right panel of the same figure, we obtain a proper DAG and hence a valid BN again, the so-called Dynamic Bayesian Network (DBN). For more details about DBNs, see Friedman *et al.* (1998); Murphy and Milan (1999) and Husmeier (2003). We want to restrict the number of parameters to ensure they can be properly inferred from the data. For this reason, we model the dynamic process as a homogeneous Markov chain, where the transition probabilities between adjacent time slices are time-invariant. Intra-slice edges are not allowed since they would represent instantaneous 'time-less' interactions. Note that due to the direction of the arrow of time, the symmetry of equivalence classes is broken: the reversal of an edge would imply that an effect is preceding its cause, which is impossible. Summarizing, with DBNs we solve three shortcomings of static BNs: it is possible to model feedback loops, the acyclicity of the graph is automatically guaranteed by construction, and the symmetries within equivalence classes are broken, thereby removing any intrinsic ambiguities. Note, however, that the intrinsic assumption of DBNs is that the data have been generated from a homogeneous Markov chain, which may not hold in practice.

When applying DBNs we need to modify Equation 1 in order to incorporate the first order Markov assumption, which implies that a node $X_i(t)$ at time t has parents $X_{\pi_i[G]}(t-1)$ at time $t-1$:

$$P(X_1, ..., X_N) = \prod_{i=1}^{N} P(X_i(t)|X_{\pi_i[G]}(t-1)) \tag{5}$$

where N is the total number of nodes.

2.3 Biological prior knowledge

As mentioned in the Introduction section, the objective of the present work is to study the integration of biological prior knowledge into the inference of gene regulatory networks. To this end, we need to define a function that measures the agreement between a given network G and the biological prior knowledge that we have at our disposal. We follow the approach proposed by Imoto *et al.* (2003) and call this measure the energy E, borrowing the name from the statistical physics community.

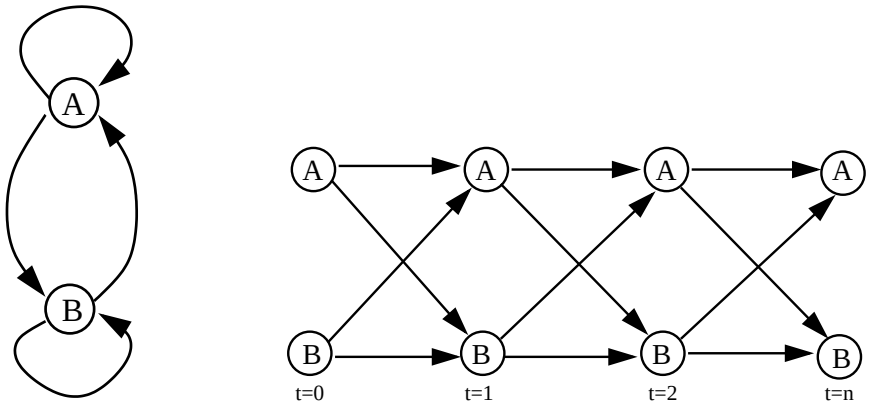


Figure 1: **Dynamic Bayesian Network**: The network on the left is not a proper DAG; the two genes interact with each other via feedback loops. Considering delays between these interactions, it is possible to imagine this network *unfolded in time* where interactions within any time slice t are not permitted. The result is a proper DAG as represented by the graph on the right.

2.3.1 The energy of a network

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 there is no directed edge between node$_i$ and node$_j$. The evidence is stronger as B_{ij} is closer to 0.
- If $0.5 < B_{ij} \leq 1$ we have prior evidence that there is a directed edge pointing from node$_i$ to node$_j$. The evidence is stronger as B_{ij} is closer to 1.

Note that despite their restriction to the unit interval, the B_{ij} are not probabilities in a stochastic sense. To obtain a proper probability distribution over networks, we have to introduce an explicit normalization procedure, as will be discussed shortly.

Having defined how to represent a network G and the biological prior knowledge B, we can now define the 'energy' of a network:

$$E(G) = \sum_{i,j=1}^{N} |B_{i,j} - G_{i,j}| \tag{6}$$

where N is the total number of nodes in the studied domain. The energy E is zero for a perfect match between the prior knowledge B and the actual network structure G, while increasing values of E indicate an increasing mismatch between B and G.

2.3.2 One source of biological prior knowledge

To integrate the prior knowledge expressed by Equation 6 into the inference procedure, we follow Imoto *et al.* (2003) and define the prior distribution over network structures G to take the form of a Gibbs distribution:

$$P(G|\beta) = \frac{e^{-\beta E(G)}}{Z(\beta)} \tag{7}$$

where the energy $E(G)$ was defined in Equation 6, β is a hyperparameter that corresponds to an inverse temperature in statistical physics, and the denominator is a normalizing constant that is usually referred to as the partition function:

$$Z(\beta) = \sum_{G \in \mathcal{G}} e^{-\beta E(G)} \tag{8}$$

Note that the summation extends over the set of all possible network structures $\mathcal{G}$. The hyperparameter β can be interpreted as a factor that indicates the strength of the influence of the biological prior knowledge relative to the data. For $\beta \to 0$, the prior distribution defined in Equation 7 becomes flat and uninformative about the network structure. Conversely, for $\beta \to \infty$, the prior distribution becomes sharply peaked at the network structure with the lowest energy.

For DBNs we can exploit the modularity of Bayesian networks and compute the sum in Equation 8 efficiently. Note that $E(G)$ in Equation 6 can be rewritten as follows:

$$E(G) = \sum_{n=1}^{N} \mathcal{E}\left(n, \pi_n\left[G\right]\right) \tag{9}$$

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

$$\mathcal{E}\left(n, \pi_n\right) = \sum_{i \in \pi_n} \left(1 - B_{in}\right) + \sum_{i \notin \pi_n} B_{in} \tag{10}$$

Inserting Equation 9 into Equation 8 we obtain:

$$\begin{aligned} Z &= \sum_{G\in\mathcal{G}} e^{-\beta E(G)} \\ &= \sum_{\pi_1} \cdots \sum_{\pi_N} e^{-\beta(\mathcal{E}(1,\pi_1)+\ldots+\mathcal{E}(N,\pi_N))} \\ &= \prod_n \sum_{\pi_n} e^{-\beta\mathcal{E}(n,\pi_n)} \end{aligned} \tag{11}$$

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 $\binom{N-1}{m}$ (where m is the maximum fan-in), that is, polynomial in N. The reason for this simplification is the fact that any modification of the parent configuration of a node in a DBN leads to a new valid DBN by construction. This convenient feature does not apply to static BNs, though, where modifications of a parent configuration π_n may lead to directed cyclic structures, which are invalid and hence have to 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, as suggested by Imoto *et al.* (2006), 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; e.g. Friedman *et al.* (2000); Friedman and Koller (2003); Husmeier (2003). We tested the viability of the approximation made for static Bayesian networks in our simulations, to be discussed in Section 5; see especially Figures 16 and 17.

2.3.3 Multiple sources of biological prior knowledge

The method described in the previous section can be generalized to multiple sources of prior knowledge. To keep the notation transparent, we restrict our discussion to two sources of prior knowledge; an extension to more than two

sources is straightforward and follows along the same line of argumentation as presented here. We assume that the biological prior knowledge from each independent source is represented by a separate prior knowledge matrix B^k, $k \in \{1, 2\}$, each satisfying the requirements laid out in the previous section. This gives us two energy functions:

$$E_1(G) = \sum_{i,j=1}^{N} \left| B^1_{i,j} - G_{i,j} \right| \tag{12}$$

$$E_2(G) = \sum_{i,j=1}^{N} \left| B^2_{i,j} - G_{i,j} \right| \tag{13}$$

where each energy is associated with its own hyperparameter β_k. The prior probability of a network G given the hyperparameters β_1 and β_2 is now defined as:

$$P(G|\beta_1, \beta_2) = \frac{e^{-\{\beta_1 E_1(G) + \beta_2 E_2(G)\}}}{Z(\beta_1, \beta_2)} \tag{14}$$

where the partition function in the denominator is given by:

$$Z(\beta_1, \beta_2) = \sum_{G \in \mathcal{G}} e^{-\{\beta_1 E_1(G) + \beta_2 E_2(G)\}} \tag{15}$$

For DBNs, the partition function can again be efficiently computed in closed form. Similarly to the discussion above Equation 11, we can rewrite Equations 12 and 13 as follows:

$$E_1(G) = \sum_{n=1}^{N} \mathcal{E}_1 \left(n, \pi_n \left[G \right] \right) \tag{16}$$

$$E_2(G) = \sum_{n=1}^{N} \mathcal{E}_2 \left(n, \pi_n \left[G \right] \right) \tag{17}$$

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

$$\mathcal{E}_1 \left(n, \pi_n \right) = \sum_{i \in \pi_n} \left(1 - B^1_{in} \right) + \sum_{i \notin \pi_n} B^1_{in} \tag{18}$$

$$\mathcal{E}_2 \left(n, \pi_n \right) = \sum_{i \in \pi_n} \left(1 - B^2_{in} \right) + \sum_{i \notin \pi_n} B^2_{in} \tag{19}$$

$$\tag{20}$$

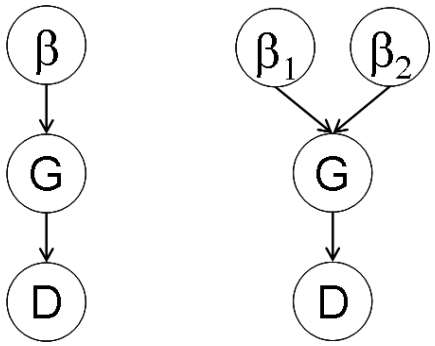


Figure 2: **Probabilistic graphical models.** The two probabilistic graphical models represent conditional independence relations between the data D, the network structure G, and the hyperparameters of the prior on G. The left graph shows the situation of a single source of prior knowledge, with one hyperparameter β. The graph in the right panel shows the situation of two independent sources of prior knowledge, associated with two separate hyperparameters β_1 and β_2. The conditional independence relations can be obtained from the graphs according to the standard rules of factorization in Bayesian networks, as discussed, e.g., in Heckerman (1999). This leads to the following expansions. Left panel: $P(D, G, \beta) = P(D|G)P(G|\beta)P(\beta)$. Right panel: $P(D, G, \beta_1, \beta_2) = P(D|G)P(G|\beta_1, \beta_2)P_1(\beta_1)P_2(\beta_2)$.

Inserting Equations 16 and 17 into Equation 15, we obtain:

$$\begin{aligned} Z &= \sum_{G\in\mathcal{G}} e^{-\{\beta_1 E_1(G)+\beta_2 E_2(G)\}} \\ &= \sum_{\pi_1}\cdots\sum_{\pi_N} e^{-\{\beta_1[\mathcal{E}_1(1,\pi_1)+\ldots+\mathcal{E}_1(N,\pi_N)]+\beta_2[\mathcal{E}_2(1,\pi_1)+\ldots+\mathcal{E}_2(N,\pi_N)]\}} \\ &= \prod_n \sum_{\pi_n} e^{-\{\beta_1\mathcal{E}_1(n,\pi_n)+\beta_2\mathcal{E}_2(n,\pi_n)\}} \end{aligned} \tag{21}$$

For static BNs, this expression provides an upper bound, which can be expected to be tight for strict fan-in restrictions; see the discussion below Equation 11.

2.4 MCMC sampling scheme

Having defined the prior probability distribution over network structures, our next objective is to extend the MCMC scheme of Equation 4 to sample both the network structure and the hyperparameters from the posterior distribution.

2.4.1 MCMC with one source of biological prior knowledge

Starting from a definition of the prior distribution on the hyperparameter β, $P(\beta)$, our aim is to sample the network structure G and the hyperparameter β from the posterior distribution $P(G, \beta|D)$. To this end, we propose a new network structure G_{new} from the proposal distribution $Q(G_{\text{new}}|G_{\text{old}})$ and, additionally, a new hyperparameter from the proposal distribution $R(\beta_{\text{new}}|\beta_{\text{old}})$. We then accept this move according to the standard Metropolis-Hastings update rule (Hastings, 1970) with the following acceptance probability:

$$A = \min\left\{\frac{P(D, G_{\text{new}}, \beta_{\text{new}})Q(G_{\text{old}}|G_{\text{new}})R(\beta_{\text{old}}|\beta_{\text{new}})}{P(D, G_{\text{old}}, \beta_{\text{old}})Q(G_{\text{new}}|G_{\text{old}})R(\beta_{\text{new}}|\beta_{\text{old}})}, 1\right\} \tag{22}$$

which owing to the conditional independence relations depicted in Figure 2 can be expanded as follows:

$$A = \min\left\{\frac{P(D|G_{\text{new}})P(G_{\text{new}}|\beta_{\text{new}})P(\beta_{\text{new}})Q(G_{\text{old}}|G_{\text{new}})R(\beta_{\text{old}}|\beta_{\text{new}})}{P(D|G_{\text{old}})P(G_{\text{old}}|\beta_{\text{old}})P(\beta_{\text{old}})Q(G_{\text{new}}|G_{\text{old}})R(\beta_{\text{new}}|\beta_{\text{old}})}, 1\right\} \tag{23}$$

To increase the acceptance probability and, hence, mixing and convergence of the Markov chain, it is advisable to break the move up into two submoves. First, we sample a new network structure G_{new} from the proposal distribution $Q(G_{\text{new}}|G_{\text{old}})$ while keeping the hyperparameter β fixed, and accept this move with the following acceptance probability:

$$A(G_{\text{new}}|G_{\text{old}}) = \min\left\{\frac{P(D|G_{\text{new}})P(G_{\text{new}}|\beta)Q(G_{\text{old}}|G_{\text{new}})}{P(D|G_{\text{old}})P(G_{\text{old}}|\beta)Q(G_{\text{new}}|G_{\text{old}})}, 1\right\} \tag{24}$$

Next, we sample a new hyperparameter β from the proposal distribution $R(\beta_{\text{new}}|\beta_{\text{old}})$ for a fixed network structure G, and accept this move with the following acceptance probability:

$$A(\beta_{\text{new}}|\beta_{\text{old}}) = \min\left\{\frac{P(G|\beta_{\text{new}})P(\beta_{\text{new}})R(\beta_{\text{old}}|\beta_{\text{new}})}{P(G|\beta_{\text{old}})P(\beta_{\text{old}})R(\beta_{\text{new}}|\beta_{\text{old}})}, 1\right\} \tag{25}$$

For a uniform prior distribution $P(\beta)$ and a symmetric proposal distribution $R(\beta_{\text{new}}|\beta_{\text{old}})$, this expression simplifies:

$$A(\beta_{\text{new}}|\beta_{\text{old}}) = \min\left\{\frac{P(G|\beta_{\text{new}})}{P(G|\beta_{\text{old}})}, 1\right\} \tag{26}$$

The two submoves are iterated until some convergence criterion (Cowles and Carlin, 1996) is satisfied.

2.4.2 MCMC with multiple sources of biological prior knowledge

The scheme presented in the previous section can be extended to multiple sources of prior knowledge. To avoid opacity in the notation, we restrict our discussion to two independent sources of prior knowledge. The generalization to more than two sources is straightforward and follows the same principles as discussed in this section. Starting from two prior distributions on the hyperparameters, $P_1(\beta_1)$ and $P_2(\beta_2)$, our objective is to sample network structures and hyperparameters from the posterior distribution $P(G, \beta_1, \beta_2|D)$. Again, we follow the standard Metropolis-Hastings scheme (Hastings, 1970). We sample a new network structure G_{new} from the proposal distribution $Q(G_{\text{new}}|G_{\text{old}})$, and new hyperparameters from the proposal distributions $R_1(\beta_{1_{\text{new}}}|\beta_{1_{\text{old}}})$ and $R_2(\beta_{2_{\text{new}}}|\beta_{2_{\text{old}}})$. The acceptance probability of this move is:

$$A = \min\left\{\frac{P(D, G_{\text{new}}, \beta_{1_{\text{new}}}, \beta_{2_{\text{new}}})Q(G_{\text{old}}|G_{\text{new}})R_1(\beta_{1_{\text{old}}}|\beta_{1_{\text{new}}})R_2(\beta_{2_{\text{old}}}|\beta_{2_{\text{new}}})}{P(D, G_{\text{old}}, \beta_{1_{\text{old}}}, \beta_{2_{\text{old}}})Q(G_{\text{new}}|G_{\text{old}})R_1(\beta_{1_{\text{new}}}|\beta_{1_{\text{old}}})R_2(\beta_{2_{\text{new}}}|\beta_{2_{\text{old}}})}, 1\right\} \tag{27}$$

From the conditional independence relations depicted in Figure 2, this expression can be expanded as follows:

$$\begin{aligned} A = \min\Bigg\{&\frac{P(D|G_{\text{new}})P(G_{\text{new}}|\beta_{1_{\text{new}}}, \beta_{2_{\text{new}}})P_1(\beta_{1_{\text{new}}})P_2(\beta_{2_{\text{new}}})}{P(D|G_{\text{old}})P(G_{\text{old}}|\beta_{1_{\text{old}}}, \beta_{2_{\text{old}}})P_1(\beta_{1_{\text{old}}})P_2(\beta_{2_{\text{old}}})} \times \\ &\frac{Q(G_{\text{old}}|G_{\text{new}})R_1(\beta_{1_{\text{old}}}|\beta_{1_{\text{new}}})R_2(\beta_{2_{\text{old}}}|\beta_{2_{\text{new}}})}{Q(G_{\text{new}}|G_{\text{old}})R_1(\beta_{1_{\text{new}}}|\beta_{1_{\text{old}}})R_2(\beta_{2_{\text{new}}}|\beta_{2_{\text{old}}})}, 1\Bigg\} \end{aligned} \tag{28}$$

As discussed in the previous section, it is advisable to break this 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_1 and β_2.
- Sample a new hyperparameter $\beta_{1_{\text{new}}}$ from the proposal distribution $R_1(\beta_{1_{\text{new}}}|\beta_{1_{\text{old}}})$ for fixed hyperparameter β_2 and fixed network structure G.
- Sample a new hyperparameter $\beta_{2_{\text{new}}}$ from the proposal distribution $R_2(\beta_{2_{\text{new}}}|\beta_{2_{\text{old}}})$ for fixed hyperparameter β_1 and fixed network structure G.

Assuming uniform prior distributions $P_1(\beta_1)$ and $P_2(\beta_2)$ as well as symmetric proposal distributions $R_1(\beta_{1_{\text{new}}}|\beta_{1_{\text{old}}})$ and $R_2(\beta_{2_{\text{new}}}|\beta_{2_{\text{old}}})$, the corresponding

acceptance probabilities are given by the following expressions:

$$A(G_{\text{new}}|G_{\text{old}}) = \min\left\{\frac{P(D|G_{\text{new}})P(G_{\text{new}}|\beta_1,\beta_2)Q(G_{\text{old}}|G_{\text{new}})}{P(D|G_{\text{old}})P(G_{\text{old}}|\beta_1,\beta_2)Q(G_{\text{new}}|G_{\text{old}})},1\right\} \tag{29}$$

$$A(\beta_{1\text{new}}|\beta_{1\text{old}}) = \min\left\{\frac{P(G|\beta_{1\text{new}},\beta_2)}{P(G|\beta_{1\text{old}},\beta_2)},1\right\} \tag{30}$$

$$A(\beta_{2\text{new}}|\beta_{2\text{old}}) = \min\left\{\frac{P(G|\beta_1,\beta_{2\text{new}})}{P(G|\beta_1,\beta_{2\text{old}})},1\right\} \tag{31}$$

2.4.3 Practical issues

In our simulations, we chose the prior distribution of the hyperparameters $P(\beta)$ to be the uniform distribution over the interval $[0, \text{MAX}]$. The proposal probability for the hyperparameters $R(\beta_{\text{new}}|\beta_{\text{old}})$ was chosen to be a uniform distribution over a moving interval of length $2l \ll \text{MAX}$, centred on the current value of the hyperparameter. Consider a hyperparameter β_{new} to be sampled in an MCMC move given that we have the current value β_{old}. The proposal distribution is uniform over the interval $[\beta_{\text{old}} - l, \beta_{\text{old}} + l]$ with the constraint that $\beta_{\text{new}} \in [0, \text{MAX}]$. If the sampled value β_{new} happens to lie outside the allowed interval, the value is reflected back into the interval. The respective proposal probabilities can be shown to be symmetric and therefore to cancel out in the acceptance probability ratio. In our simulations, we set the upper limit of the prior distribution to be $\text{MAX} = 30$, and the length of the sampling interval to be $l = 3$. Note that the choice of l only affects the convergence and mixing of the Markov chain, but has theoretically no influence on the results. While an adaptation of this parameter during burn-in could be attempted to optimize the computational efficiency of the scheme, we found that the chosen value of l gave already a fast convergence of the Markov chain that we did not deem necessary to further improve.

To test for convergence of the MCMC simulations, various methods have been developed; see Cowles and Carlin (1996) for a review. In our work, we applied the simple scheme used in Friedman and Koller (2003): 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. For the applications reported in Section 5, this led to the decision to run the MCMC simulations for a total number of 5×10^5 steps, of which the first half were discarded as the burn-in phase.

3 Simulations

The objective of this section is to explore the posterior probability landscape in the space of hyperparameters. This will help us to better interpret the values of the hyperparameters sampled with MCMC in real applications, and to assess whether these values are plausible. We pursue this objective with two different approaches. In the first approach, we design a hypothetical population of network structures for which we can analytically derive a closed-form expression of the partition function and, hence, the marginal posterior probability of the hyperparameters. These results will be presented in Subsections 3.1 and 3.3 for one and multiple sources of prior knowledge, respectively. In the second approach, we focus on a small network with a limited number of nodes. Although we cannot derive a closed-form expression for the partition function in this case, we can compute the partition function numerically via an exhaustive enumeration of all possible network structures; this again allows us to compute the marginal posterior probability of the hyperparameters. The resulting posterior probability landscapes will be presented in Subsections 3.2 and 3.4, again for one and multiple sources of prior knowledge, respectively. We compare these results with the values of hyperparameters sampled from an MCMC simulation; this approximate numerical procedure is the only approach that is viable in real-world applications with many interacting nodes.

3.1 Idealized derivation for one source of biological prior knowledge

Consider the partition of a hypothetical space of network structures, depicted in Figure 3. This Venn diagram consists of four mutually exclusive subsets, which represent networks that are characterized by different compatibilities with respect to the data and the prior knowledge. We make the idealizing assumption that the networks either completely succeed or fail in modelling the data. The networks are also assumed to be either completely consistent or inconsistent with the assumed prior knowledge. The different sizes of the subsets are related to the relative proportions of the networks they contain, which are described by the following quantities:

- **TD**: Proportion of networks that are in agreement with the data only.
- **TD1**: Proportion of networks that are in agreement with the data and with the prior.
- **T1**: Proportion of networks that are in agreement with the prior only.

Graph in agreement with:		Result		
Data	Prior	P(D|G)	E	Proportion
no	no	a	1	F
no	yes	a	0	T1
yes	no	A	1	TD
yes	yes	A	0	TD1

Table 1: **Idealized scenario for one source of prior.** This table summarizes the definitions for the idealized population of network structures when considering one source of biological prior knowledge, corresponding to the Venn diagram of Figure 3.

- **F**: Proportion of networks that are neither in agreement with the data nor with the prior.

We define that networks that are in agreement with the data have marginal likelihood $P(D|G) = A$, while those in disagreement with the data have the lower marginal likelihood $P(D|G) = a$, with $a < A$. In our experiments discussed below, we set $A = 10$ and $a = 1$. A network that is in accordance with the biological prior knowledge has zero energy $E = 0$; otherwise, the network is penalized with a higher energy of $E = 1$. Table 1 presents a summary of these definitions. We want to find the posterior distribution $P(\beta|D)$:

$$P(\beta|D) \;=\; \frac{1}{P(D)} \sum_G P(D, G, \beta) \tag{32}$$

The conditional independence relations, represented by the graphical model in the left panel of Figure 2, imply that

$$P(D, G, \beta) \;=\; P(D|G)P(G|\beta)P(\beta) \tag{33}$$

Assuming a uniform prior over β, we thus obtain

$$P(\beta|D) \;\propto\; \sum_G P(D|G)P(G|\beta) \tag{34}$$

Inserting the expression for the prior distribution, Equations 7-8, into this sum, we get:

$$\sum_G P(D|G)P(G|\beta) \;=\; \frac{\sum_G P(D|G)e^{-\beta E(G)}}{\sum_G e^{-\beta E(G)}} \tag{35}$$

Using the definitions from Table 1, we thus obtain the following expression for the posterior distribution $P(\beta|D)$:

$$P(\beta|D) \;\propto\; \frac{a \times T1 + A \times TD1 + e^{-\beta}(a \times F + A \times TD)}{TD1 + T1 + e^{-\beta}(F + TD)} \tag{36}$$

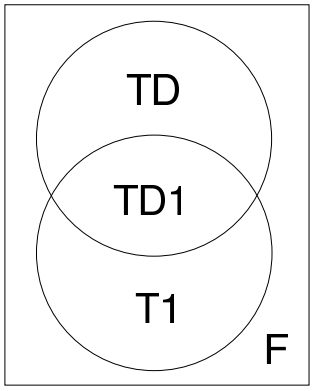


Figure 3: **Venn diagram for an idealized population of network structures and one source of prior knowledge.** The Venn diagram shows a hypothetical population of network structures. We make the idealizing assumption that the networks either completely succeed or fail in modelling the data. The networks are also assumed to be either completely consistent or inconsistent with the assumed prior knowledge. TD is the proportion of graphs that agree with the data. TD1 is the proportion of graphs that agree with the data and the biological prior knowledge. T1 is the proportion of graphs that agree with the biological prior knowledge only. F is the proportion of graphs that are neither in agreement with the data nor with the biological prior knowledge. A summary of this scenario is provided in Table 1.

where we refer to the expression on the right as the unnormalized posterior distribution. A plot of this distribution is shown in the left panel of Figure 6.

3.2 Simulation results for one source of prior knowledge

The objective of this subsection is to compare the closed form of the posterior distribution $P(\beta|D)$ from Equation 36 with that obtained from a synthetic study using real Bayesian networks. To this end, we consider a Bayesian network with a small number of nodes such that a complete enumeration of all possible network structures is possible. This allows the partition function in Equation 8 and hence the posterior distribution $P(\beta|D)$ to be computed exactly, the latter via Equations 7 and 34. We consider the two extreme scenarios of completely correct and completely wrong prior knowledge. For the idealized network population, the situation of completely correct prior knowledge is depicted in the Venn diagram on the left of Figure 4: all networks that accord with the prior also accord with the data, while networks not according with the prior also fail to accord with the data. The Venn diagram on the right of

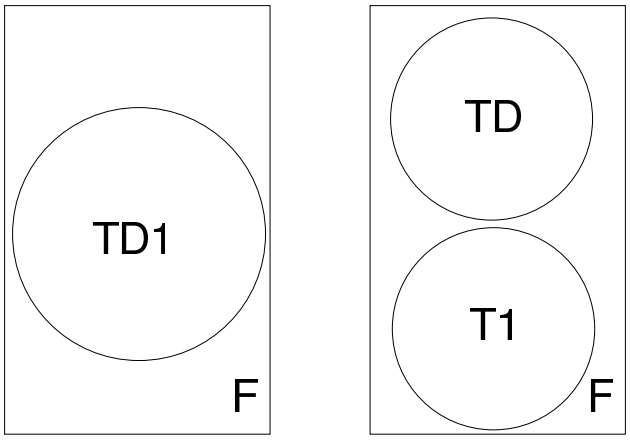


Figure 4: **Venn diagrams for a completely correct and a completely wrong source of biological prior knowledge.** The two Venn diagrams show special scenarios of the hypothetical network population depicted in Figure 3. The left panel represents the situation of completely correct prior knowledge. All networks that are consistent with the data also accord with the prior, and all networks that are in accordance with the prior also agree with the data. Hence $T1 = TD = 0$. The right panel shows the situation of a completely wrong source of prior knowledge. Networks that are consistent with the data are not supported by the prior, while networks that are in agreement with the prior contradict the findings in the data. Hence $TD1 = 0$. (For a definition of the symbols, see Table 1 and the caption of Figure 3).

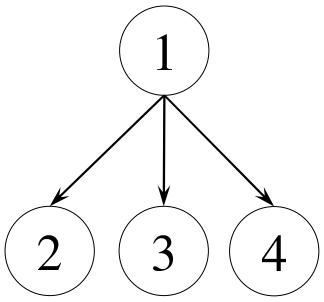


Figure 5: **HUB network.** This figure shows the network structure from which we generated data for the synthetic inference study.

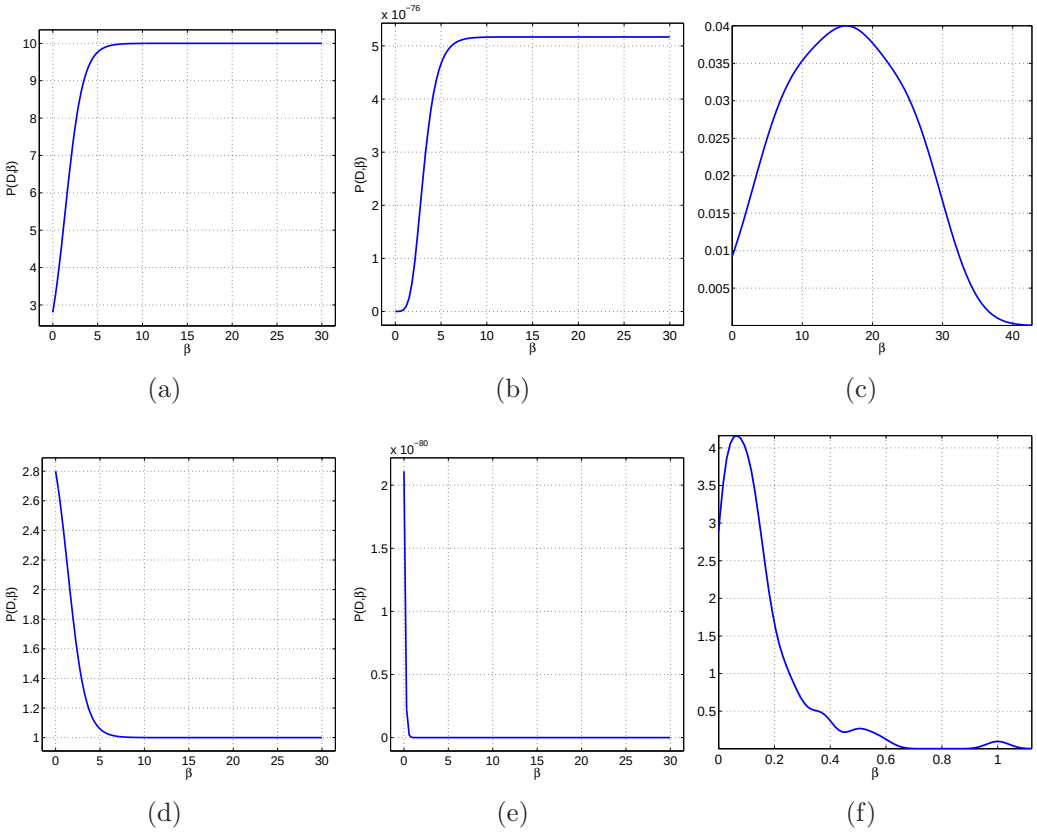


Figure 6: **Results of the simulation study for a single source of prior knowledge.** The top row shows the results when including the correct prior knowledge. The bottom row shows the results when the prior knowledge is wrong. The left column shows the unnormalized posterior probability of the hyperparameter β for the idealized network population depicted in Figure 4, computed from Equation 36 and plotted against β. The values of the network population proportions, defined in Table 1 and Figure 3, were set as follows. Correct prior (corresponding to the left panel in Figure 4): $TD = T1 = 0, TD1 = 0.2$. Wrong prior (corresponding to the right panel in Figure 4): $TD = T1 = 0.2, TD1 = 0$. The centre column shows the unnormalized posterior probability of β for the synthetic toy problem, plotted against β. For comparison, the right column shows the marginal posterior probability densities of β, estimated from the MCMC trajectories with a Parzen estimator, using a Gaussian kernel whose standard deviation was set automatically by the MATLAB function `ksdensity.m`. The MCMC scheme was discussed in Section 2.4.

Figure 4 depicts the opposite scenario of completely wrong prior knowledge: networks that accord with the data never accord with the prior while, conversely, networks that accord with the prior never accord with the data. For the synthetic toy problem, the completely correct prior corresponds to a prior knowledge matrix B that is identical to the true adjacency matrix G of the network (see Section 2.3 for a reminder of this terminology). On the contrary, completely wrong prior knowledge corresponds to a prior knowledge matrix B that is the complete complement of the network adjacency matrix G, that is, has entries indicating edges where there are none in the true network and, conversely, has zero entries for the locations of the true edges in the network.

The network that we used for the synthetic toy problem is shown in Figure 5. We treated it as a DBN and generated a time series of 100 exemplars from it, as described in Section 4.1. The results are shown in Figure 6, where the top row corresponds to the true prior, and the bottom row to the wrong prior. The left and centre columns show plots of the (unnormalized) posterior distribution of the hyperparameter β for the idealized network population and the synthetic toy problem, respectively. The graphs are similar, as expected. In both cases, when the prior is correct, $P(\beta|D)$ monotonically increases until it reaches a plateau. When the prior is wrong, $P(\beta|D)$ peaks at zero, and monotonically decreases for increasing values of β. For comparison, the right column shows the marginal posterior probability densities of β estimated from the MCMC trajectories. The MCMC scheme was discussed in Section 2.4. All results are consistent in indicating that for the true prior, high values of β are encouraged, while for the wrong prior, high values of β are suppressed. Since β represents the weight that is assigned to the prior, our finding confirms that the proposed methodology is working as expected. It also lays the foundations for investigating the more complex scenario of multiple sources of prior knowledge, to be discussed next.

3.3 Idealized derivation for two sources of biological prior knowledge

Next, we generalize the scenario of Subsection 3.1 to two independent sources of prior knowledge. Again, consider a hypothetical space of network structures, which is assumed to be partitioned into distinct regions, as depicted by the Venn diagram of Figure 7. The symbols in this diagram indicate the proportions of networks that fall into the respective regions:

- **TD** is the proportion of graphs that are in agreement with the data only.

- **TD1** is the proportion of graphs that are in agreement with the data and with the first source of prior knowledge.
- **T1** is the proportion of graphs that are in agreement with the first source of prior knowledge only.
- **T2** is the proportion of graphs that are in agreement with the second source of prior knowledge only.
- **TD2** is the proportion of graphs that are in agreement with the data and with the second source of prior knowledge.
- **TD12** is the proportion of graphs that are in agreement with the data and with both sources of prior knowledge.
- **T12** is the proportion of graphs that are in agreement with both sources of prior knowledge, but not the data.
- **F** is the proportion of graphs that are neither in agreement with the data, nor with any prior.

We define that networks that are in agreement with the data have marginal likelihood $P(D|G) = A$, while networks not in agreement with the data have the lower marginal likelihood $P(D|G) = a$, with $a < A$. In our experiments we set $A = 10$ and $a = 1$. Networks that are in accordance with the first source of prior knowledge have energy $E_1 = 0$, otherwise the energy is $E_1 = 1$. Networks that are in accordance with the second source of prior knowledge have energy $E_2 = 0$, otherwise the energy is $E_2 = 1$. Table 2 presents a summary of these definitions. Generalizing the derivation presented in Subsection 3.1, we now want to find the posterior distribution of both hyperparameters $P(\beta_1, \beta_2|D)$:

$$P(\beta_1, \beta_2|D) \;=\; \frac{1}{P(D)} \sum_G P(\beta_1, \beta_2, D, G) \tag{37}$$

From the conditional independence relations depicted by the graphical model in the right panel of Figure 2, we get:

$$P(D, G, \beta_1, \beta_2) \;=\; P(D|G)P(G|\beta_1, \beta_2)P_1(\beta_1)P_2(\beta_2) \tag{38}$$

Assuming uniform priors over the two hyperparameters β_1 and β_2, we obtain:

$$P(\beta_1, \beta_2|D) \;\propto\; \sum_G P(D|G)P(G|\beta_1, \beta_2) \tag{39}$$

Graph in agreement with:			Result			
Data	Prior 1	Prior 2	P(D|G)	E_1	E_2	Proportion
no	no	no	a	1	1	F
no	no	yes	a	1	0	T2
no	yes	no	a	0	1	T1
no	yes	yes	a	0	0	T12
yes	no	no	A	1	1	TD
yes	no	yes	A	1	0	TD2
yes	yes	no	A	0	1	TD1
yes	yes	yes	A	0	0	TD12

Table 2: **Idealized scenario for two independent sources of prior knowledge.** This table summarizes the definitions for the idealized population of network structures with two sources of prior knowledge, corresponding to Figure 7.

Inserting the expression for the prior, Equations 14-15, into this sum, we get:

$$\sum_G P(D|G)P(G|\beta_1,\beta_2) = \frac{\sum_G P(D|G)e^{[-\beta_1 E_1(G)-\beta_2 E_2(G)]}}{\sum_G e^{[-\beta_1 E_1(G)-\beta_2 E_2(G)]}} \tag{40}$$

Using the definitions from Table 2, this yields:

$$P(\beta_1,\beta_2|D) \propto \tag{41}$$
$$\frac{e^{-\beta_2}(a[T1]+A[TD1])+e^{-\beta_1}(a[T2]+A[TD2])+e^{(-\beta_1-\beta_2)}(a[F]+A[TD])+a[T12]+A[TD12]}{e^{-\beta_2}(T1+TD1)+e^{-\beta_1}(T2+TD2)+e^{(-\beta_1-\beta_2)}(TD+F)+TD12+T12}$$

where, again, we refer to the expression on the right as the unnormalized posterior distribution of the hyperparameters. A plot of this distribution is shown in the top left panel of Figure 9.

3.4 Simulation results for two sources of prior knowledge

We revisit the simulations discussed in Subsection 3.2, where we have considered two sources of prior knowledge, one being correct and the other being completely wrong. Rather than studying the effects of these priors in isolation, we now combine them and integrate them simultaneously into the inference scheme. For the idealized population of network structures, the situation is illustrated in Figure 8. The posterior probability distribution of the two hyperparameters is computed from Equation 41, using the parameter setting stated in the captions of Figures 8 and 9. For the synthetic toy problem, the prior

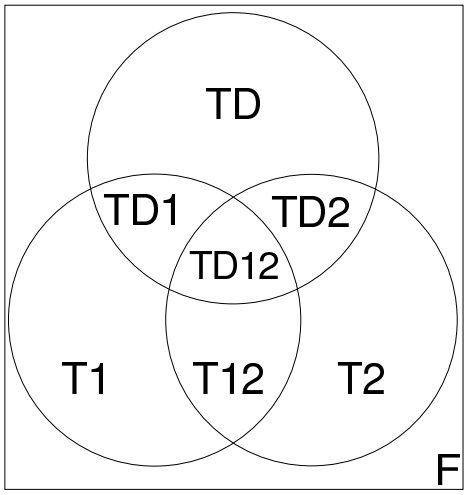


Figure 7: **Venn diagram for an idealized population of network structures and multiple sources of prior knowledge.** This Venn diagram is a generalization of Figure 3 for two independent sources of prior knowledge. TD is the proportion of networks that agree with the data. TD1 is the proportion of networks that agree with the data and prior 1. T1 is the proportion of networks that agree with prior 1 only. TD2 is the proportion of networks that agree with the data and prior 2. T2 is the proportion of networks that agree with prior 2 only. TD12 is the proportion of networks that agree with the data and with both priors. T12 is the proportion of networks that agree with both priors but not the data. F is the proportion of networks that are neither in agreement with the data nor the biological prior knowledge. A summary of this scenario can be found in Table 2.

probability distribution over network structures is computed from Equation 14, obtaining the partition function of Equation 15 from a complete enumeration of all possible network structures. The posterior distribution of the hyperparameters is then computed from Equation 39, again resorting to a complete enumeration of network structures. For comparison, we also sampled the hyperparameters from the posterior distribution numerically, using the MCMC scheme described in Section 2.4.2. The results are shown in Figure 9. The bottom left panel shows the trace plots from the MCMC simulation. The values of β_2, the hyperparameter associated with the wrong prior, are always below those of β_1, the hyperparameter associated with the true prior. This confirms our expectation that the inference scheme succeeds in distinguishing between the different priors and automatically associates a higher weight with the correct prior. Somewhat counterintuitively, though, the value of β_2 does not decay to zero, suggesting that the second prior, despite the worst-case

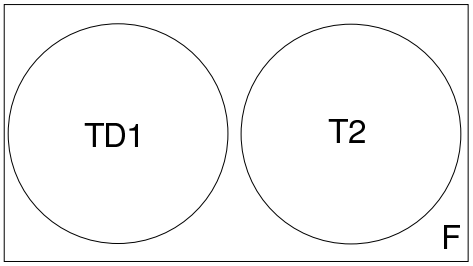


Figure 8: **Venn diagram for a completely correct and a completely wrong source of biological prior knowledge.** This Venn diagram shows a special case of Figure 7 where one source of biological prior knowledge is in complete agreement with the data while the other source of prior knowledge is completely wrong. All networks that are consistent with the data also accord with the first prior, and all networks that are in accordance with the first prior also agree with the data. Hence $T1 = TD = 0$. Networks that are consistent with the data are not supported by the second prior, while networks that are in agreement with the second prior contradict the findings in the data. Hence $TD2 = TD12 = 0$. The priors are also mutually exclusive: $T12 = 0$. Note that the scenario depicted here effectively combines the two scenarios of Figure 4. See Table 2 and the caption of Figure 7 for a definition of the symbols.

scenario of it being completely wrong, is never 'switched off' completely. This seemingly strange behaviour was also consistently found in our MCMC simulations on the real data – see the discussion in Section 5.2.2 – and provided the motivation for the synthetic simulation study discussed in the present section. An elucidation of this behaviour is obtained from the plots of the posterior distribution $P(\beta_1, \beta_2|D)$ in the left and right top panels of Figure 9. Both graphs indicate that $P(\beta_1, \beta_2|D)$ contains a ridge parallel to the line $\beta_1 = \beta_2$, dropping to zero for $\beta_1 < \beta_2$, and reaching a plateau for $\beta_1 > \beta_2$. This plateau explains the results found in our MCMC simulations. When β_1 is sufficiently larger than β_2, corresponding to a configuration on the plateau well over the ridge, there is no effective force pushing β_2 down to zero. The intuitive explanation is that for β_1 sufficiently larger than β_2, the effect of the second (wrong) prior is already negligible, so that it becomes obsolete to completely switch it off.

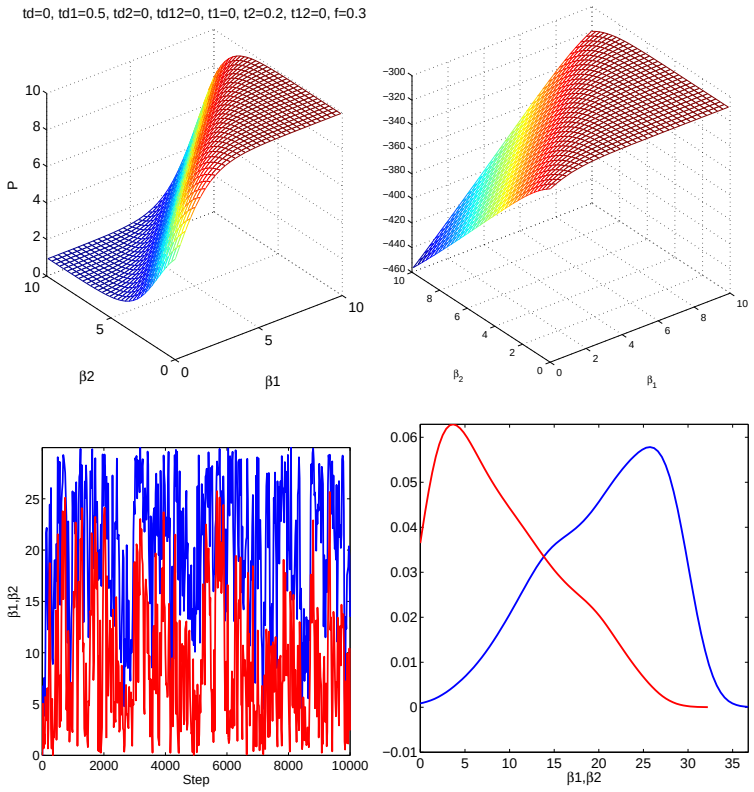


Figure 9: **Results of the simulation study for multiple sources of prior knowledge.** This figure shows the inference results for two independent sources of prior knowledge, associated with separate hyperparameters β_1 and β_2. The top left panel shows a plot of the unnormalized posterior probability distribution of β_1 and β_2 for the idealized population of network structures depicted in Figure 8. The expression was computed from Equation 41 with the following parameter settings: $TD1 = 0.5, T2 = 0.2, F = 0.3$, $TD = TD2 = TD12 = T1 = T12 = 0$ (see the caption of Figure 8 for an explanation of why the parameters were chosen in that way). The top right panel shows a plot of the unnormalized posterior distribution of β_1 and β_2 for the synthetic toy problem. The bottom left panel shows two trace plots obtained when sampling the two hyperparameters from the posterior distribution with the MCMC scheme discussed in Section 2.4.2. The horizontal axis represents the MCMC step while the vertical axis shows the sampled values of the hyperparameters. The bottom right panel shows the marginal posterior probability densities of β_1 and β_2, estimated from the MCMC trajectories with a Parzen estimator, using a Gaussian kernel whose standard deviation was set automatically by the MATLAB function `ksdensity.m`. The blue graph corresponds to β_1, the hyperparameter associated with the true prior. The red graph corresponds to β_2, the hyperparameter associated with the wrong prior.

4 Data and priors

4.1 Simulated data

The data generated for the synthetic simulations described in Section 3 were obtained from a DBN with a linear Gaussian distribution. The random variable $X_i(t+1)$ denoting the expression of node i at time $t+1$ is distributed according to

$$X_i(t+1) \ \sim \ N(\sum\nolimits_k w_{ik} x_k(t), \sigma^2) \tag{42}$$

where $N(.)$ denotes the Normal distribution, the sum extends over all parents of node i, and $x_k(t)$ represents the value of node k at time t. We set the standard deviation to $\sigma = 0.1$, and the interaction strengths to $w_{ik} = 1$. The structure of the network from which we generated data is represented in Figure 5.

4.2 Yeast cell cycle

For the evaluation of the proposed inference method, we were guided by the study of Bernard and Hartemink (2005). The authors aimed to infer regulatory networks involving 25 genes of yeast (*Saccharomyces cerevisiae*), of which 10 genes encode known transcription factors (TFs). The inference was based on gene expression data, combined with prior knowledge about transcription factor binding locations. The gene expression data were obtained from Spellman *et al.* (1998); this data set contains 73 time points collected over 8 cycles of the yeast cell cycle using four different synchronization protocols. The prior knowledge about transcription factor binding locations was obtained from the chromatin immunoprecipitation (ChIP-on-chip) assays of Lee *et al.* (2002).

In our study, we followed the approach of Bernard and Hartemink (2005), but complemented their evaluation by the inclusion of additional gene expression data and a separate source of prior knowledge. As further gene expression data we included the results of microarray experiments carried out by Tu *et al.* (2005); this data set contains 36 time points of gene expression data in yeast, collected over three consecutive metabolic cycles in intervals of 25 minutes. As additional prior knowledge, we included the TF binding locations obtained from an independent chromatin immunoprecipitation assay, reported in Harbison *et al.* (2004). In order to include these binding locations in the proposed inference scheme, we transformed the p-values obtained from the immunoprecipitation assays into probabilities, using the transformation proposed by Bernard and Hartemink (2005). These probabilities formed the entries B_{ij} of

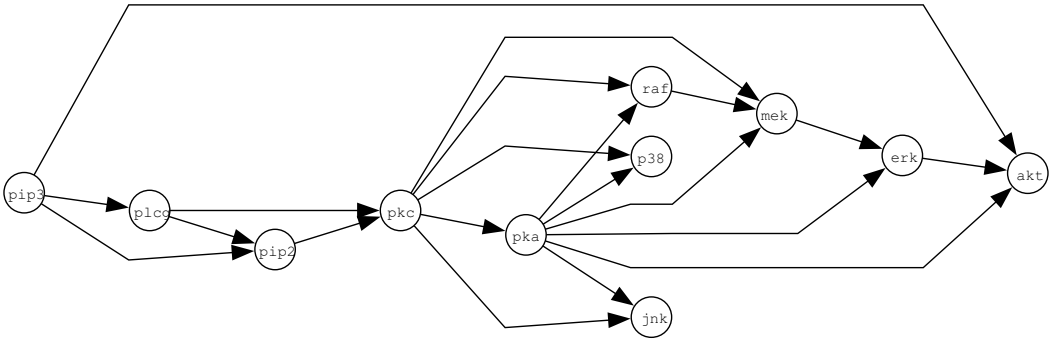


Figure 10: **Raf signalling pathway.** The graph shows the currently accepted Raf signalling network, taken from Sachs *et al.* (2005). Nodes represent proteins, edges represent interactions, and arrows indicate the direction of signal transduction.

our biological prior knowledge matrix. However, only 10 of the 25 studied genes are known to be TFs. For the remaining genes, no information about binding locations is available. The respective entries in the prior knowledge matrix were thus set to $B_{ij} = 0.5$, corresponding to the absence of prior information (see the discussion in Section 2.3).

Summarizing, we evaluated the performance of the proposed inference scheme on two sets of gene expression data and two sets of TF binding location indications. An overview is given in Table 3.

4.3 Raf signalling pathway

Sachs *et al.* (2005) have applied intracellular multicolour flow cytometry experiments to quantitatively measure protein concentrations. Data were collected after a series of stimulatory cues and inhibitory interventions targeting specific proteins in 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 (e.g. Sachs *et al.* (2005); Dougherty *et al.* (2005)); see Figure 10 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 data of Sachs *et al.* (2005). Details about how we standardized the data can be found in Werhli *et al.* (2006).

We extracted biological prior knowledge from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database (Kanehisa, 1997; Kanehisa and Goto, 2000; Kanehisa *et al.*, 2006). KEGG pathways represent cur-

	Expression Data	**1st source of Prior**	**2nd source of Prior**
1	Spellman	Lee	Harbison
2	Tu	Lee	Harbison
3	Spellman	Lee	MCMC Tu
4	Tu	Lee	MCMC Spellman

Table 3: **Yeast evaluation settings.** This table summarizes the evaluation procedures we used on the yeast data. The table shows the name of the first author of the data sets that we used. Gene expression data: Spellman *et al.* (1998) and Tu *et al.* (2005). TF binding location assays: Lee *et al.* (2002) and Harbison *et al.* (2004). The entries *MCMC Spellman* and *MCMC Tu* indicate that the prior knowledge matrix was composed of the marginal posterior probabilities of directed pairwise gene interactions (edges) obtained from running MCMC simulations without prior knowledge on the respective expression data set.

rent 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[1] 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 2.3, 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

$$B_{ij} = \frac{m_{ij}}{M_{ij}} \tag{43}$$

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.

[1]We use the term "gene" generically for all interacting nodes in the network. This may include proteins encoded by the respective genes.

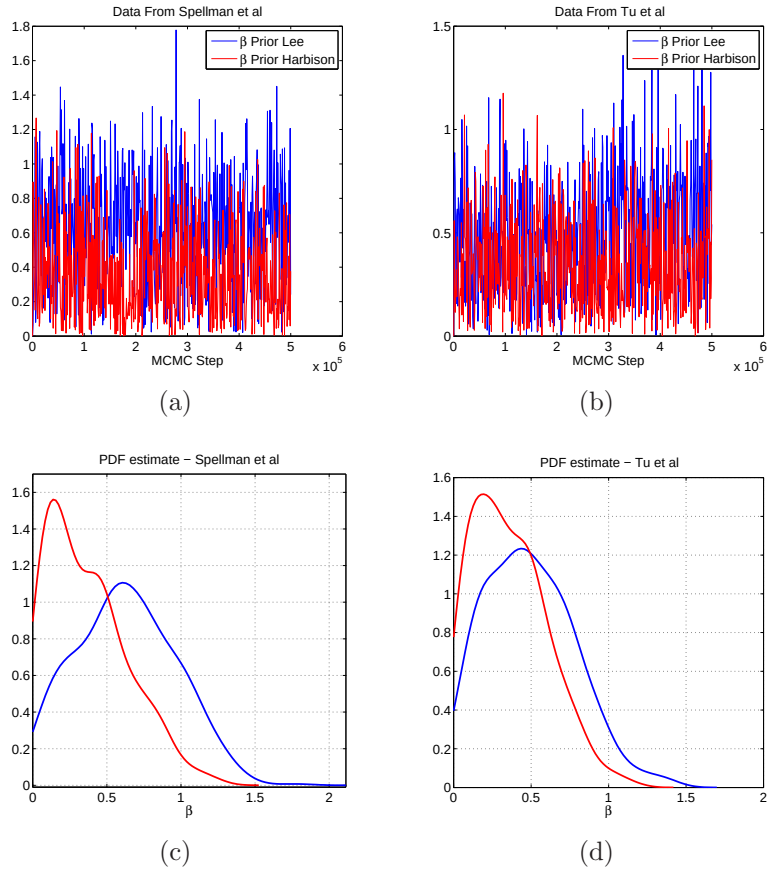


Figure 11: **Inferring hyperparameters associated with TF binding locations from gene expression data of yeast.** The top row (a,b) shows the hyperparameter trajectories for two different sources of prior knowledge, sampled from the posterior distribution with the MCMC scheme discussed in Section 2.4.2. The bottom row (c,d) shows the corresponding marginal posterior probability densities, estimated from the MCMC trajectories with a Parzen estimator, using a Gaussian kernel whose standard deviation was set automatically by the MATLAB function `ksdensity.m`. The blue line represents the hyperparameter associated with the TF binding locations of Lee *et al.* (2002). The red line shows the hyperparameter associated with the TF binding locations of Harbison *et al.* (2004). The two columns are related to different yeast microarray data. Left column: Spellman *et al.* (1998). Right column: Tu *et al.* (2005). The two experiments correspond to the first two rows of Table 3.

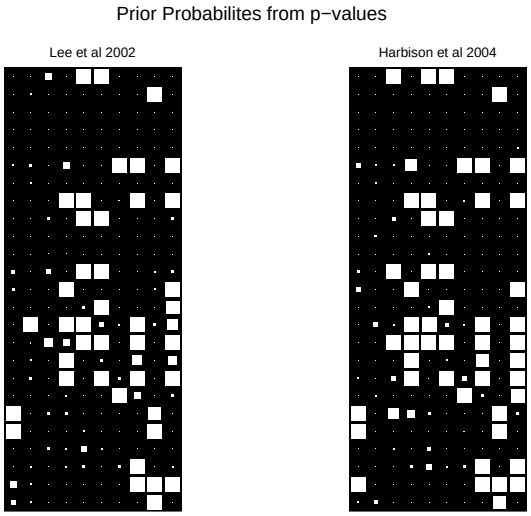


Figure 12: **Transcription factor (TF) binding locations.** The two Hinton diagrams provide a qualitative display of the TF binding location assays of Lee *et al.* (2002) (left panel) and Harbison *et al.* (2004) (right panel). The columns of the two matrices represent 10 known TFs. The rows represent 25 genes that are putatively regulated by the TFs. The size of a white square represents the probability that a TF binds to the promoter of the respective gene, with a larger square indicating a value closer to 1. These probabilities were obtained by subjecting the p-values from the original immunopreciptation experiments of Lee *et al.* (2002) and Harbison *et al.* (2004) to the transformation proposed by Bernard and Hartemink (2005).

5 Results

5.1 Yeast cell cycle

For evaluating the performance of the proposed Bayesian inference scheme on the yeast cell cycle data, we followed Bernard and Hartemink (2005) with the extension described in Section 4.2. We associated the edges of the BN with conditional probabilities of the multinomial distribution family. In this case, the marginal likelihood $P(D|G)$ of Equation 3 is given by the so-called BDe score; see Heckerman (1999) for details. The chosen form of conditional probabilities requires a discretization of the data. Like Bernard and Hartemink (2005), we discretized the gene expression data into three levels using the information bottleneck algorithm, proposed by Hartemink (2001). We represented information about the cell cycle phase with a separate node, which was forced to be a root node connected to all the nodes in the domain. In all our MCMC

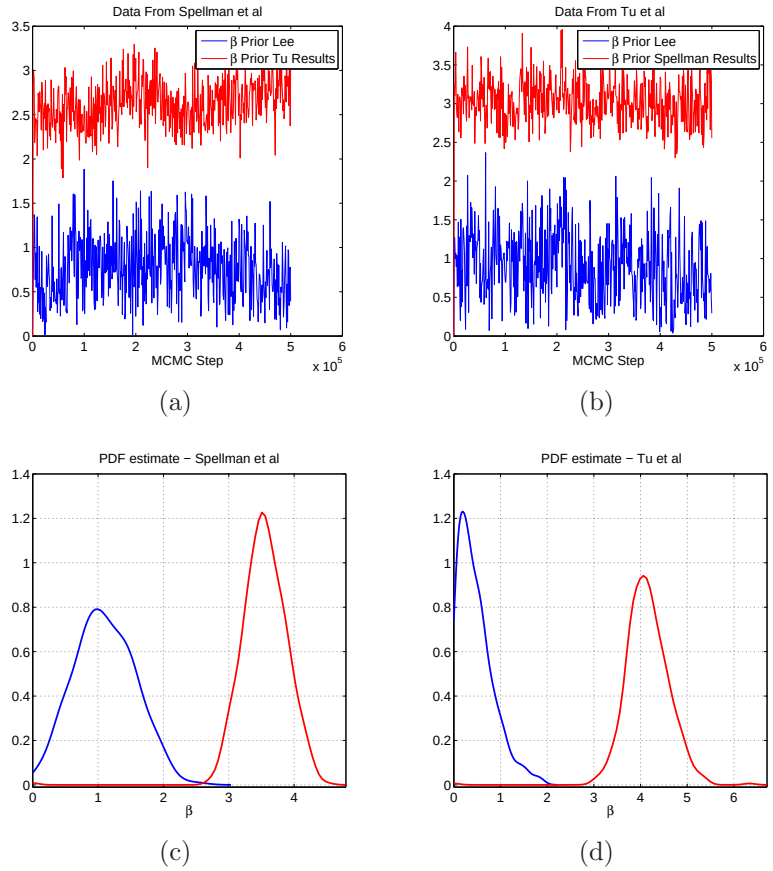


Figure 13: **Inferring hyperparameters associated with priors of different nature.** The graphs are similar to those of Figure 11, but were obtained for different sources of prior knowledge. The blue lines show the MCMC trace plots (top row) and estimated marginal posterior probability distributions (bottom row) of the hyperparameter associated with the TF binding locations from Lee *et al.* (2002). The red lines correspond to the hyperparameter associated with prior knowledge obtained from an independent microarray experiment in the way described in Section 5.1. The left column shows the results obtained from the experiment corresponding to the third row of Table 3. The right column shows the results obtained from the experiment corresponding to the fourth row of Table 3. For an explanation of the graphs, see the caption of Figure 11.

simulations, we combined gene expression data with two independent sources of prior knowledge, and sampled networks and hyperparameters from the conditional probability distribution according to the MCMC scheme described in Section 2.4.2.

Table 3 presents a summary of the simulation settings we used. In our first application, corresponding to the first row of Table 3, the gene expression data were taken from Spellman *et al.* (1998). In our second application, corresponding to the second row of Table 3, the gene expression data came from Tu *et al.* (2005). In both applications, we used the same two independent sources of prior knowledge in the form of transcription factor (TF) binding locations (Lee *et al.*, 2002; Harbison *et al.*, 2004), as described in Section 4.2.

The MCMC trajectories of the hyperparameters associated with the two sources of biological prior knowledge are presented in Figure 11. The figure also shows the estimated marginal posterior probability distributions of the two hyperparameters. These distributions, as well as the MCMC trace plots, do not appear to be very different, which suggests that the two priors are similar. A closer inspection of the results from the two TF binding assays, shown in Figure 12, reveals that the indications of putative TF binding locations obtained independently by Lee *et al.* (2002) and Harbison *et al.* (2004) are, in fact, very similar. This finding confirms that the results obtained with the proposed Bayesian inference scheme are consistent and in accordance with our expectation. From Figure 11 we also note that the sampled values of the hyperparameters are rather small, and that the estimated marginal posterior distributions – compared to those presented in the next section – are quite close to zero. This suggests that the prior information included is not in strong agreement with the data. There are two possible explanations for this effect. First, the TF activities might be controlled by post-translational modifications, which implies that the gene expression data obtained from microarray experiments might not contain sufficient information for inferring regulatory interactions between TFs and the genes they regulate. Second, there might be relevant regulatory interactions between genes that do not belong to the set of a priori known TFs, which are hence inherently undetectable by the binding assays.

One might therefore assume that prior knowledge obtained on the basis of a preceding microarray experiment might be more informative about a subsequent second microarray experiment than TF binding locations. To test this conjecture, we took one of the two gene expression data sets, assumed a uniform prior on network structures (subject to the usual fan-in restriction), and sampled networks from the posterior distribution with MCMC. From this sample, we obtained the marginal posterior probabilities of all edges, and used the

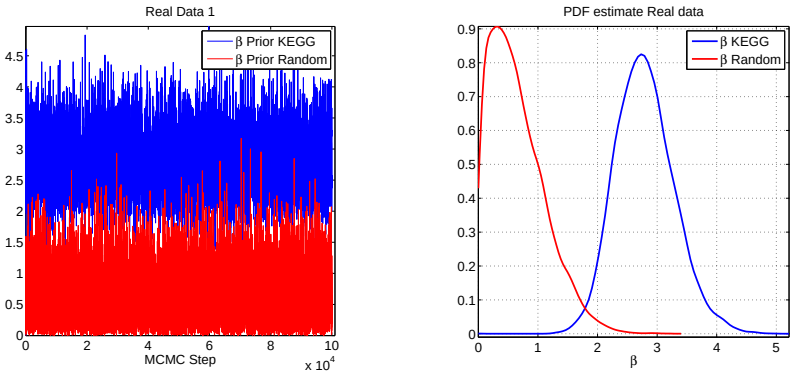
resulting matrix as a source of prior knowledge for the subsequent microarray experiment. We proceeded with the settings shown in the third and fourth row of Table 3. First, we combined the results obtained from the gene expression data of Spellman *et al.* (1998) with the binding locations from Lee *et al.* (2002) and applied these two sources of prior knowledge to the gene expression data from Tu *et al.* (2005). Second, we combined the results obtained from the gene expression data of Tu *et al.* (2005) with the binding locations from Lee *et al.* (2002) and applied these two sources of prior knowledge to the gene expression data from Spellman *et al.* (1998). The resulting hyperparameter trajectories are presented in Figure 13 together with their estimated probability densities. Compared with the previous results of Figure 11, there is now a much clearer separation between the two distributions. The sampled values of the hyperparameter associated with the second, independent source of microarray data significantly exceed those of the hyperparameter associated with the binding data. This suggests that prior knowledge that is more consistent with the data is given a stronger weight by the Bayesian inference scheme, in confirmation of our conjecture.

The critical question to ask next is: by how much does the accuracy of network reconstruction improve as a consequence of integrating prior knowledge into the inference scheme? Unfortunately, this evaluation cannot be done for yeast owing to our lack of knowledge about the true gene regulatory interactions and the absence of a proper gold-standard network. To answer this question, we therefore turn to a second application, for which more biological knowledge about the true regulatory processes exists.

5.2 Raf signalling pathway

5.2.1 Motivation

As described in Section 4.3, 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 10. Additionally, recent work by Sachs *et al.* (2005) 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 Werhli *et al.* (2006) and downsampled the data to a sample size representative of current microarray experiments (100 exemplars). As described in Section 4.3, the objective of



(a) (b)

Figure 14: **Inferring hyperparameters from the cytometry data of the Raf pathway.** The left panel (a) shows the hyperparameter trajectories for two different sources of prior knowledge, sampled from the posterior distribution with the MCMC scheme discussed in Section 2.4.2. The right panel (b) shows the corresponding posterior probability densities, estimated from the MCMC trajectories with a Parzen estimator, using a Gaussian kernel whose standard deviation was set automatically by the MATLAB function `ksdensity.m`. The blue lines refer to the hyperparameter associated with the prior knowledge extracted from the KEGG pathways. The red lines refer to completely random and hence vacuous prior knowledge. The data, on which the inference was based, consisted of 100 concentrations of the 11 proteins in the Raf pathway, subsampled from the observational cytometry data of Sachs *et al.* (2005).

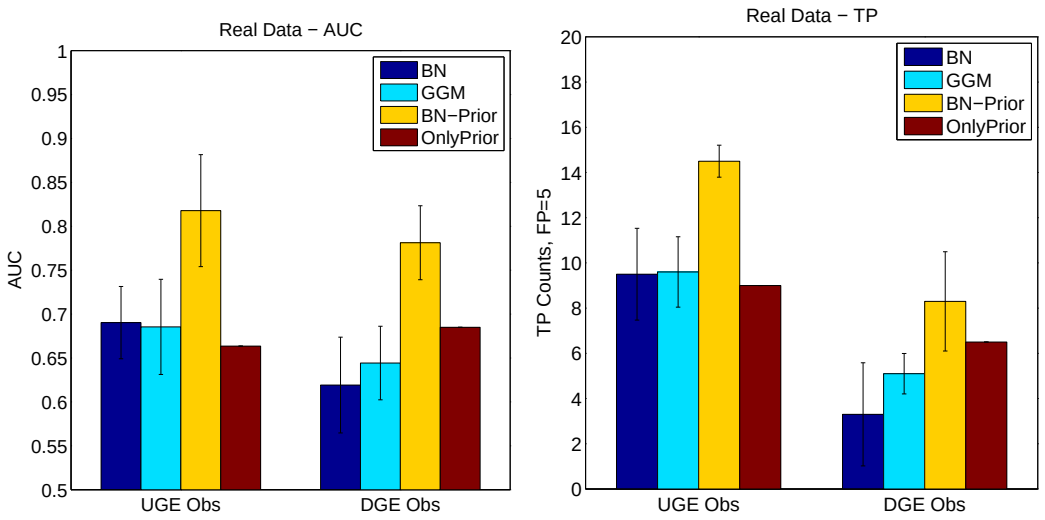


Figure 15: **Reconstruction of the Raf signalling pathway with different machine learning methods.** The figure evaluates the accuracy of inferring the Raf signalling pathway from cytometry data and prior information from KEGG. Two evaluation criteria were used. The left panel shows the results in terms of the area under the ROC curve (AUC scores), while the right panel shows the number of predicted true positive (TP) edges for a fixed number of 5 spurious edges. Each evaluation was carried out twice: with and without taking the edge direction into consideration (UGE: undirected graph evaluation, DGE: directed graph evaluation). 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 (Only Prior). In the latter case, the elements of the prior knowledge matrix (introduced in Section 2.3) were computed from Equation 43. The histogram bars represent the mean values obtained by averaging the results over five data sets of 100 protein concentrations each, independently sampled from the observational cytometry data of Sachs *et al.* (2005). The error bars show the respective standard deviations.

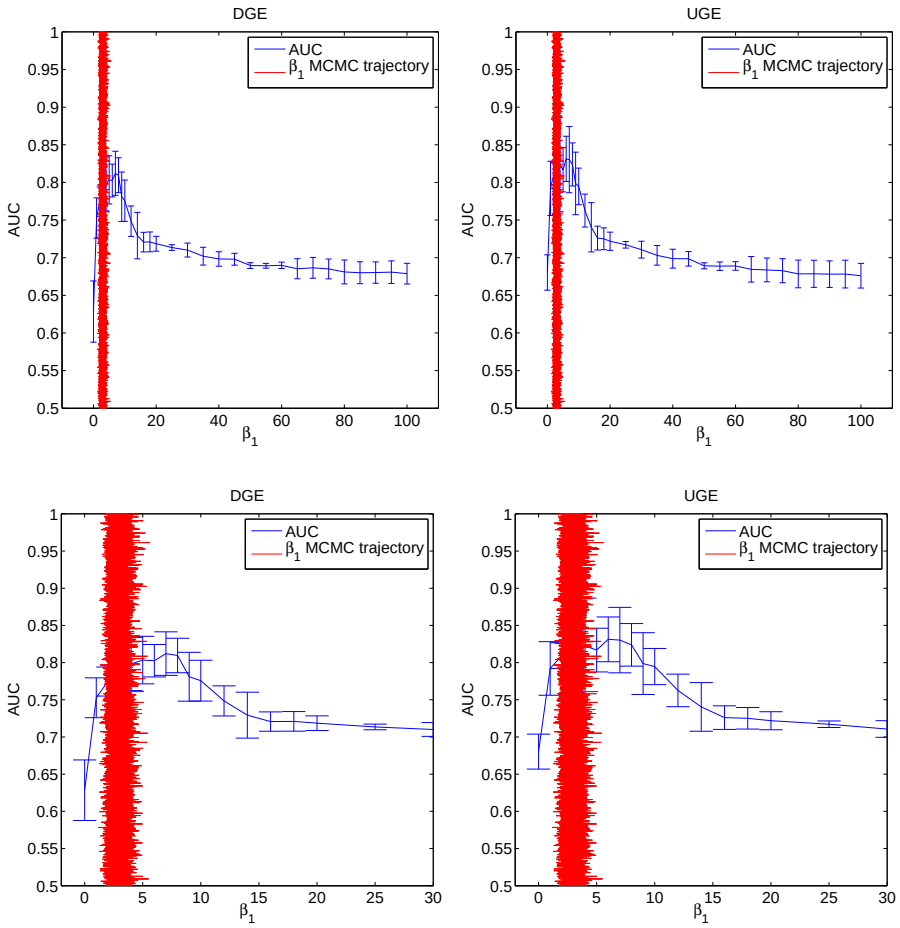


Figure 16: **Learning the hyperparameter associated with the prior knowledge from KEGG.** The horizontal axis represents the value of β_1, the hyperparameter associated with the prior knowledge from KEGG. The vertical axis represents the area under the ROC curve (AUC). The blue line shows the mean AUC score for fixed values of β_1, obtained by sampling network structures from the posterior distribution with MCMC. The results were averaged over five data sets of 100 protein concentrations each, independently sampled from the observational cytometry data of Sachs *et al.* (2005). The error bars show the respective standard deviations. The vertical red lines show trace plots of β_1 obtained with the MCMC scheme described in Section 2.4.2, where networks and hyperparameters are sampled from the posterior distribution. Each evaluation was carried out twice, with and without taking the edge direction into consideration. Right panel: undirected graph evaluation (UGE). Left panel: directed graph evaluation (DGE). The bottom row presents a magnified view of the left-most part of the graph.

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 (down-sampled) cytometry data with prior knowledge from the KEGG pathway database. To this end, we compared the results obtained with the methodology described in Section 2 with our earlier results from Werhli *et al.* (2006), where we had evaluated the performance of Bayesian networks (BNs) and Graphical Gaussian models (GGMs) without the inclusion of prior knowledge. We applied GGMs as described in Schäfer and Strimmer (2005). For comparability with Werhli *et al.* (2006), we used BNs with the family of linear Gaussian distributions, for which the marginal likelihood $P(D|G)$ of Equation 3 is given by the so-called BGe score; see Geiger and Heckerman (1994) for details. Note that the cytometry data of Sachs *et al.* (2005) are not taken from a time course; hence, BNs were treated as static rather than dynamic models.

5.2.2 Discriminating between different priors

We wanted to test whether the proposed Bayesian inference method can discriminate between different sources of prior knowledge and automatically assess their relative merits. To this end, we complemented the prior from the KEGG pathway database with a second prior, for which the entries in the prior knowledge matrix B were chosen completely at random. Hence, this second source of prior knowledge is vacuous and does not include any useful information for reconstructing the regulatory network. Figure 14 presents the MCMC trajectories of the hyperparameters β_1 and β_2 together with their respective estimated probability distributions. The hyperparameter associated with the KEGG prior, β_1, takes on substantially larger values than the hyperparameter associated with the vacuous prior, β_2. The estimated posterior distribution of β_1 covers considerably larger values than the estimated posterior distribution of β_2. This suggests that the proposed method successfully discriminates between the two priors and effectively suppresses the influence of the vacuous prior. Note that the vacuous prior is not completely 'switched off', though, and that the sampled values of β_2 are still substantially larger than zero. This seemingly counterintuitive behaviour is not a failure of the method, but rather an intrinsic feature of the posterior probability landscape; see Figure 9 and the discussion in Section 3.4.

5.2.3 Reconstructing the regulatory network

While the true network is a directed graph, our reconstruction methods may lead to undirected, directed, or partially directed graphs[2]. 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. Ideally, we would like to evaluate the methods on the basis of the whole ROC curves. Unfortunately, this approach would not allow us to concisely summarize the results. We therefore pursued two different approaches. The first approach is based on integrating the ROC curve so as to obtain the area under the curve (AUC), with larger areas indicating, overall, a better performance. The second approach is based on the selection of an arbitrary threshold on the edge scores, from which a specific network prediction is obtained. Following our earlier study (Werhli *et al.*, 2006), we set the threshold such that it led to a fixed count of 5 FPs. From the predicted network, we determined the number of correctly predicted (TP) edges, and took this score as our second figure of merit.

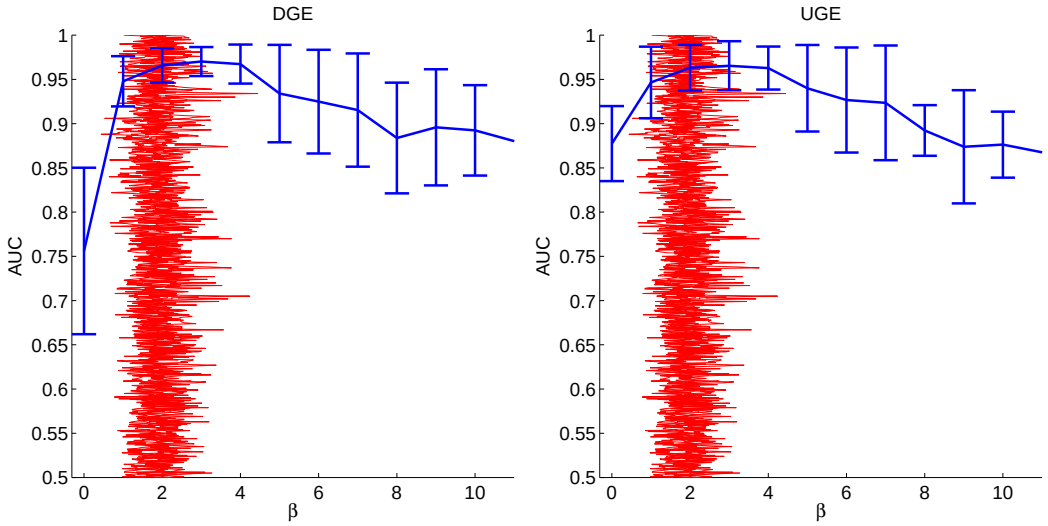
The results are shown in Figure 15. The proposed Bayesian inference scheme clearly outperforms the methods that do not include the prior knowledge from the KEGG database (BNs and GGMs). It also clearly outperforms the prediction that is solely based on the KEGG pathways alone without taking account of the cytometry data. The improvement is significant for all four evaluation criteria: AUC and TP scores for both directed (DGE) and

[2]GGMs 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 at the end of Section 2.1.

undirected (UGE) graph evaluations. This suggests that the network reconstruction accuracy can be substantially improved by systematically integrating expression data with prior knowledge about pathways, as extracted from the literature or databases like KEGG.

5.2.4 Learning the hyperparameters

While the study described in Section 5.2.2 suggests that the proposed Bayesian inference scheme succeeds in suppressing irrelevant prior knowledge, we were curious to see whether the hyperparameter associated with the relevant prior (from KEGG) was optimally inferred. To this end, we chose a large set of fixed values for β_1, while keeping the hyperparameter associated with the vacuous prior fixed at zero: $\beta_2 = 0$. For each fixed value of β_1, we sampled BNs from the posterior distribution with MCMC, and evaluated the network reconstruction accuracy using the evaluation criteria described in Section 5.2.3. 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 2.4.2. The results are shown in Figure 16. The blue lines show plots of the various prediction criteria obtained for fixed hyperparameters, plotted against β_1. Plotted along the vertical direction, the red lines show MCMC trace plots for the sampled values of β_1. These results suggest that the inferred values of β_1 are close to those that achieve the best network reconstruction accuracy. However, there is a small yet significant bias: the sampled values of β_1 lie systematically below those that optimize the reconstruction performance. There are two possible explanations for this effect. First, recall that for static BNs as considered here, the partition function of Equation 15 is only approximated by Equation 21, which could lead to a systematic bias in the inference scheme. Second, it has to be noted that the gold-standard Raf pathway reported in the literature is not guaranteed to be the true biological regulatory network. Interestingly, a recent study (Dougherty *et al.*, 2005) reports evidence for a negative feedback involving Raf, which is not included in the assumed gold standard network taken from Sachs *et al.* (2005). The existence of a hidden feedback loop acting on a putative feedforward path may lead to some systematic error in the edge directions, as static BNs are intrinsically restricted to the modelling of directed acyclic graphs. To shed further light on this issue, we therefore decided to carry out an additional synthetic study.



(a) (b)

Figure 17: **Learning the hyperparameter from synthetic data.** The graphs correspond to those of Figure 16, but were obtained from five independently generated synthetic data sets. These data were generated from the gold-standard Raf signalling pathway reported in Sachs *et al.* (2005), as described in Section 5.3. The prior knowledge was set to a corrupted version of the gold-standard network, in which 6 (out of the 20) true edges had been removed and replaced by wrong edges. For an explanation of the graphs and symbols, see the caption of Figure 16.

5.3 Comparison with simulated data

We simulated synthetic data from the Raf signalling network, depicted in Figure 10, as follows. Let $X_i(t)$ denote a random variable that represents the hypothetical protein concentration of node i. We generated data from a linear Gaussian distribution

$$X_i \sim N(\sum_k w_{ik} x_k, \sigma^2) \tag{44}$$

where the sum extends over all parents of node i (that is, those nodes with an edge pointing to node i), and x_k represents the value that node k takes on. We took the set of parents from the Raf signalling network, depicted in Figure 10, always initiating the process described by Equation 44 at the root, that is, node *pip3*. We set the standard deviation to $\sigma = 0.1$, and the interaction strengths to $w_{ik} = 1$. To mimic the situation described in the previous section, we generated 5 independent data sets with 100 samples each. As prior knowledge, we used a corrupted version of the true network, in which 6 (out of the 20) true edges had been removed and replaced by wrong edges. We then proceeded with the inference in the same way as described in Section 5.2. The results are shown in Figure 17, which corresponds to Figure 16 for the real cytometry data. From a comparison of these two figures, we note that the small bias in the inference of the hyperparameter has disappeared, and that values of the hyperparameter are sampled in the range where the reconstruction accuracy is optimized. This suggests that the small bias observed in Figure 16 might not be caused by the approximation of the partition function in Equation 21, but seems more likely to be a consequence of the other two effects discussed at the end of Section 5.2 (errors in the gold-standard network and putative feedback loops).

6 Discussion

The work presented here is based on pioneering work by Imoto *et al.* (2003) on learning gene regulatory networks from expression data and biological prior knowledge, which has recently found a variety of applications (Tamada *et al.*, 2003; Nariai *et al.*, 2004; Tamada *et al.*, 2005; Imoto *et al.*, 2006). The idea is to express the available prior knowledge in terms of an energy function, from which a prior distribution over network structures is obtained in the form of a Gibbs distribution. The hyperparameter of this distribution, which corresponds to an inverse temperature in statistical physics, represents the weight associated with the prior knowledge relative to the data. Our work complements the work of Imoto *et al.* (2003) in various respects. We have extended

the framework to multiple sources of prior knowledge; we have derived and tested an MCMC scheme for sampling networks and hyperparameters simultaneously from the posterior distribution; we have elucidated intrinsic features of this scheme from an idealized network population amenable to a closed-form derivation of the posterior distribution; and we have assessed the viability of the proposed Bayesian inference approach on various synthetic and real-world data.

Our findings can be summarized as follows. When including two sources of prior knowledge of similar nature, the marginal posterior distributions of the associated hyperparameters are similar (Figure 11). When the two sources of prior knowledge are different, higher weight is assigned to the prior that is more consistent with the data (Figure 13). When including an irrelevant prior with vacuous information, its influence will be automatically suppressed (Figure 14) in that the marginal posterior distribution of the corresponding hyperparameter is shifted towards zero. The irrelevant prior is not completely switched off, though. This would correspond to a delta distribution sitting at zero, which is never observed, not even for the worst-case scenario of prior knowledge that is in complete contradiction to the true network and the data (Figures 9c and d). To elucidate this unexpected behaviour, we carried out two types of analysis. In the first case, we considered an idealized population of network structures for which the prior distribution could be computed in closed form (Equation 41). In the second case, we considered networks composed of a small number of nodes (Figure 5), for which the partition function of Equation 15, and hence the prior distribution over networks structures (Equation 14), could be numerically computed by exhaustive enumeration of all possible structures. Both types of analysis reveal that the posterior distribution over hyperparameters contains a flat plateau (Figure 9a-b), which accounts for our seemingly counter-intuitive observations.

We evaluated the accuracy of reconstructing the Raf protein signalling network, which has been extensively studied in the literature. To this end, we combined protein concentrations from cytometry experiments with prior knowledge from the KEGG pathway database. The findings of our study clearly demonstrate that the proposed Bayesian inference scheme outperforms various alternative methods that either take only the cytometry data or only the prior knowledge from KEGG into account (Figure 15). We inspected the values of the sampled hyperparameters. Encouragingly, we found that their range was close to the optimal value that maximizes the network reconstruction accuracy (Figure 16). A small systematic deviation would be expected owing to the approximation we have made for computing the partition function of the prior (Equations 11 and 21). Interestingly, a comparison between

real and simulated cytometry data – Figure 16 versus Figure 17 – revealed that the small bias only occurred in the former case. This suggests that other confounding factors, like errors in the gold-standard network and as yet unaccounted feedback loops, might have a stronger effect than the approximation made for computing the partition function.

A certain shortcoming of the proposed method is the intrinsic asymmetry between *prior knowledge* and *data*, which manifests itself in the fact that the hyperparameters of the prior are inferred from the data. Ultimately, the prior knowledge is obtained from some data also; for instance, prior knowledge about TF binding sites is obtained from immunoprecipitation data. A challenging topic for future research, hence, is to treat both *prior* and *data* on a more equal footing, and to develop more systematic methods of postgenomic data integration.

References

Bernard,A. and Hartemink,A.J. (2005) Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. In *Pacific Symposium on Biocomputing* pp. 459–470 World Scientific, New Jersey.

Butte,A.S. and Kohane,I.S. (2003) Relevance networks: a first step toward finding genetic regulatory networks within microarray data. In *The Analysis of Gene Expression Data*, (Parmigiani,G., Garett,E.S., Irizarry,R.A. and Zeger,S.L., eds), pp. 428–446 Springer, New York.

Chickering,D.M. (1995) A transformational characterization of equivalent Bayesian network structures. *International Conference on Uncertainty in Artificial Intelligence (UAI),* **11**, 87–98.

Cowles,M.K. and Carlin,B.P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association,* **91**, 883–904.

Dougherty,M.K., Müller,J., Ritt,D.A., Zhou,M., Zhou,X.Z., Copeland,T.D., Conrads,T.P., Veenstra,T.D., Lu,K.P. and Morrison,D.K. (2005) Regulation of Raf-1 by direct feedback phosphorylation. *Molecular Cell,* **17**, 215–224.

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.

Friedman,N., Murphy,K. and Russell,S. (1998) Learning the structure of dynamic probabilistic networks. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, (Cooper,G.F. and Moral,S., eds), pp. 139–147 Morgan Kaufmann, San Francisco, CA.

Geiger,D. and Heckerman,D. (1994) Learning Gaussian networks. pp. 235–243 Morgan Kaufmann, San Francisco, CA.

Giudici,P. and Castelo,R. (2003) Improving Markov chain Monte Carlo model search for data mining. *Machine Learning,* **50**, 127–158.

Harbison,C.T., Gordon,B.D., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J., Jennings,E.G., Zeitlinger,J., Pokholok,D.K., Kellis,M., Rolfe,A.P., Takusagawa,K.T., Lander,E.S., Gifford,D.K., Fraenkel,E. and Young,R.A. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature,* **431** (7004), 99–104.

Hartemink,A.J. (2001). *Principled computational methods for the validation and discovery of genetic regulatory networks.* PhD thesis, MIT.

Hartemink,A.J., Gifford,D.K., Jaakkola,T.S. and Young,R.A. (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing,* **6**, 422–433.

Hastings,W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* **57**, 97–109.

Heckerman,D. (1999) A tutorial on learning with Bayesian networks. In *Learning in Graphical Models*, (Jordan,M.I., ed.), Adaptive Computation and Machine Learning pp. 301–354 MIT Press, Cambridge, Massachusetts.

Heckerman,D., Geiger,D. and Chickering,D.M. (1995) Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning,* **20**, 245–274.

Husmeier,D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics,* **19**, 2271–2282.

Husmeier,D., Dybowski,R. and Roberts,S. (2005) *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Advanced Information and Knowledge Processing, Springer, New York.

Imoto,S., Higuchi,T., Goto,T., Kuhara,S. and Miyano,S. (2003) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proceedings IEEE Computer Society Bioinformatics Conference,* **(CSB'03)**, 104–113.

Imoto,S., Higuchi,T., Goto,T. and Miyano,S. (2006) Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks. *Statistical Methodology,* **3** (1), 1–16.

Imoto,S., Kim,S., Goto,T., , Aburatani,S., Tashiro,K., Kuhara,S. and Miyano,S. (2003) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of Bioinformatics and Computational Biology,* **1** (2), 231–252.

Kanehisa,M. (1997) A database for post-genome analysis. *Trends Genet,* **13**, 375–376.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research,* **28**, 27–30.

Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research,* **34**, D354–357.

Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I., Zeitlinger,J., Jennings,E.G., Murray,H.L., Gordon,D.B., Ren,B., Wyrick,J.J., Tagne,J.B., Volkert,T.L., Fraenkel,E., Gifford,D.K. and Young,R.A. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. . *Science,* **298** (5594), 799–804.

Madigan,D. and York,J. (1995) Bayesian graphical models for discrete data. *International Statistical Review,* **63**, 215–232.

Murphy,K. and Milan,S. (1999). Modelling gene expression data using dynamic Bayesian networks. Technical report MIT artificial intelligence laboratory.

Nariai,N., Kim,S., Imoto,S. and Miyano,S. (2004) Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pacific Symposium on Biocomputing,* **9**, 336–347.

Sachs,K., Perez,O., Pe'er,D., Lauffenburger,D.A. and Nolan,G.P. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science, Vol 308, Issue 5721, 523-529 , 22 April 2005,* **308** (5721), 523–529.

Schäfer,J. and Strimmer,K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology,* **4**. Article 32.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell,* **9** (12), 3273–3297.

Tamada,Y., Bannai,H., Imoto,S., Katayama,T., Kanehisa,M. and Miyano,S. (2005) 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.

Tamada,Y., Kim,S., Bannai,H., Imoto,S., Tashiro,K., Kuhara,S. and Miyano,S. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics,* **19**, ii227–ii236.

Tu,B.P., Kudlicki,A., Rowicka,M. and Mcknight,S.L. (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science,* **310** (5751), 1152–1158.

Werhli,A.V., Grzegorczyk,M. and Husmeier,D. (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics,* **22** (20), 2523–2531.