Journal of Bioinformatics and Computational Biology
Vol. 6, No. 3 (2008) 543–572
© Imperial College Press

Imperial College Press
www.icpress.co.uk

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

GENE REGULATORY NETWORK RECONSTRUCTION BY BAYESIAN INTEGRATION OF PRIOR KNOWLEDGE AND/OR DIFFERENT EXPERIMENTAL CONDITIONS

ADRIANO V. WERHLI

Department of Computing Science
Pontifical Catholic University of Rio Grande do Sul
Porto Alegre, Brazil
werhli@gmail.com

DIRK HUSMEIER

Biomathematics and Statistics Scotland (BIOSS)
and Centre for Systems Biology at Edinburgh (CSBE)
Edinburgh, United Kingdom
dirk@bioss.ac.uk

Received 1 August 2007
Revised 1 December 2007
Accepted 3 January 2008

There have been various attempts to improve the reconstruction of gene regulatory networks from microarray data by the systematic integration of biological prior knowledge. Our approach is based on pioneering work by Imoto *et al.*[11] where the prior knowledge is expressed in terms of energy functions, from which a prior distribution over network structures is obtained in the form of a Gibbs distribution. The hyperparameters of this distribution represent the weights associated with the prior knowledge relative to the data. We have derived and tested a Markov chain Monte Carlo (MCMC) scheme for sampling networks and hyperparameters simultaneously from the posterior distribution, thereby automatically learning how to trade off information from the prior knowledge and the data. We have extended this approach to a Bayesian coupling scheme for learning gene regulatory networks from a combination of related data sets, which were obtained under different experimental conditions and are therefore potentially associated with different active subpathways. The proposed coupling scheme is a compromise between (1) learning networks from the different subsets separately, whereby no information between the different experiments is shared; and (2) learning networks from a monolithic fusion of the individual data sets, which does not provide any mechanism for uncovering differences between the network structures associated with the different experimental conditions. We have assessed the viability of all proposed methods on data related to the Raf signaling pathway, generated both synthetically and in cytometry experiments.

Keywords: Gene regulatory networks; Bayesian networks; Bayesian inference; Markov chain Monte Carlo; gene expression data; Raf pathway; KEGG; data integration.

1. Introduction

Bayesian networks have received increasing attention from the computational biology community as models of gene regulatory networks, following up on pioneering work by Friedman *et al.*[1] and Hartemink *et al.*[2] Several tutorials on Bayesian networks have been published.[3–5] We therefore only qualitatively recapitulate some aspects that are of relevance for the present study, and refer the reader to the above tutorials for a thorough and more rigorous introduction.

The structure of a Bayesian network is defined by a directed acyclic graph (DAG) indicating how different variables of interest, represented by nodes, interact. The word "interact" has a causal connotation, which is ultimately of interest to the biologist, but has to be taken with caution in this context, as explained shortly. The edges of a Bayesian network are associated with conditional probabilities, defined by a functional family and their parameters. The interacting entities are associated with random variables, which represent some measured entities of interest like relative gene expression levels or protein concentrations. We denote the set of all the measurements of all the random variables as the data, represented by the letter D. As a consequence of the acyclicity of the network structure, the joint probability of all the random variables can be factorized into a product of lower-complexity conditional probabilities according to conditional independence relations defined by the graph structure $\mathcal{M}$. Under certain regularity conditions, the parameters associated with these conditional probabilities can be integrated out analytically. This allows us to compute the marginal likelihood or evidence $P(D|\mathcal{M})$, which captures how well the network structure $\mathcal{M}$ explains the data D. In the present study, we computed $P(D|\mathcal{M})$ under the assumption of a linear Gaussian distribution. The resulting score was derived by Geiger and Heckerman[6] and is referred to as the BGe score.

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

There are two ways to break the symmetries of the equivalence classes. One approach is to use active interventions, like gene knockouts or overexpressions. If knocking out gene A affects gene B, but knocking out gene B does not affect gene A, then $A \rightarrow B$ will tend to have a higher evidence than $A \leftarrow B$. For more detail, see Pournara and Wernisch[7] and Werhli *et al.*[8] An alternative way to break the

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

symmetries, investigated in this paper, is to use prior information. If genes A and B are conditionally dependent, and we have prior knowledge that A is a transcription factor that regulates genes in the functional category B belongs to, then we will presumably favor $A \rightarrow B$ over $A \leftarrow B$. To formalize this notion, we score networks by the posterior probability

$$P(\mathcal{M}|D) \propto P(D|\mathcal{M})P(\mathcal{M}), \tag{1}$$

where $P(D|\mathcal{M})$ is the evidence, and $P(\mathcal{M})$ is the prior distribution over network structures; the latter distribution captures the biological knowledge that we have prior to measuring the data D. While different graphs might have identical scores in light of the data, $P(D|\mathcal{M})$, symmetries can be broken by the inclusion of prior knowledge, $P(\mathcal{M})$, and these two sources of information are systematically integrated into the posterior distribution $P(\mathcal{M}|D)$. Our ultimate objective, therefore, is to find the network structure $\mathcal{M}$ that maximizes $P(\mathcal{M}|D)$. Unfortunately, the number of structures increases superexponentially with the number of nodes. Also, in systems biology, where we aim to learn complex interaction patterns involving many components, the amount of information from the data and the prior is usually not sufficient to render the distribution $P(\mathcal{M}|D)$ sharply peaked in a single graph; instead, the distribution is usually diffusely spread over a large set of networks. Summarizing this distribution by a single network would not be appropriate. Instead, we aim to sample network structures from the posterior distribution $P(\mathcal{M}|D)$ so as to obtain a typical collection of high-scoring networks and, thereby, capture intrinsic inference uncertainty. Direct sampling from this distribution is usually intractable, though. Hence, we resort to a Markov chain Monte Carlo (MCMC) scheme,[9] which under fairly general regularity conditions is theoretically guaranteed to converge to the posterior distribution of Eq. (1). Given a network structure $\mathcal{M}_{\text{old}}$, a new network structure $\mathcal{M}_{\text{new}}$ is proposed from a proposal distribution $Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}})$, which is then rejected or accepted according to the standard Metropolis–Hastings scheme[10] with the following acceptance probability:

$$A = \min\left\{ \frac{P(D|\mathcal{M}_{\text{new}})P(\mathcal{M}_{\text{new}})}{P(D|\mathcal{M}_{\text{old}})P(\mathcal{M}_{\text{old}})} \times \frac{Q(\mathcal{M}_{\text{old}}|\mathcal{M}_{\text{new}})}{Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}})}, 1 \right\}. \tag{2}$$

The functional form of the proposal distribution $Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}})$ depends on the chosen type of proposal moves. In the present paper, we consider three edge-based proposal operations: creating, deleting, and inverting an edge. The computation of the Hastings factor $Q(\mathcal{M}_{\text{old}}|\mathcal{M}_{\text{new}})/Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}})$ is, for instance, discussed in Husmeier *et al.*[4]

2. Methodology A: Integration of Prior Knowledge

2.1. *Biological prior knowledge*

To integrate biological prior knowledge into the inference of gene regulatory networks, we define a function that measures the agreement between a given network

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

$\mathcal{M}$ and our biological prior knowledge. Following an approach first proposed by Imoto *et al.*[11] and subsequently applied in Refs. 12–15, we call this measure the energy E, borrowing the name from statistical physics. We split E into two components. One of the components, E_0, is associated with the absence of edges; the other component, E_1, is associated with the presence of edges. A network $\mathcal{M}$ is represented by a binary adjacency matrix, where each entry $\mathcal{M}_{ij}$ can be either 0 or 1. A zero entry, $\mathcal{M}_{ij} = 0$, indicates the absence of an edge between node$_i$ and node$_j$; conversely, if $\mathcal{M}_{ij} = 1$, there is a directed edge from node$_i$ to node$_j$. We define the biological prior knowledge matrix B to be a matrix in which the entries $B_{ij} \in [0, 1]$ represent our knowledge about interactions between nodes as follows: If entry $B_{ij} = 0.5$, we do not have any prior knowledge about the presence or absence of the directed edge between node$_i$ and node$_j$. If $0 \leq B_{ij} < 0.5$, we have prior evidence that the directed edge between node$_i$ and node$_j$ is absent; the evidence is stronger as B_{ij} is closer to 0. If $0.5 < B_{ij} \leq 1$, we have prior evidence that the directed edge pointing from node$_i$ to node$_j$ is present; the evidence is stronger as B_{ij} is closer to 1. Having defined how to represent a network $\mathcal{M}$ and the biological prior knowledge B, we now define the energies associated with the presence and absence of edges as follows:

$$E_0(\mathcal{M}) = \sum_{\substack{i,\,j\,=\,1 \\ B_{i,j}\,<\,0.5}}^{N} |B_{i,j} - \mathcal{M}_{i,j}| \tag{3}$$

$$E_1(\mathcal{M}) = \sum_{\substack{i,\,j\,=\,1 \\ B_{i,j}\,>\,0.5}}^{N} |B_{i,j} - \mathcal{M}_{i,j}|\,, \tag{4}$$

where N is the total number of nodes.

To integrate the prior knowledge expressed by Eqs. (3) and (4) into the inference procedure, we follow Imoto *et al.*[11] and define the prior distribution over network structures $\mathcal{M}$ to take the form of a Gibbs distribution:

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

where β_0 and β_1 are hyperparameters that indicate the weight of the respective source of prior knowledge relative to the data, and the partition function is defined as

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

with $\mathbb{M}$ denoting the set of all valid (i.e. directed and acyclic) network structures. Unfortunately, the number of structures increases superexponentially with the number of nodes, rendering the computation of Z not viable for large networks. To

proceed, we define

$$E_0(\mathcal{M}) = \sum_{n=1}^{N} \mathcal{E}_0(n, \pi_n[\mathcal{M}]) \tag{7}$$

$$E_1(\mathcal{M}) = \sum_{n=1}^{N} \mathcal{E}_1(n, \pi_n[\mathcal{M}]), \tag{8}$$

where $\pi_n[\mathcal{M}]$ is the set of parents of node n in the graph $\mathcal{M}$, and we have defined

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

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

Akin to the perfect gas approximation in statistical physics (e.g. Chapter 7 in Balian[16]), we now approximate the partition function of the whole network by a product of single-node partition functions:

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

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 a constraint on their cardinality (for which we chose an upper bound of 3, as in Refs. 1, 17, and 18). The consequence of the perfect gas approximation is a considerable reduction in the computational complexity. However, structures with directed cycles, i.e. invalid DAGs, are not excluded from the sum. Consequently, the price to pay for the reduced computational complexity is a small yet systematic overestimation of the partition function. According to one of our previous studies,[19] this bias does not appear to be critical, though.

2.2. *MCMC sampling scheme*

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

Starting from a definition of the prior distributions on the hyperparameters β_0 and β_1, $P(\beta_0)$ and $P(\beta_1)$, our aim is to sample the network structure $\mathcal{M}$ and the hyperparameters β_0 and β_1 from the posterior distribution $P(\mathcal{M}, \beta_0, \beta_1|D)$. To this end, we propose a new network structure $\mathcal{M}_{\text{new}}$ from the proposal distribution $Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}})$ and, additionally, new hyperparameters from the proposal distributions $R(\beta_{0_{\text{new}}}|\beta_{0_{\text{old}}})$ and $R(\beta_{1_{\text{new}}}|\beta_{1_{\text{old}}})$. We then accept this move according

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

to the standard Metropolis–Hastings update rule[10] with the following acceptance probability:

$$A = \min\left\{ \frac{P(D, \mathcal{M}_{\text{new}}, \beta_{0_{\text{new}}}, \beta_{1_{\text{new}}})}{P(D, \mathcal{M}_{\text{old}}, \beta_{0_{\text{old}}}, \beta_{1_{\text{old}}})} \times \frac{Q(\mathcal{M}_{\text{old}}|\mathcal{M}_{\text{new}}) R(\beta_{0_{\text{old}}}|\beta_{0_{\text{new}}})}{Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}}) R(\beta_{0_{\text{new}}}|\beta_{0_{\text{old}}})} \right.$$
$$\left. \times \frac{R(\beta_{1_{\text{old}}}|\beta_{1_{\text{new}}})}{R(\beta_{1_{\text{new}}}|\beta_{1_{\text{old}}})}, 1 \right\}$$
$$= \min\left\{ \frac{P(D|\mathcal{M}_{\text{new}}) P(\mathcal{M}_{\text{new}}|\beta_{0_{\text{new}}}, \beta_{1_{\text{new}}})}{P(D|\mathcal{M}_{\text{old}}) P(\mathcal{M}_{\text{old}}|\beta_{0_{\text{old}}}, \beta_{1_{\text{old}}})} \times \frac{P(\beta_{0_{\text{new}}}) P(\beta_{1_{\text{new}}}) Q(\mathcal{M}_{\text{old}}|\mathcal{M}_{\text{new}})}{P(\beta_{0_{\text{old}}}) P(\beta_{1_{\text{old}}}) Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}})} \right.$$
$$\left. \times \frac{R(\beta_{0_{\text{old}}}|\beta_{0_{\text{new}}}) R(\beta_{1_{\text{old}}}|\beta_{1_{\text{new}}})}{R(\beta_{0_{\text{new}}}|\beta_{0_{\text{old}}}) R(\beta_{1_{\text{new}}}|\beta_{1_{\text{old}}})}, 1 \right\}. \tag{12}$$

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

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

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

$$A(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}}) = \min\left\{ \frac{P(D|\mathcal{M}_{\text{new}})}{P(D|\mathcal{M}_{\text{old}})} \times \frac{P(\mathcal{M}_{\text{new}}|\beta_0, \beta_1)}{P(\mathcal{M}_{\text{old}}|\beta_0, \beta_1)} \times \frac{Q(\mathcal{M}_{\text{old}}|\mathcal{M}_{\text{new}})}{Q(\mathcal{M}_{\text{new}}|\mathcal{M}_{\text{old}})}, 1 \right\} \tag{13}$$

$$A(\beta_{0\text{new}}|\beta_{0\text{old}}) = \min\left\{ \frac{P(\mathcal{M}|\beta_{0\text{new}}, \beta_1)}{P(\mathcal{M}|\beta_{0\text{old}}, \beta_1)}, 1 \right\} \tag{14}$$

$$A(\beta_{1\text{new}}|\beta_{1\text{old}}) = \min\left\{ \frac{P(\mathcal{M}|\beta_0, \beta_{1\text{new}})}{P(\mathcal{M}|\beta_0, \beta_{1\text{old}})}, 1 \right\}. \tag{15}$$

The three submoves are iterated until some convergence criterion is satisfied, discarding an initial burn-in phase before sampling configurations. In our simulations, we chose the prior distribution of each hyperparameter $P(\beta_i)$, $i \in \{0, 1\}$, to be the uniform distribution over the interval $[0, \text{MAX}]$, with $\text{MAX} = 30$. The proposal distribution of the hyperparameters $R(\beta_{i_{\text{new}}}|\beta_{i_{\text{old}}})$ was chosen to be a uniform distribution over a moving interval of length $L = 6 \ll \text{MAX}$, centered on the current value of the respective hyperparameter and using reflection to satisfy the constraint $\beta_{i_{\text{new}}} \in [0, \text{MAX}]$. Note that L only affects the convergence and mixing of the Markov chain — that is, the computational efficiency — and could, in principle, be adjusted during the burn-in phase. To test for convergence of the MCMC

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

simulations, various methods have been developed.[20] In our work, we applied the scheme used in Werhli *et al.*[8]: each MCMC run was repeated from independent initializations, and consistency in the marginal posterior probabilities of the edges was taken as an indication of sufficient convergence, leading to a typical trajectory length of 5×10^5 steps, of which the first half was discarded as the burn-in phase.

3. Methodology B: Active Pathways Under Different Experimental Conditions

The assumption so far has been that the molecular biological system of interest can be characterized by a unique regulatory network. What we are actually aiming to infer, though, are the active parts of this network, which may differ under different experimental conditions. To illustrate this point, consider a transcription factor that potentially upregulates a group of genes further downstream in the regulatory chain. If the experimental conditions are chosen such that the gene coding for this transcription factor is never expressed itself, then the respective subnetwork will never be activated and hence cannot be inferred from the data. When aiming to infer regulatory networks related to an organism's immune system, we would expect certain pathways to be activated only upon infection and remain invisible when gene expression profiles are only taken in the healthy state. In fact, some preliminary analysis in Werhli[21] related to the challenging of macrophages with interferon-gamma (IFNγ) and viral infection has revealed differences in the active pathways under the conditions of viral infection, IFNγ treatment, and viral infection plus IFNγ treatment. This suggests that a regulatory network is not an immutable entity, but may vary in response to changes in the experimental and/or environmental conditions.

When aiming to reconstruct a network from gene expression profiles taken under different experimental conditions, there seem to be two principled approaches we may pursue. The first is to ignore the changes in the experimental conditions altogether and merge the data into one monolithic set. The problem with this approach is that it inevitably blurs the differences between the different conditions and thereby obscures the biological insight we are aiming to gain; for instance, we would not be able to tell the difference between the state of a network in infected, healthy, and IFNγ-treated cells. The second approach is to keep the data obtained under different conditions separate, and to infer separate regulatory networks active under these different conditions. While this approach has the potential to reveal the differences between the regulatory networks in different states, e.g. infection versus treatment, it will almost inevitably result in a considerable reduction in statistical power and reconstruction accuracy. Current postgenomic data sets are usually sparse, e.g. the number of microarray experiments biologists can afford to carry out is usually limited to the order of a few dozen, which compromises the extent to which networks can be reconstructed. Breaking a sparse data set up into smaller

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

units will inevitably aggravate this situation and increase the uncertainty about inferred network structures.

In the present work, we aim to pursue a compromise between the two extreme procedures described above. The motivation is given by the insight gained from our earlier study described in Chapter 2 of Werhli.[21] Although we found differences between the active pathways under the different conditions of infection and treatment with IFNγ, the networks shared considerable features in common. Our conjecture is that this holds in general, and that a cell's regulatory networks, while potentially transitioning between different active states in response to different external cues, share substantial features owing to a common generic network architecture. Our objective is to formulate this proposition mathematically so as to integrate it into the probabilistic modeling process.

As it turns out, this objective can be achieved by a modification of the probabilistic model described in Sec. 2. Recall that the objective of Sec. 2 was the integration of explicit prior knowledge into the inference scheme by softly constraining the inferred network to be similar to the *a priori* known network. In modification of this scheme, we now propose learning separate regulatory networks from disjunct gene expression data, but tying these networks together by softly constraining them to be similar to a shared underlying generic network. This approach overcomes the rigidity of the first scenario described above, which would obscure the differences between the network states in different experimental conditions. By sharing information between the different network states, the problem of the second scenario described above is also averted; that is, the statistical power and accuracy of the reconstruction should be considerably enhanced.

3.1. *Probabilistic model*

In order to integrate information from I different data sets $(\mathcal{D}_1 \dots \mathcal{D}_I)$ obtained under different experimental conditions, we use the probabilistic graphical model presented in Fig. 1. Each data set $(\mathcal{D}_1 \dots \mathcal{D}_I)$ is associated with its own hyperparameter $(\beta_1, \dots, \beta_I)$ and network structure $(\mathcal{M}_1, \dots, \mathcal{M}_I)$. The latent graph $\mathcal{M}^\star$, which is not directly associated with the data, leads to a coupling between the individual network structures $(\mathcal{M}_1, \dots, \mathcal{M}_I)$ and encourages them to be similar. Note that Fig. 1 constitutes a hierarchical Bayesian model, in which the β_i's and $\mathcal{M}^\star$ correspond to hyperparameters that determine the prior distribution on the network structures $\mathcal{M}_i$'s. Further note that $\mathcal{M}^\star$ is not just a variable, but a complex entity representing a whole network itself. We therefore refer to $\mathcal{M}^\star$ as the hypernetwork.

The joint probability of the probabilistic graphical model of Fig. 1 is given by

$$P(\mathcal{M}_1, \dots, \mathcal{M}_I, \mathcal{D}_1 \dots \mathcal{D}_I, \beta_1, \dots, \beta_I, \mathcal{M}^\star) = \prod_{i=1}^{I} P(\mathcal{D}_i|\mathcal{M}_i) P(\mathcal{M}_i|\beta_i, \mathcal{M}^\star) P(\beta_i) P(\mathcal{M}^\star), \tag{16}$$

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

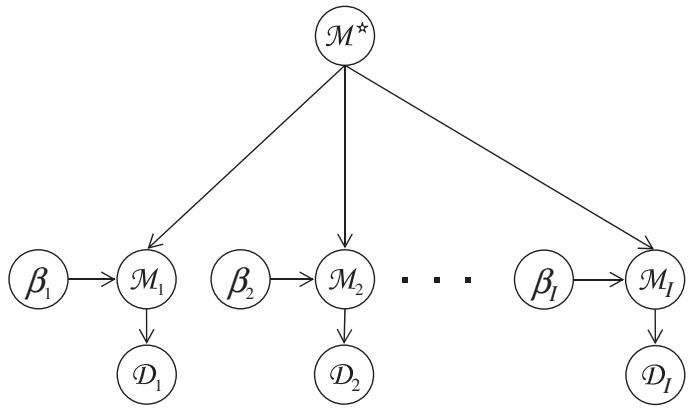


Fig. 1. Probabilistic model for learning active subnetworks under different experimental conditions. $(\mathcal{D}_1 \ldots \mathcal{D}_I)$ are data sets obtained under different experimental conditions. Each of these data sets is associated with its own hyperparameter $(\beta_1, \ldots, \beta_I)$ and network structure $(\mathcal{M}_1, \ldots, \mathcal{M}_I)$. The hypernetwork $\mathcal{M}^\star$ leads to a coupling between the individual network structures $(\mathcal{M}_1, \ldots, \mathcal{M}_I)$ and encourages them to be similar.

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

where the prior distribution over network structures, $P(\mathcal{M}_i|\beta_i, \mathcal{M}^\star)$, takes the form of a Gibbs distribution:

$$P(\mathcal{M}_i|\beta_i, \mathcal{M}^\star) = \frac{e^{-\beta_i(|\mathcal{M}_i - \mathcal{M}^\star|)}}{Z(\beta_i, \mathcal{M}^\star)}. \tag{17}$$

Recall that the hyperparameter β_i corresponds to an inverse temperature in statistical physics, and the term $|\mathcal{M}_i - \mathcal{M}^\star|$ measures the similarity between the graphs $\mathcal{M}_i$ and $\mathcal{M}^\star$, for instance in terms of the Hamming distance (i.e. the number of different edges). This scheme introduces a coupling between the individual networks $\mathcal{M}_i$: deviations between $\mathcal{M}_i$ and $\mathcal{M}^\star$ are penalized, which implies an indirect penalty for deviations between $\mathcal{M}_i$ and $\mathcal{M}_k$, $i \neq k$. The denominator in Eq. (17) is a normalizing constant, also known as the partition function:

$$Z(\beta_i, \mathcal{M}^\star) = \sum_{\mathcal{M}_i \in \mathbb{M}} e^{-\beta_i(|\mathcal{M}_i - \mathcal{M}^\star|)}, \tag{18}$$

where $\mathbb{M}$ is the set of all valid network structures. The summation over all possible models $\mathcal{M}_i$ can be performed efficiently using Eq. (11), as discussed in the text below that equation.

The hyperparameter β_i can be interpreted as a factor that indicates the strength of the influence of the hypernetwork $\mathcal{M}^\star$ relative to the data. For $\beta_i \to 0$, the prior distribution defined in Eq. (17) becomes flat and uninformative about the network structure. Conversely, for $\beta_i \to \infty$, the prior distribution becomes sharply peaked, forcing the network structure $\mathcal{M}_i$ to be equal to the hypernetwork $\mathcal{M}^\star$.

3.2. *MCMC sampling scheme*

Our goal is to sample all network structures $\mathcal{M}_i$, all of the hyperparameters β_i, and the hypernetwork $\mathcal{M}^\star$ from the posterior distribution. In order to achieve this objective, we propose new structures $\mathcal{M}_{i_{\text{new}}}$ from the proposal distribution $Q_i(\mathcal{M}_{i_{\text{new}}}|\mathcal{M}_{i_{\text{old}}})$, new hyperparameters from the proposal distribution $R_i(\beta_{i_{\text{new}}}|\beta_{i_{\text{old}}})$, and a new hypernetwork from the proposal distribution $W(\mathcal{M}^\star_{\text{new}}|\mathcal{M}^\star_{\text{old}})$. We then accept these moves according to the standard Metropolis–Hastings update rule[10] with the following acceptance probability:

$$A = \min\left\{\prod_{i=1}^{I} \frac{P(\mathcal{D}_i, \mathcal{M}_{i_{\text{new}}}, \beta_{i_{\text{new}}}, \mathcal{M}^\star_{\text{new}}) Q_i(\mathcal{M}_{i_{\text{old}}}|\mathcal{M}_{i_{\text{new}}}) R_i(\beta_{i_{\text{old}}}|\beta_{i_{\text{new}}})}{P(\mathcal{D}_i, \mathcal{M}_{i_{\text{old}}}, \beta_{i_{\text{old}}}, \mathcal{M}^\star_{\text{old}}) Q_i(\mathcal{M}_{i_{\text{new}}}|\mathcal{M}_{i_{\text{old}}}) R_i(\beta_{i_{\text{new}}}|\beta_{i_{\text{old}}})} \times \frac{W(\mathcal{M}^\star_{\text{old}}|\mathcal{M}^\star_{\text{new}})}{W(\mathcal{M}^\star_{\text{new}}|\mathcal{M}^\star_{\text{old}})}, 1\right\}. \tag{19}$$

For symmetric proposal distributions $R_i(\beta_{i_{\text{new}}}|\beta_{i_{\text{old}}})$ and $W(\mathcal{M}^\star_{\text{new}}|\mathcal{M}^\star_{\text{old}})$, this expression simplifies to

$$A = \min\left\{\prod_{i=1}^{I} \frac{P(\mathcal{D}_i, \mathcal{M}_{i_{\text{new}}}, \beta_{i_{\text{new}}}, \mathcal{M}^\star_{\text{new}}) Q_i(\mathcal{M}_{i_{\text{old}}}|\mathcal{M}_{i_{\text{new}}})}{P(\mathcal{D}_i, \mathcal{M}_{i_{\text{old}}}, \beta_{i_{\text{old}}}, \mathcal{M}^\star_{\text{old}}) Q_i(\mathcal{M}_{i_{\text{new}}}|\mathcal{M}_{i_{\text{old}}})}, 1\right\}. \tag{20}$$

The prior distribution $P(\mathcal{M}^\star)$ can be chosen according to Eq. (5) so as to include explicit biological prior knowledge. However, for the sake of simplicity of the exposition and in order to focus on the coupling aspects of the proposed method, we assume that both prior distributions $P(\beta_i)$ and $P(\mathcal{M}^\star)$ are uniform; this leads to the following expression:

$$A = \min\left\{\prod_{i=1}^{I} \frac{P(\mathcal{D}_i|\mathcal{M}_{i_{\text{new}}}) P(\mathcal{M}_{i_{\text{new}}}|\beta_{i_{\text{new}}}, \mathcal{M}^\star_{\text{new}}) Q_i(\mathcal{M}_{i_{\text{old}}}|\mathcal{M}_{i_{\text{new}}})}{P(\mathcal{D}_i|\mathcal{M}_{i_{\text{old}}}) P(\mathcal{M}_{i_{\text{old}}}|\beta_{i_{\text{old}}}, \mathcal{M}^\star_{\text{old}}) Q_i(\mathcal{M}_{i_{\text{new}}}|\mathcal{M}_{i_{\text{old}}})}, 1\right\}, \tag{21}$$

where we have expanded the joint probability according to the conditional independence relations shown in Fig. 1. Note that $\mathcal{M}_i$'s, as opposed to $\mathcal{M}^\star$, need to be proper DAGs. For this reason, we include the corresponding Hastings factor — the last term in the equation — as it is not necessarily equal to one. In our simulations, to be discussed below, we have used edge-based proposal moves: the creation, deletion, and reversal of an edge. When enforcing these moves to be valid, that is, to lead to proper DAGs, the two proposal probabilities do not necessarily cancel out and therefore have to be explicitly computed; see Husmeier *et al.*[4] for further details.

In order to increase the acceptance probability and, hence, mixing and convergence of the Markov chain, we break the move up into submoves. First, we propose new structures for each of the networks $\mathcal{M}_i$ in turn, while keeping all of the other variables fixed. The new structures are accepted with the following acceptance

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

probabilities:

$$A(\mathcal{M}_{i_{\text{new}}}|\mathcal{M}_{i_{\text{old}}}) = \min\left\{\frac{P(\mathcal{D}_i|\mathcal{M}_{i_{\text{new}}})P(\mathcal{M}_{i_{\text{new}}}|\beta_i,\mathcal{M}^\star)Q_i(\mathcal{M}_{i_{\text{old}}}|\mathcal{M}_{i_{\text{new}}})}{P(\mathcal{D}_i|\mathcal{M}_{i_{\text{old}}})P(\mathcal{M}_{i_{\text{old}}}|\beta_i,\mathcal{M}^\star)Q_i(\mathcal{M}_{i_{\text{new}}}|\mathcal{M}_{i_{\text{old}}})},1\right\}$$
$$= \min\left\{\frac{P(\mathcal{D}_i|\mathcal{M}_{i_{\text{new}}})e^{-\beta_i(|\mathcal{M}_{i_{\text{new}}}-\mathcal{M}^\star|)}Q_i(\mathcal{M}_{i_{\text{old}}}|\mathcal{M}_{i_{\text{new}}})}{P(\mathcal{D}_i|\mathcal{M}_{i_{\text{old}}})e^{-\beta_i(|\mathcal{M}_{i_{\text{old}}}-\mathcal{M}^\star|)}Q_i(\mathcal{M}_{i_{\text{new}}}|\mathcal{M}_{i_{\text{old}}})},1\right\}, \tag{22}$$

where Eq. (17) has been used. Next, we propose new values for the trade-off hyperparameters β_i. Each of the trade-off hyperparameters is accepted with the following acceptance probability:

$$A(\beta_{i_{\text{new}}}|\beta_{i_{\text{old}}}) = \min\left\{\frac{P(\mathcal{M}_i|\beta_{i_{\text{new}}},\mathcal{M}^\star)}{P(\mathcal{M}_i|\beta_{i_{\text{old}}},\mathcal{M}^\star)},1\right\}$$
$$= \min\left\{\frac{e^{-\beta_{i_{\text{new}}}(|\mathcal{M}_i-\mathcal{M}^\star|)}Z(\beta_{i_{\text{old}}},\mathcal{M}^\star)}{e^{-\beta_{i_{\text{old}}}(|\mathcal{M}_i-\mathcal{M}^\star|)}Z(\beta_{i_{\text{new}}},\mathcal{M}^\star)},1\right\}. \tag{23}$$

Finally, a new hypernetwork is proposed and accepted with the following acceptance probability:

$$A(\mathcal{M}^\star_{\text{new}}|\mathcal{M}^\star_{\text{old}}) = \min\left\{\prod_{i=1}^{I}\frac{P(\mathcal{M}_i|\beta_i,\mathcal{M}^\star_{\text{new}})}{P(\mathcal{M}_i|\beta_i,\mathcal{M}^\star_{\text{old}})},1\right\}$$
$$= \min\left\{\prod_{i=1}^{I}\frac{e^{-\beta_i(|\mathcal{M}_i-\mathcal{M}^\star_{\text{new}}|)}Z(\beta_i,\mathcal{M}^\star_{\text{old}})}{e^{-\beta_i(|\mathcal{M}_i-\mathcal{M}^\star_{\text{old}}|)}Z(\beta_i,\mathcal{M}^\star_{\text{new}})},1\right\}. \tag{24}$$

To illustrate the plausibility of this sampling scheme, consider the sampling of the hyperparameters β_i according to Eq. (23). We would assume that, for a network $\mathcal{M}_i$ which consistently differs from the hypernetwork $\mathcal{M}^\star$, the corresponding hyperparameter β_i should be driven to small values (indicating weak coupling); while conversely, β_i should be driven to large values (indicating strong coupling) when a network $\mathcal{M}_i$ is consistently similar to $\mathcal{M}^\star$. This is indeed the case. In the first scenario, $|\mathcal{M}_i - \mathcal{M}^\star|$ tends to be large, and high values of β_i are repressed by the exponential term in Eq. (23). In the second scenario, $|\mathcal{M}_i - \mathcal{M}^\star|$ becomes small, and the exponential term tends toward a constant, indiscriminative with respect to selecting β_i. Note, however, that the partition function $Z(\beta_i, \mathcal{M}^\star)$ is a monotonically decreasing function in β_i, as seen from Fig. 2. This monotonicity provides a penalty for small values of β_i, driving β_i up to high values, as expected.

4. Data

4.1. *Cytometry data*

Sachs *et al.*[22] have applied intracellular multicolor flow cytometry experiments to measure protein concentrations related to the Raf pathway. Raf is a critical signaling protein involved in regulating cellular proliferation in human immune system

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

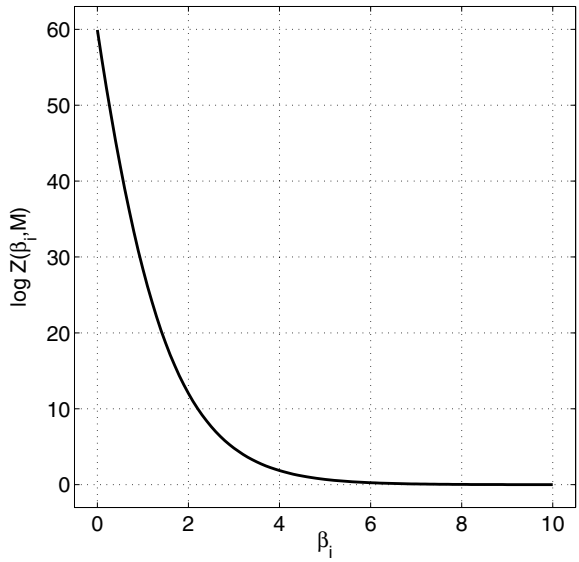


Fig. 2. Partition function example. The figure shows a plot of the partition function $Z(\beta_i, \mathcal{M}^\star)$ as a function of the hyperparameter β_i for a fixed hypernetwork $\mathcal{M}^\star$, chosen to be the gold standard Raf network of Fig. 3.

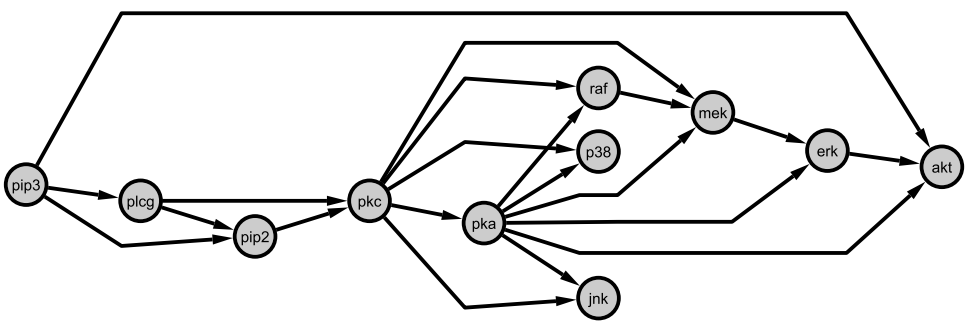


Fig. 3. Raf signaling pathway. The graph shows the currently accepted Raf signaling network, taken from Sachs *et al.*[22] Nodes represent proteins, edges represent interactions, and arrows indicate the direction of signal transduction.

cells. The deregulation of the Raf pathway can lead to carcinogenesis, and this pathway has therefore been extensively studied in the literature[22,23]; see Fig. 3 for a representation of the currently accepted gold standard network. In our experiments, we used five data sets with 100 measurements each, obtained by randomly sampling subsets from the original observational (i.e. unintervened) data of Sachs *et al.*[22] This subsampling was motivated by the fact that we wanted to investigate the learning performance on sample sizes typical of current microarray experiments, which do not provide the abundance of experimental conditions that one gets from

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

cytometry experiments. Details about how we standardized the data can be found in Werhli *et al.*[8]

4.2. *Synthetic data*

A simple synthetic way of generating data from the gold standard network of Fig. 3 is to sample them from a linear Gaussian distribution. The random variable X_i denoting the expression of node i is distributed according to

$$X_i \sim N\Big(\sum_k w_{ik} x_k, \sigma^2\Big), \tag{25}$$

where $N(.)$ denotes the normal distribution, the sum extends over all parents of node i, and x_k represents the value of node k. We set the standard deviation to $\sigma = 0.1$, sampled the interaction strength $|w_{ik}|$ from the uniform distribution over the interval $[0.5, 2]$, and randomly varied the sign of w_{ik}.

A more realistic simulation more typical of signals measured in molecular biology is based on treating the interactions in the network as enzyme-substrate reactions in organic chemistry. From chemical kinetics, it is known that the concentrations of the molecules involved in these reactions can be described by a system of ordinary differential equations (ODEs).[24,25] Assuming equilibrium and adopting a steady-state approximation, it is possible to derive a set of closed-form equations that describe the product concentrations as nonlinear (sigmoidal) functions of combinations of substrates. However, instead of solving the steady-state approximation to ODEs explicitly, we approximate the solution with a qualitatively equivalent combination of multiplications and sums of sigmoidal transfer functions. The resulting sigma-pi formalism has been implemented in the software package Netbuilder,[26,27] which we have used for simulating the data from the Raf signaling pathway, displayed in Fig. 3.

We used the same amount of data as for the flow cytometry experiments and created five simulated data sets with 100 measurements each. To model the stochastic influences, all nodes were subjected to additive Gaussian noise with zero mean and standard deviation equal to 0.1. More details about the generation of these data can be found in Werhli *et al.*[8]

4.3. *Biological prior knowledge*

We extracted biological prior knowledge from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database.[28–30] KEGG pathways represent the current knowledge of the molecular interaction and reaction networks related to metabolism, other cellular processes, and human diseases. As KEGG contains different pathways for different diseases, molecular interactions, and types of metabolism, it is possible to find the same pair of genes[a] in more than one pathway. We therefore

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

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

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 Sec. 2.1, as follows. Define by M_{ij} the total number of times a pair of genes i and j appears in a pathway, and by m_{ij} the number of times the genes are connected by a (directed) edge in the KEGG pathway. The elements B_{ij} of the prior knowledge matrix are then defined by

$$B_{ij} = \frac{m_{ij}}{M_{ij}}. \tag{26}$$

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.

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

4.4. *Simulating different active pathways*

To simulate active pathways under different experimental conditions, we combined five individual data sets as follows. For the synthetic data, three of the data sets were generated from the gold-standard Raf regulatory network, shown in Fig. 3. A fourth data set was generated from a slightly modified version of this network, in which the following four edges had been deleted: PKC → Raf, PKC → PKA, PKA → MEK, and PLCg → PIP2. An illustration of this network is shown in Werhli[21] and the supplementary material of Werhli *et al.*[8] The deletion of these edges corresponds to changes in the active subpathways under different external conditions, as described above. As a fifth data set, we included a purely random data set. This corresponds to either a drastic change of the external conditions that deactivates the whole pathway, or to a flawed experiment that has corrupted the data. We want to investigate whether the proposed method succeeds in identifying this outlying data set and prevents it from adversely affecting the overall inference. We are also interested in whether the proposed method can distinguish between the data from the gold standard and the modified Raf regulatory network. All of the synthetic data were taken from Werhli *et al.*,[8] where each subset contained 100 exemplars. For the cytometry data, we took four subsets of unintervened data, randomly selected from the data in Sachs *et al.*[22] and preprocessed as described in Werhli *et al.*[8] Each subset contained 20 measurements. To these data sets, we added a fifth data set of equal size, consisting of pure noise.

5. Evaluation Criteria

As described in Sec. 1, not all of the edge directions in a Bayesian network can always be inferred, which may lead to a partially directed graph. We compared the performance of Bayesian networks with graphical Gaussian models (GGMs), as proposed by Schäfer and Strimmer,[31] with the regularization approach described in Schäfer and Strimmer.[32] Note that GGMs are undirected graphs; hence, the network

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

reconstruction methods that we compared may lead to undirected, directed, or partially directed graphs. To assess the performance of these methods, we applied two different criteria. The first approach, referred to as *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 a 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 *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 network reconstruction methods considered in our study leads to a matrix of scores associated with the edges in a network. For Bayesian networks sampled from the posterior distribution with MCMC, these scores are the marginal posterior probabilities of the edges. For GGMs, these are partial correlation coefficients. Both scores define a ranking of the edges. This ranking defines a receiver operator characteristics (ROC) curve, where the relative number of true-positive (TP) edges is plotted against the relative number of false-positive (FP) edges. The results are shown in Fig. 4.

6. Results A: Integrating Prior Knowledge

The objective of our first study was the assessment of the method proposed in Sec. 2, where the objective is the integration of biological prior knowledge into the inference scheme. Figure 4 shows the ROC curves for four different network reconstruction methods: using the prior knowledge from KEGG only, according to Eq. (26); learning Bayesian networks and graphical Gaussian models from the protein concentration data alone; and using the proposed Bayesian inference scheme for integrating prior knowledge and data. The figure also distinguishes between learning the skeleton of the graph only (UGE) and considering the direction of the edges also (DGE). Recall that larger areas under the ROC curves indicate a better prediction performance overall, although the slope on the left is also of interest, as we are usually interested in keeping the number of false positives bounded at low values. The figure suggests that the systematic integration of prior knowledge with the proposed Bayesian inference scheme leads, overall, to a considerable improvement in the prediction performance over the three alternative schemes based on either the data or the prior knowledge from KEGG alone.

There are various interesting trends to be noted, though. For learning the skeleton of the graph (UGE), the improvement obtained on the real cytoflow data is more substantial than on the synthetic data (see left panel of Fig. 4). This is a consequence of the fact that on the synthetic data, Bayesian networks without prior knowledge already show a strong performance on learning the skeleton of the network, leaving not much room for further improvement. On the cytoflow data, on the other hand, the performance is much poorer; consequently, the integration of

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

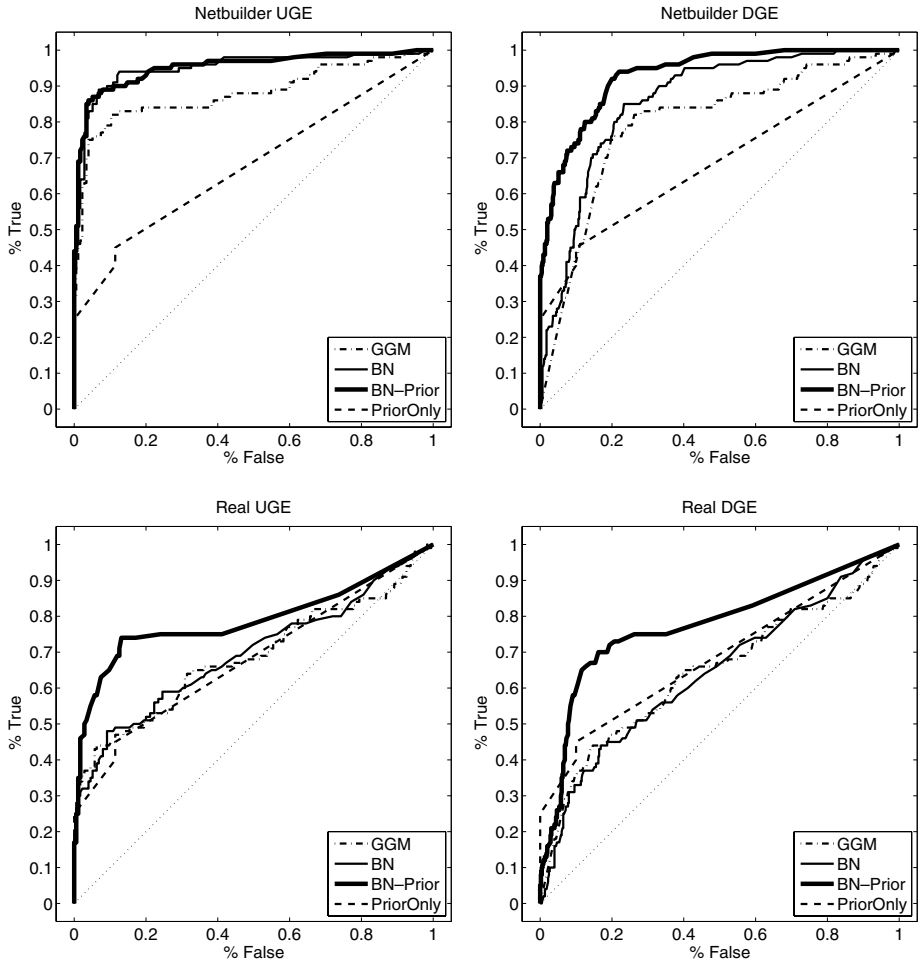


Fig. 4. Reconstruction of the Raf signaling pathway. The figure evaluates the accuracy of inferring the Raf signaling network from cytometry data (bottom row) and from simulated Netbuilder data (top row), each combined with prior information from KEGG. This evaluation was carried out twice: with and without taking the edge direction into account (UGE: undirected graph evaluation; 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), where the hyperparameters β_0 and β_1 were sampled from the posterior distribution with the MCMC scheme discussed in Sec. 2.2; and prior knowledge from KEGG only (PriorOnly). In the latter case, the elements of the prior knowledge matrix (introduced in Sec. 2.1) were computed from Eq. (26). The ROC curves presented are the mean ROC curves obtained by averaging the results over five different data sets.

prior knowledge leads to a more substantial improvement. When taking the edge directions into consideration (DGE), the proposed Bayesian integration scheme outperforms all other methods on the synthetic data (see Fig. 4, top right). This result is consistent with what has been discussed in Sec. 1: when learning Bayesian

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

networks from nondynamical noninterventional data (as considered here) without prior knowledge, there is inherent uncertainty about the direction of edges owing to intrinsic symmetries within network equivalence classes (see Sec. 1). These symmetries are broken by the inclusion of prior knowledge — hence, the improvement in the prediction performance. This improvement is also observed on the real cytoflow data (Fig. 4, bottom right), but to a lesser extent. Although the area under the ROC curve related to the Bayesian integration scheme exceeds that of all other ROC curves, the prediction based on prior knowledge alone shows a steeper slope in the very left region of the false-positive axis. This implies that for very high values of the threshold on the edge scores, a network learned from prior knowledge alone is more accurate than a network learned with any of the three methods that make use of the data. While the resulting network itself would not be particularly interesting — it would only contain a very small number (3 or 4) of the highest scoring edges — this observation is nevertheless interesting and can be explained as follows. The discrepancy between the UGE and DGE scores indicates that the Bayesian network learns the skeleton of the graph more accurately than the direction of the interactions, with some of the edge directions systematically inverted. A possible explanation is errors in the gold standard network.

The recent literature describes evidence for a negative feedback loop between Raf and ERK via MEK. Active Raf phosphorylates and activates MEK, which in turn activates ERK. This corresponds to the directed regulatory path shown in Fig. 3. However, through a hypothesized negative feedback mechanism involving ERK, Raf is phosphorylated on inhibitory sites, generating an inactive, desensitized Raf. Details can be found in Dougherty *et al.*[23] This feedback loop is not included in the gold standard network reported by Sachs *et al.*,[22] shown in Fig. 3. Such as-yet unaccounted feedback loops, as suggested in Dougherty *et al.*,[23] could explain systematic deviations between the predicted and the gold standard network, not only because the structure of a Bayesian network is constrained to be acyclic, but also because we ultimately do not have a reliable gold standard to assess the quality of the predictions. This example points to a fundamental problem inherent in any evaluation based solely on real biological data, and illustrates clearly the advantage of our combined evaluation based on both laboratory and simulated data.

It is obviously of interest to investigate the accuracy of the inference of the hyperparameters β_0 and β_1, especially as this inference depends on the partition function Z of Eq. (6), which can only be computed approximately (see Eq. (11)). To this end, we repeated the MCMC simulations for a large set of fixed values of β_0 and β_1, selected from the grid $[0, 20] \times [0, 20]$. For each pair of fixed values (β_0, β_1), we sampled Bayesian networks from the posterior distribution with MCMC, and evaluated the network reconstruction accuracy using the evaluation criteria described in Sec. 5. 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 Sec. 2.2. The results are shown in Fig. 5. The gray shading of the contour plots indicates the network

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

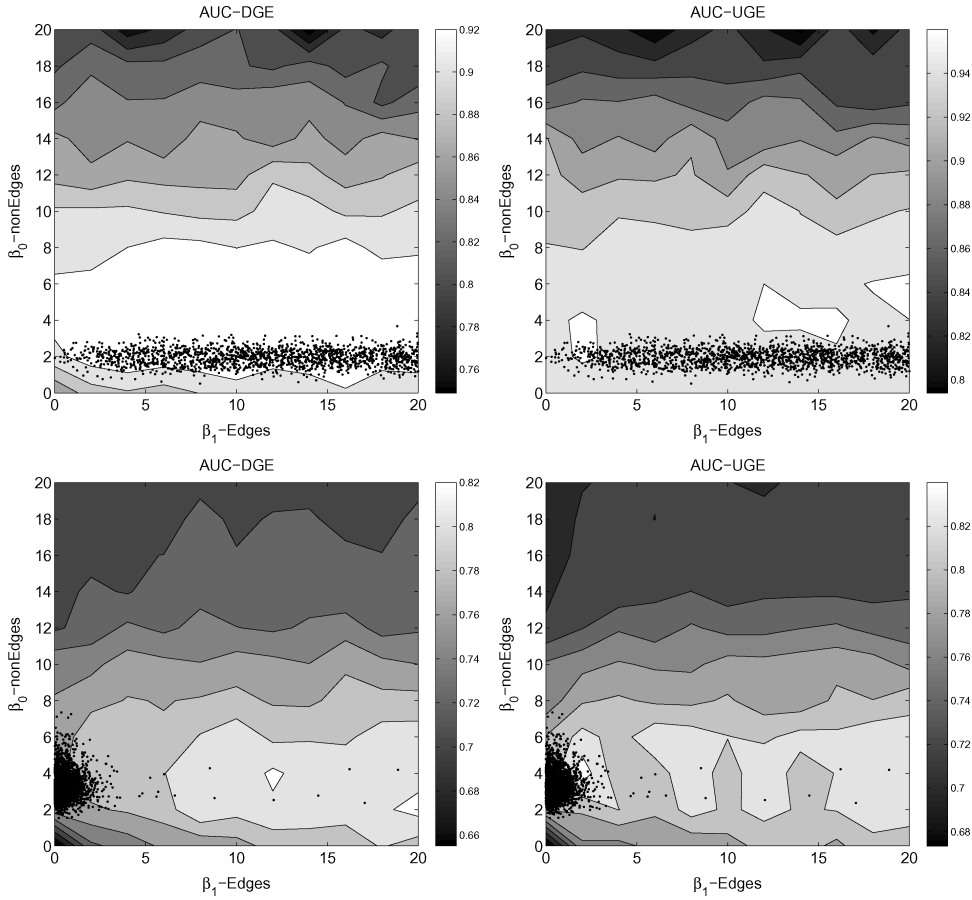


Fig. 5. Learning the hyperparameters associated with the prior knowledge from KEGG on simulated Netbuilder data and real flow cytometry data. The MCMC simulations described in Sec. 2.2 were repeated twice: for a large set of fixed values of the hyperparameters β_0 and β_1, and with the hyperparameters sampled from the posterior distribution with MCMC, using Eqs. (14) and (15). The gray shading of the contour plots represents the mean area under the ROC curve (AUC value) — averaged over five different data sets — as a function of the fixed values of the hyperparameters β_0 and β_1. The black dots show the values of these hyperparameters that were sampled in the MCMC simulations. The top row shows the results obtained on the simulated data. The bottom row shows the results obtained on the real flow cytometry protein concentrations. The left column shows the results for the directed graph evaluation (DGE), while the column on the right shows the results obtained when ignoring edge directions and only taking the skeleton of the network into account (UGE: undirected graph evaluation).

reconstruction accuracy in terms of DGE and UGE, obtained from the synthetic (top panels) and real cytometry data (bottom panels). The black dots show the hyperparameter values sampled with the MCMC simulations. While the distribution of β_0, the hyperparameter associated with the non-edges, is fairly peaked, the distribution of β_1, the hyperparameter associated with the edges, is rather diffuse.

This diffusion is particularly noticeable on the synthetic data. However, even on the real cytometry data, the distribution of β_1 has a long tail, with values being sampled across the whole permissible spectrum. An inspection of the prior knowledge matrix B extracted from KEGG according to Eq. (26) reveals that the prior knowledge associated with the energy function E_1 — Eq. (4) — accounts for only 25% of the true edges in the gold standard network of Fig. 3, while the prior knowledge associated with the energy function E_0 — Eq. (3) — accounts for 92% of the non-edges. Consequently, it appears that E_0 captures more relevant information for network reconstruction than E_1, which is reflected by the tighter distribution of the respective hyperparameter. The location of the sampled values of the hyperparameters β_0 and β_1 falls into the region of high network reconstruction scores. This suggests that the proposed Bayesian sampling scheme succeeds in finding hyperparameter values that lead to a high network reconstruction accuracy. A certain deviation from the optimal reconstruction would be expected owing to the approximation made for computing the partition function (see Eq. (11)). However, this deviation is small for both scores (UGE and DGE) on the synthetic data, and for the UGE score on the cytometry data. A noticeable deviation occurs for the DGE score on the cytometry data, though (see Fig. 5, bottom left panel). This deviation indicates a systematic mismatch between the DGE score and the posterior probability of the hyperparameters, which suggests that the cytometry data do not support all of the edge directions in the gold standard network of Fig. 3. Two possible explanations are either wrong edge directions in the gold standard network or the existence of as-yet unaccounted feedback loops, in confirmation of what has been discussed above.

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

7. Results B: Integrating Data Under Different Experimental Conditions

The objective of our second simulation study was the assessment of the method proposed in Sec. 3, where the aim is the integration of disjunct data sets corresponding to different experimental conditions.

7.1. *Inferring the hyperparameters*

Figure 6 shows various MCMC trace plots obtained on the linear Gaussian data, where the columns refer to different simulations. The first row shows trace plots of the log-likelihood,[b] while the remaining rows show trace plots of the hyperparameters β_i associated with the different data sets. The question of interest is whether the proposed method can identify the corrupted data set (pure noise),

[b]By "likelihood", we refer to the expression in Eq. (16), which strictly speaking should be called the joint probability of the probabilistic model defined in Fig. 1.

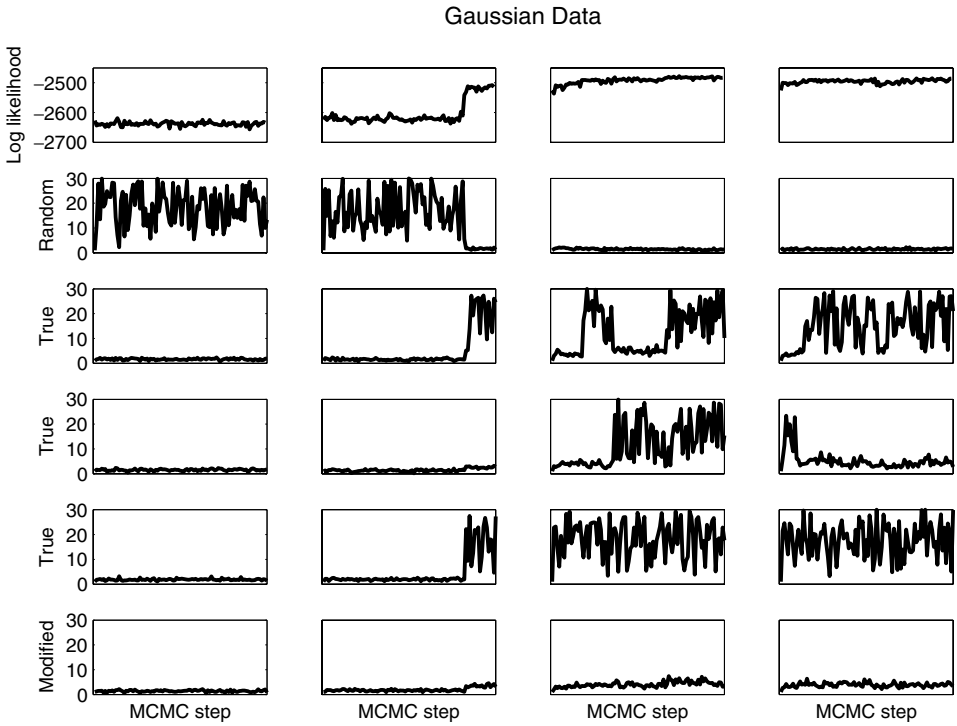


Fig. 6. MCMC trace plots for Gaussian data. The columns represent different simulations. The first row shows trace plots of the log-likelihood of Eq. (16), while the remaining rows show trace plots of the hyperparameters β_i associated with the five different data sets used. These data sets are of a different nature. *Random*: Corrupted data consisting of pure noise. *True*: Data sets generated from the gold standard Raf network, shown in Fig. 3. *Modified*: Data generated from the modified Raf network, in which four edges had been deleted, as described in the text. Note that insufficient convergence of the MCMC simulation represented by the first column is clearly indicated by the low likelihood scores (top row). The columns on the right show trace plots of MCMC simulations with significantly improved convergence; however, mixing problems are still evident, as indicated by the transitions between alternate stretches of high and low values of the hyperparameters β_0 and β_1. All MCMC simulations were run for 5×10^5 Metropolis–Hastings steps.

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

and distinguish between the data generated from the true network and those generated from the modified network. The first simulation (column 1) fails in this respect. In fact, the value of the hyperparameter β_{rand} associated with the corrupted data consistently exceeds the values of the other hyperparameters. However, the log-likelihood is consistently low, suggesting that the MCMC simulations have not yet converged. This conjecture is corroborated by the second simulation, which shows a behavior similar to the first simulation at the beginning, but then undergoes a sharp phase transition, during which β_{rand} is suddenly suppressed while the other hyperparameters shoot up to high values. A concomitant transition in the log-likelihood indicates that the Markov chain is escaping from

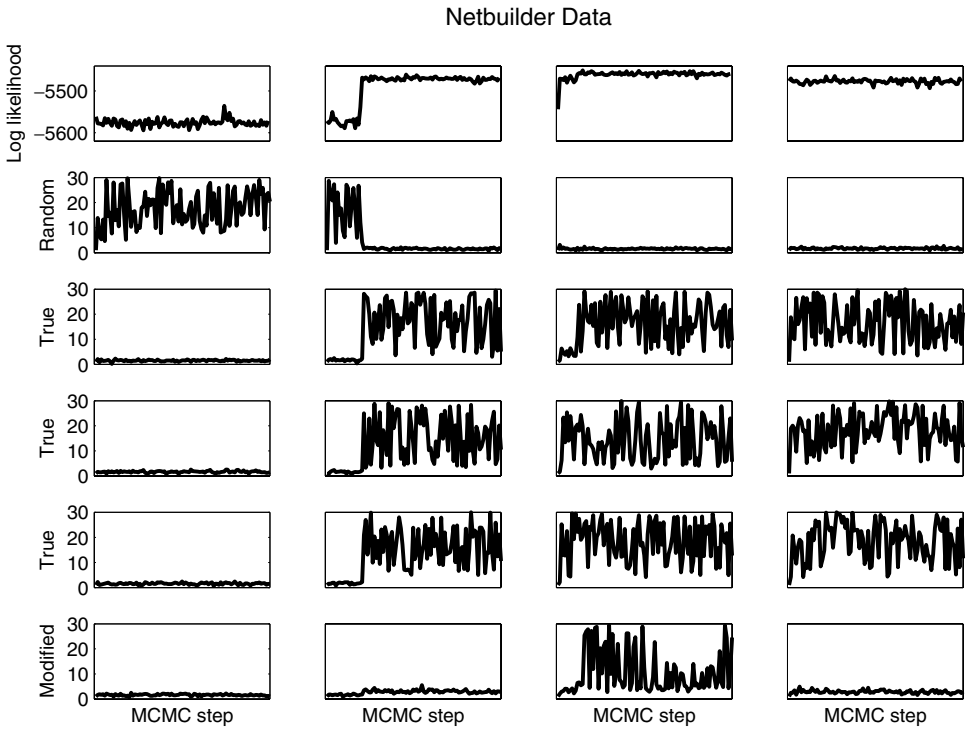


Fig. 7. MCMC trace plots for Netbuilder data. The graphs correspond to those of Fig. 6, but were obtained on the nonlinear synthetic data rather than the Gaussian data. See the caption of Fig. 6 for further details.

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

a metastable low-probability state in which it was trapped. The two remaining simulations, corresponding to columns 3 and 4 of Fig. 6, show a better convergence from the outset, with β_{rand} being consistently suppressed, and the hyperparameter associated with the modified network taking on values below those of the hyperparameters associated with the true network. A similar behavior can be found in Fig. 7, which was obtained from four MCMC simulations on the nonlinear synthetic data.

Figure 8 shows the estimated posterior distributions of the five hyperparameters for the best-converged MCMC simulations on both the linear and nonlinear synthetic data. These plots suggest that the proposed method succeeds in identifying the corrupted data, whose associated hyperparameter is significantly suppressed, as well as the data generated from the modified network. In the latter case, the distribution of the respective hyperparameter is shifted to lower values than the distributions of the hyperparameters associated with the true network. The amount of shift varies between the two data sets, which we suspect is more related to different degrees of convergence of the Markov chains than intrinsic differences between the linear and nonlinear data. The upshot of this study is that the proposed

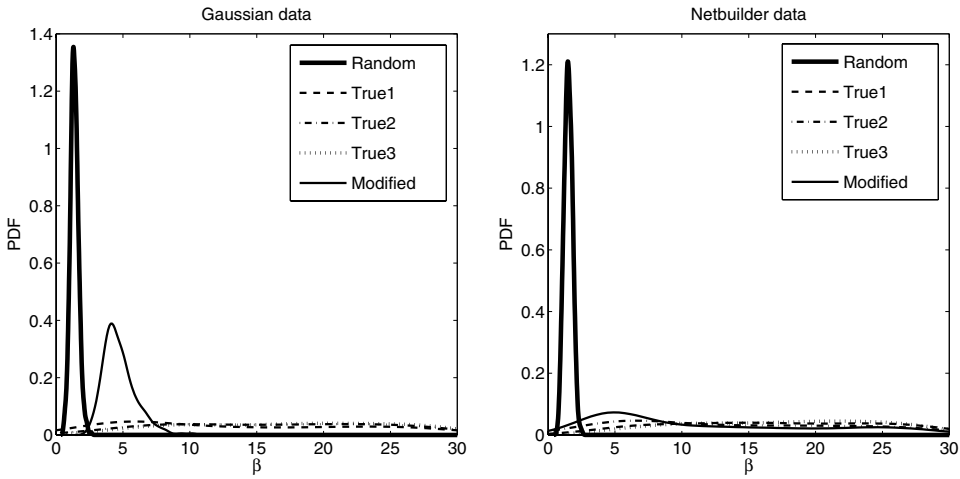


Fig. 8. Posterior distributions of the hyperparameters. These figures show the posterior distributions of the five hyperparameters $\beta_1, \ldots, \beta_5$, estimated with a kernel estimator applied to the samples obtained from the MCMC simulations with the best convergence characteristic, that is, column 3 in Fig. 6, and column 3 in Fig. 7. (a) Linear Gaussian data. (b) Nonlinear data generated with Netbuilder. The different line styles correspond to the different data types, as described in the text and the caption of Fig. 6. Note that the figure shows the qualitatively correct behavior of the hyperparameter distributions. The distribution of β_1, which is associated with the random data, is centered on small values close to zero. The hyperparameters β_2, β_3, and $\beta 4$, which are associated with the true data, have a broad distribution reaching up to very large values. The distribution of β_5, which is associated with the modified network, is expected to lie between these two distributions, and this is in fact borne out in our simulations. However, there is some disagreement between the two panels with respect to the exact format of the latter distribution; this disagreement is most likely a consequence of the poor mixing and insufficient convergence of the Markov chains. If the MCMC simulations were to be run until proper convergence (at increased computational costs), one would expect that the discrepancies between the two panels would disappear.

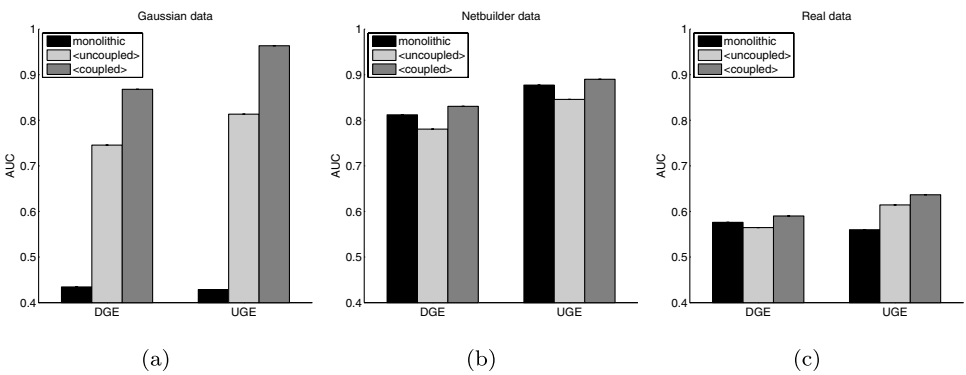
J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

method works successfully, but convergence problems of the MCMC simulations can become an issue. One problem in our first set of simulations was that we initialized all networks as empty graphs. This gives the hyperparameter associated with the corrupted data a certain "headstart": high-scoring networks inferred from the corrupted data will only contain a few edges, as there are no true associations between randomized nodes. This makes these networks similar to the hypernetwork (which was initialized as an empty graph), explaining the high value of β_{rand} at the beginning of some of our simulations. A better strategy is to pretrain the individual networks, e.g. using a greedy optimization, and setting the hypernetwork to their consensus network. While this has led to a modest improvement, there is still considerable scope for the development of more efficient MCMC sampling schemes, as discussed in Sec. 8.

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

7.2. *Network reconstruction*

We are particularly interested in whether the proposed coupling scheme leads to any improvement in terms of network reconstruction accuracy over the two alternative approaches described above: learning a single network from a merged, monolithic data set; and learning separate networks from the individual data sets without coupling. In what follows, we will refer to these methods as the *monolithic* and the *uncoupled* approaches, respectively. To summarize the results succinctly, we take the area under the ROC curve as a performance criterion, for both the DGE and the UGE scores, with larger areas indicating a better performance. The results are shown in Fig. 9. They indicate that the proposed coupling scheme consistently outperforms the other two approaches. The improvement is most pronounced on the synthetic Gaussian data. For these data, the control strength parameters associated with the edges in the regulatory network — w_{ik} of Eq. (25) — were different for each individual data set, which implies that even when the network structure itself did not change, the nature of the associated regulation processes could vary in both strength and sign (corresponding to an activation versus an inhibition). This



	Gaussian		Netbuilder		Cytometry	
	DGE	UGE	DGE	UGE	DGE	UGE
Monolithic	0.43	0.42	0.82	0.88	0.57	0.56
Uncoupled	0.74	0.81	0.78	0.84	0.56	0.61
Coupled	0.86	0.96	0.83	0.89	0.59	0.64

Fig. 9. Network reconstruction accuracy. The histograms and the table show a comparison of the network reconstruction accuracy in terms of AUC (area under the curve) scores for three different methods: the monolithic approach (black), the uncoupled approach (light gray), and the proposed Bayesian coupling scheme (dark gray); see the main text for further details. The three panels correspond to different data sets: linear Gaussian synthetic data (left panel), nonlinear synthetic data generated with Netbuilder (central panel), and protein concentrations from cytometry experiments (right panel). Each panel contains two histograms, evaluating only the reconstruction of the skeleton of the graph (UGE score) and additionally taking the edge direction into account (DGE score).

explains the poor performance of the monolithic approach, which intrinsically does not allow for any such variation. The difference in performance is less pronounced for the nonlinear synthetic data generated with Netbuilder, where only the instantiation of the noise rather than the parameters associated with the edges differed between different data sets. It appears that the slight performance improvement obtained with the proposed method is mainly a consequence of the inclusion of the corrupted data, whose influence is suppressed as a consequence of the adaptation of the associated hyperparameter, as discussed in the previous subsection. For the cytometry data, the amount of performance improvement achieved with the proposed method lies between the two synthetic data sets, with the improvement being more noticeable for the reconstruction of the skeleton of the graph (UGE score) than the reconstruction of the edge directions (DGE score).

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

7.3. *Convergence of the Markov chains*

A possible reason for the occasionally only modest performance improvement of the proposed method over the two alternative approaches is a lack of convergence of the MCMC simulations. Convergence problems have already been discussed in Sec. 7.1, and become more obvious in Fig. 10. The panels in this figure show scatter plots of

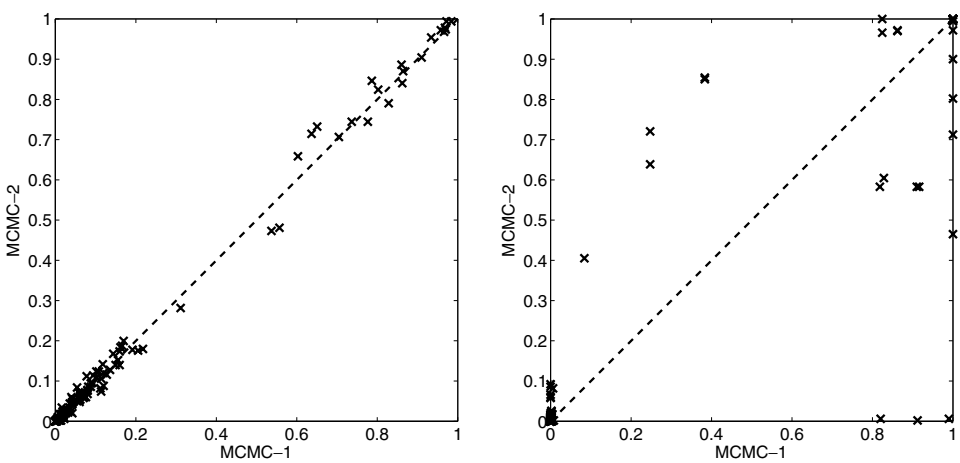


Fig. 10. MCMC convergence indication. Each of the two panels shows a scatter plot of the marginal posterior probabilities of the edges, obtained from two separate MCMC simulations applied to a subset of the nonlinear synthetic Netbuilder data. (a) The conventional approach, which aims to learn a separate Bayesian network from each subset of the data. (b) The proposed method, whereby Bayesian networks learned from different subsets of the data are coupled. The scatter plot was obtained from one of these coupled networks, corresponding to one of the $\mathcal{M}_i$'s in Fig. 1. For the conventional scheme, there is a clear consistency between the results from the two independent MCMC simulations, that is, there is no indication of any convergence difficulties. For the proposed coupling scheme, however, the marginal posterior probabilities obtained from the two independent MCMC simulations differ, which indicates a lack of convergence.

the marginal posterior probabilities of the edges obtained from two separate MCMC simulations, started from different initializations. Figure 10(a) was obtained from the conventional uncoupled MCMC scheme. The marginal posterior probabilities obtained from two independent simulations are very similar, indicating consistency of the predictions irrespective of the initialization. However, Fig. 10(b) — obtained from the proposed coupling scheme — shows a noticeable difference between the two independent MCMC simulations, which clearly indicates a lack of convergence. This behavior was found consistently throughout our simulations.

To shed more light on the convergence characteristics, we computed the average acceptance ratios of the MCMC moves during the whole simulation. Table 1 shows the acceptance ratios for the conventional scheme without coupling. Table 2 shows the acceptance ratios for the proposed coupling scheme. A comparison between

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

Table 1. MCMC acceptance ratios for uncoupled learning of network structures. This table shows the MCMC acceptance ratios (in percent) for the conventional scheme in which network structures $\mathcal{M}_1$ to $\mathcal{M}_5$ are learned independently from separate data sets. The higher acceptance ratio in the first row results from the fact that $\mathcal{M}_1$ was learned from random data, where the likelihood surface is relatively flat. The higher acceptance ratio in the last column results from the smaller sample size of the cytometry data (20 rather than 100 exemplars), which again leads to a flatter likelihood surface.

Network	Gaussian	Netbuilder	Cytometry
$\mathcal{M}_1$	54.4	54.4	55.9
$\mathcal{M}_2$	13.0	13.7	33.9
$\mathcal{M}_3$	13.4	15.3	24.9
$\mathcal{M}_4$	15.2	15.2	32.0
$\mathcal{M}_5$	12.8	13.5	31.6

Table 2. MCMC acceptance ratios for the proposed Bayesian coupling scheme. This table is to be compared with Table 1. It shows the MCMC acceptance ratios (in percent) for learning five network structures $\mathcal{M}_1$ to $\mathcal{M}_5$ from five separate data sets. As opposed to Table 1, the networks are coupled via a hypernetwork $\mathcal{M}^\star$ according to the proposed coupling scheme illustrated in Fig. 1. It is seen that as a consequence of this coupling, the MCMC acceptance probabilities have substantially decreased.

Network	Gaussian	Netbuilder	Cytometry
$\mathcal{M}_1$	25.6	18.5	0.4
$\mathcal{M}_2$	1.2	2.3	14.2
$\mathcal{M}_3$	0.3	2.7	2.7
$\mathcal{M}_4$	0.05	2.4	13.3
$\mathcal{M}_5$	1.2	4.4	1.8
$\mathcal{M}^\star$	4.5	11.0	10^{-3}

these two tables suggests that as a consequence of coupling, the acceptance probabilities have significantly decreased. This can be understood intuitively in that as a result of coupling, a local modification of a network structure is penalized not only when moving into regions of lower posterior probability, but also when increasing the difference between the network structures. The result is a higher rigidity of the Markov chain, which shows poorer mixing and convergence than the uncoupled scheme. A possible approach to deal with this rigidity is to adopt a simulated annealing scheme. Alternatively, more sophisticated sampling schemes could be explored, as briefly discussed below.

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

8. Conclusion

Our paper complements the work of Imoto *et al.*[11] on improving the reconstruction of regulatory networks from postgenomic data by the systematic integration of prior knowledge. The idea is to express the prior knowledge in terms of energy functions, from which a prior distribution over network structures is obtained in the form of a Gibbs distribution. The hyperparameters of this distribution represent the weights associated with the various sources of prior knowledge relative to the data. We have developed a Bayesian approach to inferring these hyperparameters, based on MCMC. We have tested the viability of this approach by trying to reconstruct the Raf pathway from flow cytometry protein concentrations and prior knowledge from KEGG. As an independent source of validation, we repeated the evaluation on synthetic data generated from the gold standard network. Our findings suggest that the Bayesian integration scheme systematically improves the network reconstruction over approaches that use either the protein concentrations only or the prior knowledge from KEGG alone. Also, the hyperparameters are sampled in regions close to those that yield the best possible network reconstruction, suggesting that the ideal gas approximation made for computing the partition function does not adversely affect the performance of the scheme. Learning the undirected skeleton graph from the cytometry data led to results that were systematically better than those obtained when learning the directed graph from these data, though. This difference between the directed and undirected graph reconstruction did not occur on the synthetic data, which suggests that either certain edge directions in the gold standard network are wrong or certain feedback loops are missing, in corroboration of the findings reported by Dougherty *et al.*[23]

Our simulations did not achieve any improvement in terms of network reconstruction accuracy over our earlier scheme described in Werhli and Husmeier,[19] in which no distinction between present and absent interactions in the network had been made. This suggests that more flexibility in the presentation of the prior knowledge does not automatically guarantee a performance improvement. One of the reasons for this lack of improvement is presumably related to the fact that most of the useful prior information was contained in the absence of edges, whereas only

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

little information was contained in the presence of interactions (as suggested by Fig. 5). The decision of whether an edge is present or absent depends on the choice of the threshold, though, which was rather arbitrarily set to a fixed value of 0.5 (see Eqs. (9) and (10)). A different choice of threshold parameter might have led to a smaller disparity between the two subsets of edges with respect to the information content, which suggests that sampling this parameter from the posterior distribution with MCMC might have led to a clearer performance enhancement. This further suggests, on a more general basis, that the flexibility and presentation of the prior knowledge about network structures, e.g. related to the subdivision of nodes and edges into subgroups, could be included in the MCMC scheme, which would provide an interesting avenue for future research.

In the present work, we have also proposed a Bayesian coupling scheme for learning gene regulatory networks from a combination of related data sets, which were obtained under different experimental conditions and are therefore potentially associated with different active subpathways. The proposed coupling scheme is a compromise between two extreme scenarios: (1) learning networks from the different subsets separately, whereby no information between the different experiments is shared; and (2) learning networks from a monolithic fusion of the individual data sets, which does not provide any mechanism for uncovering differences between the network structures associated with the different experimental conditions. Our proposed method combines the flexibility of the first approach with the data-merging aspect inherent in the second approach. The essential idea is that the networks associated with the different experimental conditions are softly constrained to be similar, where the strength of this constraint is defined by a hyperparameter that is automatically inferred from the data. Inference of these hyperparameters as well as the network structures is carried out in the Bayesian framework by approximately sampling from the posterior distribution with MCMC.

We have tested the proposed method on three types of data related to the Raf signaling pathway: two synthetic data sets, generated from the gold standard network, either under a linear Gaussian distribution or under a nonlinear distribution using Netbuilder; and real protein concentrations from cytometry experiments. Our results can be summarized as follows. Given sufficient convergence of the MCMC simulations, a random data set deliberately included with the proper data is clearly detected. The hyperparameter associated with the random data is automatically set to very small values; this suggests that the proposed Bayesian coupling scheme is effective in switching off the influence of corrupted data. A data set generated from a modified network structure is also automatically detected. The associated hyperparameter is sampled from a distribution placed between those associated with the random data and the data from the unmodified network, successfully distinguishing it from both. In terms of network reconstruction accuracy, the proposed Bayesian coupling scheme consistently outperformed the two competing approaches. The performance difference was most noticeable on those synthetic data where the

individual data sets corresponded to different activation levels of the regulatory subpathways (owing to different settings of the interaction parameters). The difference was less pronounced when only adding corrupted data to data from homogeneous experimental (cytometry data) or simulation (Netbuilder) conditions.

A problem intrinsic to Method B, the data integration coupling scheme of Sec. 3, is a deterioration of the convergence and mixing of the Markov chain. In fact, some of the results presented in the second study, described in Sec. 7, were obtained from MCMC simulations that had incompletely converged, suggesting that the performance improvement achieved with coupling could be further improved upon proper convergence. Unfortunately, this aspect has to be left to future research; the present paper was invited to be submitted to a special issue of the journal, which required the study to be completed by a given deadline. As future research, we will explore novel proposal moves, which swap substructures between the individual networks and the hypernetwork, allowing the latter to change in a more systematic way at (hopefully) a higher acceptance probability. The running of parallel Metropolis-coupled Markov chains, as described in Geyer[33] and Gilks *et al.*[34] and successfully applied in phylogenetics,[35] will also be attempted, especially as it will allow the exploitation of modern PC clusters, and might offer ways to more efficiently design highly accepted proposal moves based on information obtained from the whole population of Markov chains.[36]

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

Acknowledgments

A. Werhli was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). D. Husmeier was supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD).

References

1. Friedman N, Linial M, Nachman I, Pe'er D, Using Bayesian networks to analyze expression data, *J Comput Biol* **7**:601–620, 2000.
2. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA, Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks, *Pac Symp Biocomput* **6**:422–433, 2001.
3. Heckerman D, A tutorial on learning with Bayesian networks, in Jordan MI (ed.), *Learning in Graphical Models*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, pp. 301–354, 1999.
4. Husmeier D, Dybowski R, Roberts S, *Probabilistic Modeling in Bioinformatics and Medical Informatics*, Advanced Information and Knowledge Processing, Springer, New York, 2005.
5. Krause PJ, Learning probabilistic networks, *Knowledge Eng Rev* **13**:321–351, 1998.
6. Geiger D, Heckerman D, Learning Gaussian networks, in *Proc 10th Conf on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, pp. 235–243, 1994.
7. Pournara I, Wernisch L, Reconstruction of gene networks using Bayesian learning and manipulation experiments, *Bioinformatics* **20**:2934–2942, 2004.

8. Werhli AV, Grzegorczyk M, Husmeier D, Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks, *Bioinformatics* **22**:2523–2531, 2006.
9. Madigan D, York J, Bayesian graphical models for discrete data, *Int Stat Rev* **63**:215–232, 1995.
10. Hastings WK, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**:97–109, 1970.
11. Imoto S, Higuchi T, Goto T, Kuhara S, Miyano S, Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks, in *Proc IEEE Computer Society Bioinformatics Conference* (CSB'03), pp. 104–113, 2003.
12. Imoto S, Higuchi T, Goto T, Miyano S, Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks, *Stat Method* **3**(1):1–16, 2006.
13. Nariai N, Kim S, Imoto S, Miyano S, Using protein–protein interactions for refining gene networks estimated from microarray data by Bayesian networks, *Pac Symp Biocomput* **9**:336–347, 2004.
14. Tamada Y, Bannai H, Imoto S, Katayama T, Kanehisa M, Miyano S, Utilizing evolutionary information and gene expression data for estimating gene networks with Bayesian network models, *J Bioinform Comput Biol* **3**(6):1295–1313, 2005.
15. Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, Miyano S, Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection, *Bioinformatics* **19**:ii227–ii236, 2003.
16. Balian R, *From Microphysics to Macrophysics. Methods and Applications of Statistical Physics*, Vol. 1, Springer-Verlag, Berlin, 1982.
17. Friedman N, Koller D, Being Bayesian about network structure, *Mach Learn* **50**:95–126, 2003.
18. Husmeier D, Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks, *Bioinformatics* **19**:2271–2282, 2003.
19. Werhli A, Husmeier D, Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge, *Stat Appl Genet Mol Biol* **6**(1):article 15, 2007.
20. Cowles MK, Carlin BP, Markov chain Monte Carlo convergence diagnostics: A comparative review, *J Am Stat Assoc* **91**:883–904, 1996.
21. Werhli AV, Reconstruction of gene regulatory networks from postgenomic data, Ph.D. thesis, Biomathematics & Statistics Scotland (BioSS) and University of Edinburgh, 2007.
22. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP, Protein-signaling networks derived from multiparameter single-cell data, *Science* **308**:523–529, 2005.
23. Dougherty MK, Müller, J, Ritt DA, Zhou M, Zhou XZ, Copeland TD, Conrads TP, Veenstra TD, Lu KP, Morrison DK, Regulation of Raf-1 by direct feedback phosphorylation, *Mol Cell* **17**:215–224, 2005.
24. Atkins PW, *Physical Chemistry*, 3rd ed., Oxford University Press, Oxford, 1986.
25. Yang C-R, Shapiro BE, Mjolsness ED, Hatfield GW, An enzyme mechanism language for the mathematical modeling of metabolic pathways, *Bioinformatics* **21**(6):774–780, 2005.
26. Yuh CH, Bolouri H, Davidson EH, Genomic *cis*-regulatory logic: Experimental and computational analysis of a sea urchin gene, *Science* **279**:1896–1902, 1998.
27. Yuh CH, Bolouri H, Davidson EH, *cis*-regulatory logic in the endo16 gene: Switching from a specification to a differentiation mode of control, *Development* **128**:617–629, 2001.

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

28. Kanehisa M, A database for post-genome analysis, *Trends Genet* **13**:375–376, 1997.
29. Kanehisa M, Goto S, KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res* **28**:27–30, 2000.
30. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M, From genomics to chemical genomics: New developments in KEGG, *Nucleic Acids Res* **34**:D354–357, 2006.
31. Schäfer J, Strimmer K, An empirical Bayes approach to inferring large-scale gene association networks, *Bioinformatics* **21**(6):754–764, 2005.
32. Schäfer J, Strimmer K, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Stat Appl Genet Mol Biol* **4**(1):article 32, 2005.
33. Geyer CJ, Markov chain Monte Carlo maximum likelihood, in Keramidas EM (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Interface Foundation, Fairfax Station, VA, pp. 156–163, 1991.
34. Gilks WR, Richardson S, Spiegelhalter DJ, Strategies for improving MCMC, in Gilks WR, Roberts GO (eds.), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, Suffolk, UK, pp. 89–114, 1996.
35. Huelsenbeck JP, Ronquist F, Mr Bayes: Bayesian inference of phylogenetic trees, *Bioinformatics* **17**:754–755, 2001.
36. Laskey KB, Myers JW, Population Markov chain Monte Carlo, *Mach Learn* **50**(1–2):175–196, 2003.

J. Bioinform. Comput. Biol. 2008.06:543-572. Downloaded from www.worldscientific.com
by FUNDACAO UNIVERSIDADE DO RIO GRANDE on 11/20/14. For personal use only.

Adriano V. Werhli graduated in Physics at the University of Unisinos, Brazil, in 1998. After gaining industrial experience at a petrochemical plant (Copesul), he received his Master's degree in Applied Computing from the same university in 2003. He was a postgraduate student at Biomathematics and Statistics Scotland (BioSS) from 2004 to 2007, and received his Ph.D. in Informatics from the University of Edinburgh, UK, in October 2007. Dr. Werhli currently works as a postdoctoral research fellow at Pontifícia Universidade Católica Rio Grande do Sul, Brazil. His area of research interest is in bioinformatics, particularly the application of machine learning techniques to the discovery of pathways and regulatory networks.

Dirk Husmeier graduated in Physics (Dipl.-Phys.) at the University of Bochum (Germany) in 1991, and received both his M.Sc. (Information Processing and Neural Networks) and Ph.D. (Applied Mathematics and Neural Computation) degrees from King's College London in 1994 and 1997, respectively. After working as a postdoctoral research fellow in the Electrical Engineering Department of Imperial College London from 1997 to 1999, he joined Biomathematics and Statistics Scotland (BioSS) as a research scientist in October 1999. Since 2006, Dr. Husmeier has been leading the statistical bioinformatics research theme at BioSS.