

Systems biology

Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks

Adriano V. Werhli^{1,2,*}, Marco Grzegorzczak³ and Dirk Husmeier¹¹Biomathematics and Statistics Scotland, Edinburgh, UK, ²School of Informatics, University of Edinburgh, UK and ³Department of Statistics, University of Dortmund, Germany

Received on May 19, 2006; revised on July 7, 2006; accepted on July 10, 2006

Advance Access publication July 14, 2006

Associate Editor: John Quackenbush

ABSTRACT

Motivation: An important problem in systems biology is the inference of biochemical pathways and regulatory networks from postgenomic data. Various reverse engineering methods have been proposed in the literature, and it is important to understand their relative merits and shortcomings. In the present paper, we compare the accuracy of reconstructing gene regulatory networks with three different modelling and inference paradigms: (1) Relevance networks (RNs): pairwise association scores independent of the remaining network; (2) graphical Gaussian models (GGMs): undirected graphical models with constraint-based inference, and (3) Bayesian networks (BNs): directed graphical models with score-based inference. The evaluation is carried out on the Raf pathway, a cellular signalling network describing the interaction of 11 phosphorylated proteins and phospholipids in human immune system cells. We use both laboratory data from cytometry experiments as well as data simulated from the gold-standard network. We also compare passive observations with active interventions.

Results: On Gaussian observational data, BNs and GGMs were found to outperform RNs. The difference in performance was not significant for the non-linear simulated data and the cytoflow data, though. Also, we did not observe a significant difference between BNs and GGMs on observational data in general. However, for interventional data, BNs outperform GGMs and RNs, especially when taking the edge directions rather than just the skeletons of the graphs into account. This suggests that the higher computational costs of inference with BNs over GGMs and RNs are not justified when using only passive observations, but that active interventions in the form of gene knockouts and over-expressions are required to exploit the full potential of BNs.

Availability: Data, software and supplementary material are available from <http://www.bioss.sari.ac.uk/staff/adriano/research.html>.

Contact: adriano@bioss.ac.uk, dirk@bioss.ac.uk, Grzegorc@statistik.uni-dortmund.de

1 INTRODUCTION

Traditional approaches to systems biology are based on a mathematical description of putative pathways in terms of coupled differential equations with the objective to obtain a deeper understanding of the exact nature of the regulatory circuits and their regulation mechanisms. However, the availability of high-throughput

postgenomic data has recently prompted substantial interest in reverse engineering the networks and pathways in an inferential way from the data themselves. One of the first seminal papers promoting this approach aimed to learn gene regulatory networks in *Saccharomyces cerevisiae* from gene expression profiles with Bayesian networks (Friedman *et al.*, 2000). Since then, several authors have applied Bayesian networks to infer regulatory networks from postgenomic data of different nature (for instance, Imoto *et al.*, 2003a; Nariai *et al.*, 2005). Various alternative methods, like relevance networks (Butte and Kohane, 2003) and graphical Gaussian models (Schäfer and Strimmer, 2005a) have been proposed and applied to the inference of gene regulatory networks from gene expression data. Given the diversity of proposed reverse engineering methods, it is important for the systems biology community to obtain a better understanding of their relative strengths and weaknesses. One of the first major evaluation studies was carried out by Smith, *et al.* (2002). The authors simulated a complex biological system at different levels of organization, involving behaviour, neural anatomy, and gene expression of songbirds. They then tried to infer the structure of the known true genetic network from the simulated gene expression data with Bayesian networks. In a related study, Husmeier (2003) evaluated the accuracy of reverse engineering gene regulatory networks with Bayesian networks from data simulated from realistic molecular biological pathways, where the latter were modelled with a system of coupled differential equations. This network was also used in an earlier study by Zak *et al.* (2001), who investigated the inference accuracy of deterministic linear and log-linear models. While all three papers shed some light on the accuracy of reconstructing regulatory networks, they only investigated a particular inference method and do not include a cross-method comparison.

In order to address this shortcoming, an extensive evaluation study was carried out by Pournara (2005). The author compared graphical Gaussian models and Bayesian networks on synthetic data generated from networks with random structures and different gene regulation mechanisms, where the latter differed with respect to the cooperative or competitive interactions between transcription factors regulating the same gene. The approach we present in our paper is motivated by Pournara (2005) and complements this work in four important respects. First, the learning algorithm for Bayesian networks has been improved. In order to capture the uncertainty inherent in learning from sparse and noisy data, we sample network

*To whom correspondence should be addressed.

structures from the posterior distribution with MCMC. This approach is methodologically more consistent than the optimization scheme applied in Pournara (2005). For the practical realization, we apply a novel sampling strategy based on node orders (Friedman and Koller, 2003), which achieves faster mixing and convergence than conventional sampling in the space of network structures (Madigan and York, 1995). Second, we use improved inference methods for graphical Gaussian models. The approach adopted in Pournara (2005) is based on the PC algorithm of Spirtes *et al.* (2001). In the present work, we apply a more recent algorithm proposed by Schäfer and Strimmer (2005b), which the authors have developed after extensive experimentation with methods for stabilizing covariance matrix estimations (Schäfer and Strimmer, 2005a). Third, we include another reverse engineering method in our comparison: the approach of relevance networks proposed by Butte and Kohane (2000, 2003). This approach is appealing owing to its low computational costs, and we investigate to what extent the results can be improved with the more complex alternative algorithms mentioned above. Fourth, rather than evaluating the performance on randomly generated network structures, we base our comparison on the Raf pathway, a critical protein signalling network involved in regulating cellular proliferation in human immune system cells (Sachs *et al.*, 2005). Our evaluation exploits four types of data, distinguishing between passive observations and active interventions, and using data from both laboratory experiments as well as synthetic simulations. We have organized our paper as follows. After a brief review of the methods evaluated in our study (Section 2), we describe the data (Section 3) and simulation studies (Section 4) and justify our evaluation procedure (Section 5). We present our results in Section 6, followed by a discussion (Section 7) and the final conclusions (Section 8). Owing to space restrictions, some results, discussions and elaborations have been relegated to the Supplementary Material.

2 METHODS

We review briefly the three methods compared in our study. We conceive of a network as a generic interaction between nodes. Depending on the nature of the biological problem, these nodes may represent genes, proteins, metabolites, etc. The nodes are associated with experimentally observed measurements, like gene expression levels, protein concentrations or metabolic profiles.

2.1 Relevance Networks (RNs)

The method of RNs, proposed by Butte and Kohane (2000, 2003), is based on pairwise association scores. These scores are computed for all pairs of nodes from the signals associated with the nodes. The authors propose the mutual information and the Pearson correlation as appropriate association scores. This approach is straightforward to implement, and its computational costs are comparatively low. The principled disadvantage of RNs, however, is that the inference of an interaction between two nodes is not done in the context of the whole system. Consequently, we expect that RNs are not particularly powerful in distinguishing between direct (Fig. 1, left) and indirect (Fig. 1, centre) interactions.

2.2 Graphical Gaussian models (GGMs)

GGMs are undirected probabilistic graphical models that allow the identification of conditional independence relations among the nodes under the assumption of a multivariate Gaussian distribution of the data. The inference of GGMs is based on a (stable) estimation of the covariance matrix of this

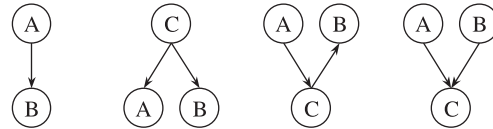


Fig. 1. Elementary interaction patterns. Left: Direct interaction between two nodes. Centre left: Regulation of two nodes by a common regulator. Centre right: Signalling chain via an intermediate regulator. Right: Coregulation of a node by two regulators (*v*-structure).

distribution. The element C_{ik} of the covariance matrix C is related to the correlation coefficient between nodes X_i and X_k . A high correlation coefficient between two nodes may indicate a direct interaction (Fig. 1, left), an indirect interaction (Fig. 1, centre right) or a joint regulation by a common (possibly unknown) factor (Fig. 1, centre left). However, only the direct interactions are of interest to the construction of a regulatory network. The strengths of these direct interactions are measured by the partial correlation coefficient ρ_{ik} , which describes the correlation between nodes X_i and X_k conditional on all the other nodes in the network. From the theory of normal distributions it is known that the matrix of partial correlation coefficients ρ_{ik} is related to the inverse of the covariance matrix C , C^{-1} (with elements C_{ik}^{-1}) (Edwards, 2000):

$$\rho_{ik} = - \frac{C_{ik}^{-1}}{\sqrt{C_{ii}^{-1} C_{kk}^{-1}}} \quad (1)$$

To infer a GGM, one typically employs the following procedure. From the given data, the empirical covariance matrix is computed, inverted and the partial correlations ρ_{ik} are computed from (1). The distribution of $|\rho_{ik}|$ is inspected, and edges (i, k) corresponding to significantly small values of $|\rho_{ik}|$ are removed from the graph. The critical step in the application of this procedure is the stable estimation of the covariance matrix and its inverse. In the present evaluation study, we apply the method of Schäfer and Strimmer (2005b). The authors propose a novel covariance matrix estimator regularized by a shrinkage approach after extensively exploring alternative regularization methods based on bagging (Schäfer and Strimmer, 2005a).

2.3 Bayesian networks (BNs)

BNs are directed graphical models for representing probabilistic relationships between multiple interacting entities. Formally, a BN is defined by a graphical structure M , 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 M 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. If 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 M of a BN has to be a directed acyclic graph (DAG), i.e. 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, \dots, X_n be a set of random variables represented by the nodes $i \in \{1, \dots, n\}$ in the graph, define $pa[i]$ to be the parents of node X_i , and let $X_{pa[i]}$ represent the set of random variables associated with $pa[i]$. Then

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | X_{pa[i]}). \quad (2)$$

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

$$P(M | D) \propto P(D | M)P(M) \quad (3)$$

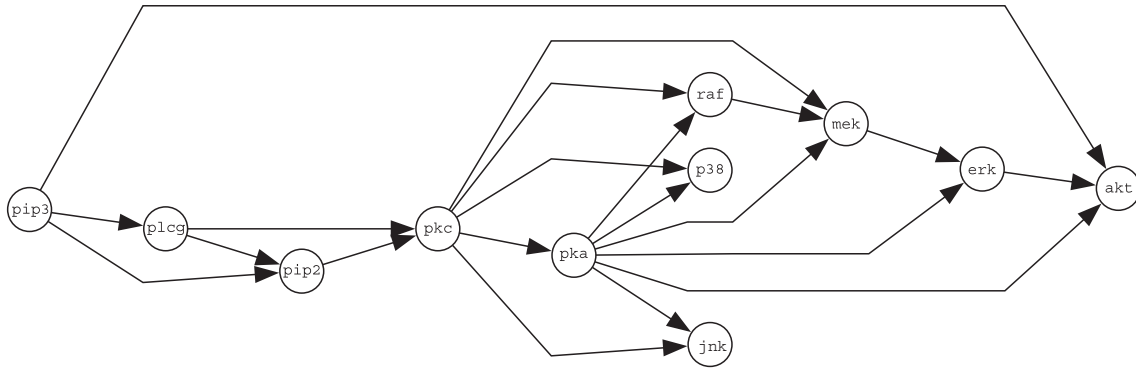


Fig. 2. Raf signalling pathway. The graph shows the currently accepted signalling network, taken from Sachs *et al.* (2005). Nodes represent proteins, edges represent interactions and arrows indicate the direction of signal transduction. In the interventional studies, the following nodes were targeted. Activations: PKA and PKC. Inhibitions: PIP2, AKT, PKC and MEK.

which requires a marginalization over the parameters q :

$$P(D|M) = \int P(D|q,M)P(q|M) dq \quad (4)$$

If certain regulatory conditions, discussed in Heckerman (1999), are satisfied and the data are complete, then the integral in (4) 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 $P(D|M)$ are usually referred to as the BDe (discretized data, multinomial distribution) or the BGe (continuous data, linear Gaussian distribution) score. Direct sampling from the posterior distribution (3) is analytically intractable, though. Hence, a Markov Chain Monte Carlo (MCMC) scheme is adopted (Madigan and York, 1995), for which an efficient proposal algorithm based on node orders has recently been proposed (Friedman and Koller, 2003). The final note in this brief summary concerns the problem of equivalence classes. Two Bayesian networks are equivalent if they show alternative ways of representing the same set of conditional independence relations. For instance, the two central-graphs in Figure 1 represent the same independence relation, namely, that nodes A and C are conditionally independent given B. This relation is different from the one shown on the right, where the two parent nodes are marginally independent—but conditionally dependent given the child. In general, it can be shown that networks are equivalent if they have the same skeleton and the same v-structure, where the latter denotes a configuration of two directed edges converging on the same node without an edge between the parents (Chickering, 1995). An equivalence class can be uniquely represented by a partially directed acyclic graph (PDAG), which is a graph that contains both directed and undirected edges with the former indicating that all network in the class concur about that edge direction. For instance, the PDAG corresponding to Figure 1 is a network in which all edges are undirected, except for the two edges in the v-structure on the right. For a more detailed discussion, see Chickering (1995).

2.4 Observational versus interventional data

Modern molecular biology possesses an extensive inventory of techniques for targeted interventions, for instance, knocking genes down with RNA interference or transposon mutagenesis. The consequence is that targeted nodes are no longer subject to the internal dynamics of the system under investigation, and the respective terms have to be excluded from the expansion in (2). This may break the symmetries within the equivalence classes; while equivalent structures have equal posterior probabilities under passive observations, this no longer holds when subjecting the system to external interventions. Consequently, edge directions that are ambiguous under

passive observations can be retrieved, and this forms the basis for learning putative causal interactions; see Pe'er *et al.* (2001) and Pournara and Wernisch (2004) for further details.

2.5 Comparison between the methods

GGMs and BNs potentially distinguish between direct and indirect interactions and therefore provide a more powerful modelling approach than RNs. BNs have the potential to present a more refined picture of interactions among nodes than GGMs owing to the directed nature of the edges; see the Supplementary Material for more details. Moreover, the inference procedure we adopt for learning BNs is score-based and more complex than the constraint-based approach adopted for GGMs [see Pournara (2005) for a comprehensive exposition of the difference between these two learning paradigms]. The latter approach aims to ‘explain away’ an observed correlation between two nodes by testing whether this correlation is not the effect of a regulation by other nodes. To this end, the partial correlations are computed, that is, the correlations conditional on all the other nodes in the system. This approach does not take into account whether network configurations that explain away these correlations are truly present. The score-based approach is in principle more powerful in that it marginalizes over all possible network configurations. However, the respective integral is analytically intractable, and the numerical approximation with MCMC is computationally expensive. In fact, the robust estimation of a rank-deficient covariance matrix proposed by Schäfer and Strimmer (2005b) turns constraint-based inference with GGMs into an extremely fast and attractive approach. Hence, the objective of the present study is to investigate whether the application of the more complex score-based approach to learning BNs is of any practical benefit for reverse engineering gene regulatory networks.

3 DATA

We base the evaluation of the three reverse engineering methods (RN, GGM and BN) on the Raf signalling network, depicted in Figure 2. 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 the pathway has therefore been extensively studied in the literature (e.g. Sachs *et al.*, 2005; Dougherty *et al.*, 2005). We use four types of data for our evaluation. First, we distinguish between passive observations and active interventions. Second, we use both real laboratory data as well as synthetic simulations. This combination of data is based on the following rationale. For simulated data, the true structure of the regulatory network is known; this allows us, in

principle, to faithfully evaluate the prediction results. However, the model used for data-generation is a simplification of real molecular-biological processes, and this might lead to systematic deviations and a biased evaluation. The latter shortcoming is addressed using real laboratory data. In this case, however, we ultimately do not know the true signalling network; the current gold-standard might be disputed in light of future experimental findings. By combining both approaches, we are likely to obtain a more reliable picture of the performance of the competing methods. Below, we will briefly summarize the main features of the data. For space restrictions we relegate a more detailed description to the Supplementary Material.

3.1 Linear Gaussian distribution

A simple synthetic way of generating data from the gold standard network of Figure 2 is to sample them from a linear-Gaussian distribution. The random variable X_i denoting the expression of node i is distributed according to

$$X_i \sim N\left(\sum_k w_{ik}x_k, \sigma\right), \quad (5)$$

where $N(\cdot)$ 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} . For simulating (noisy) interventions, we replaced the conditional distribution (5) by the following unconditional distributions. For inhibitions, we sampled X_i from a zero-mean Gaussian distribution, $N(0, \sigma)$. For activations, we sampled X_i from the tails of the empirical distribution of X_i , beyond the 2.5 and the 97.5 percentiles.

3.2 Realistic non-linear simulation

A more realistic simulation of data typical of signals measured in molecular biology is the following approach. The expression of a gene is controlled by the interaction of various transcription factors, which may have an inhibitory or activating influence. Ignoring time delays inherent in transcription and translation, these interactions can be compared with 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) (Atkins, 1986). 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 non-linear (sigmoidal) functions of combinations of substrates. However, instead of solving the steady-state approximation to ODEs explicitly, as pursued in Pournara (2005), we approximate the solution with a qualitatively equivalent combination of multiplications and sums of sigmoidal transfer functions. The resulting sigma- π formalism has been implemented in the software package Netbuilder (Yuh et al., 1998, 2001), which we have used for simulating the data from the gold standard Raf networks (see Supplementary Material for further information). To model stochastic influences, we subjected all nodes to additive Gaussian noise, and repeated the simulations for three different noise levels. Interventions were simulated by drawing values from a peaked Gaussian distribution ($\sigma = 0.01$) around the maximum (activation) and minimum (inhibition) values of the domain.

3.3 Cytometry data

Sachs et al. (2005) have applied intracellular multicolour flow cytometry experiments to quantitatively measure protein expression levels. Data were collected after a series of stimulatory cues and inhibitory interventions targeting specific proteins in the Raf pathway. A summary is given in the caption of Figure 2; see Sachs et al. (2005) for a more detailed description. The data are available from the following website: <http://www.sciencemag.org/cgi/content/full/308/5721/519/DC1>.

3.4 Dataset size

Flow cytometry allows the simultaneous measurement of the protein expression levels in thousands of individual cells. Sachs et al. (2005) have shown that for such a large dataset, it is possible to reverse engineer a network that is very similar to the known gold standard Raf signalling network. However, for many other types of current postgenomic data, such abundance of data is not available. We therefore sampled the data of Sachs et al. (2005) down to 100 data points; this is a representative figure for the typical number of different experimental conditions in current microarray experiments. We averaged the results over five independent samples. We used the same sample size and the same number of replications for the synthetic data. For observational data, all nodes were unperturbed. Interventional data were obtained by perturbing each of the six target nodes (described in the caption of Fig. 2) in turn, taking 14 measurement for each type of intervention, and including a further set of 16 unperturbed measurements.

4 SIMULATIONS

As opposed to GGMs, RNs and BNs do not require the assumption of a Gaussian distribution. However, deviations from the Gaussian incur an information loss as a consequence of data discretization (mutual information for RNs, BDe score for BNs). Alternatively, when avoiding the discretization with the heteroscedastic regression approach of Imoto et al. (2003b), the integral in (4) becomes analytically intractable and has to be approximated. It would obviously be interesting to evaluate the merits and shortcomings of these non-linear approaches. However, the main objective of the present study is the comparison of three modelling paradigms: (1) pairwise association scores independent of all other nodes (RNs), (2) undirected graphical models with constraint-based inference (GGMs) and (3) directed graphical models with score-based inference (BNs). To avoid the perturbing influence of additional decision factors, e.g. related to data discretization, and to enable a fair comparison with GGMs, we use the Gaussian assumption throughout. To minimize the deviation from this assumption, we subjected the data to a quantile normalization, ensuring that all marginal distributions of individual nodes were Normal.

Applying the Gaussian assumption to BNs, with the normal-Wishart distribution as a conjugate prior on the parameters, the integral in (4) has a closed-form solution, referred to as the BGe score. Details are given in Geiger and Heckerman (1994). The score depends on various hyperparameters, which can be interpreted as pseudocounts from a prior network. To make the prior probability over parameters— $P(q|M)$ in Equation (4)—as uninformative as possible, we set the prior network to a completely unconnected graph with an equivalent sample size as small as possible subject to the constraint that the covariance matrix is non-singular. For the

prior over network structures— $P(M)$ in Equation (3)—we followed Friedman and Koller (2003) and chose a distribution that is uniform over parent cardinalities subject to a fan-in restriction of 3. We carried out MCMC over node orders, as proposed in Friedman and Koller (2003). To test for convergence, each MCMC run was repeated from two independent initializations. Consistency in the marginal posterior probabilities of the edges was taken as indication of sufficient convergence. We found that a burn-in period of 20 000 steps was usually sufficient, and followed this up with a sampling period of 80 000 steps, keeping samples in intervals of 200 MCMC steps. For RNs, we computed the pairwise node associations with the Pearson correlation. We computed the covariance matrix in GGMs with the shrinkage approach proposed by Schäfer and Strimmer (2005b), choosing a diagonal matrix as the shrinkage target. Note that this target corresponds to the empty prior network; hence the effect of shrinkage is equivalent to the selected prior for the computation of the BGe score in BNs. The practical computations were carried out with the software provided by Schäfer and Strimmer (2005b). The MCMC simulations were carried out with our own MATLAB programs, which are available from our website.

5 EVALUATION

While the true network is a directed graph, our reconstruction methods may lead to undirected, directed, or partially directed graphs. To assess the performance of these methods, we apply 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.

Each of the three reverse engineering methods compared in our study leads to a matrix of scores associated with the edges in a network. These scores are of different nature: correlation coefficients for RNs, partial correlation coefficients for GGMs and marginal posterior probabilities for BNs. However, all three 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 obtained from applying several methods to many datasets. 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 scores indicating, overall, a better performance. While this approach does not require us to commit ourselves to the adoption of any (arbitrary) decision criterion, it does not lead to a specific network prediction. It also ignores the fact that, in practice, one is particularly interested in the performance for low FP rates. Our second performance criterion, hence, is based on the selection of a threshold on the edge scores, from which a specific network prediction is obtained. The question, then, is how to define this threshold. Schäfer and Strimmer (2005a) discuss a method for

converting the (partial) correlation coefficients of RNs and GGMs into q -values [i.e. p -values corrected for multiple testing; see Storey and Tibshirani (2003)] and ‘posterior probabilities’. However, these posterior probabilities are not equivalent to those defined for BNs. Imposing the same threshold on both leads to different rates of TPs and FPs, and hence different operating points on the ROC curves. We also found that controlling the false discovery rate at the typical value of $q = 0.05$ turned out to be too conservative; the numbers of predicted edges were very low, and sometimes zero. We therefore chose the threshold such that it led to a fixed count of five FPs. This procedure is guaranteed to compare the competing methods at the same operation point on the ROC curves, and the evaluation can therefore simply be based on the TP counts.

6 RESULTS

For a concise summary, we present our results visually in terms of scatter plots. A complete set of tables is available from our supplementary material.

Figure 3 compares the performance of BNs and GGMs on the synthetic Gaussian data and the protein concentrations from the cytometry experiment. The two panels on the left refer to the Gaussian data. Without interventions, BNs and GGMs achieve a similar performance in terms of both AUC and TP scores. Interventions lead to improved predictions with BNs. As a consequence of interventions, the number of correctly predicted undirected edges increases slightly from 15.8 to 18.5; this is not significant, though ($p = 0.097$). However, the number of correctly predicted directed edges shows a significant increase from 4.9 to 18.4 ($p < 10^{-4}$). On the intervened data, BNs outperform GGMs, and this improvement is significant when the edge directions are taken into account (AUC: $p = 0.0002$, TP: $p = 0.0005$).

The two columns on the right of Figure 3 summarize the results obtained for the cytometry data. Without interventions, GGMs and BNs show a similar performance. As a consequence of interventions, the performance of BNs improves, but less substantially than for the Gaussian data. For instance, the number of correctly predicted directed edges increases from 3.3 to 6.9, which is just significant ($p = 0.013$). With interventions, BNs tend to outperform GGMs. This improvement is only significant for the DGE-TP score, though ($p = 0.007$); while the UGE-AUC score for BNs is consistently better than for GGMs, its p -value of 0.055 is above the standard significance threshold.

To obtain a deeper understanding of the models’ performance, we applied them to the non-linear simulated data (Netbuilder) with different noise levels. The results are shown in Figure 4. When comparing the performance of BNs and GGMs on observational data, we observe the following trend. For low noise levels, GGMs slightly outperform BNs, although this difference is only significant for the DGE-TP score ($p = 0.008$); all other p -values are >0.05 . When increasing the noise level, the situation is reversed. BNs outperform GGMs, and the differences are significant for all scores except for DGE-TP (UGE-AUC: $p = 0.025$, DGE-AUC: $p = 0.029$, UGE-TP: $p = 0.016$, DGE-TP: $p = 0.067$). For large noise levels, GGMs and BNs show a similar performance, without a significant difference in any score. Interventions lead to an improvement in the performance of BNs when taking the edge direction into account. The improvement is significant in both scores, DGE-TP

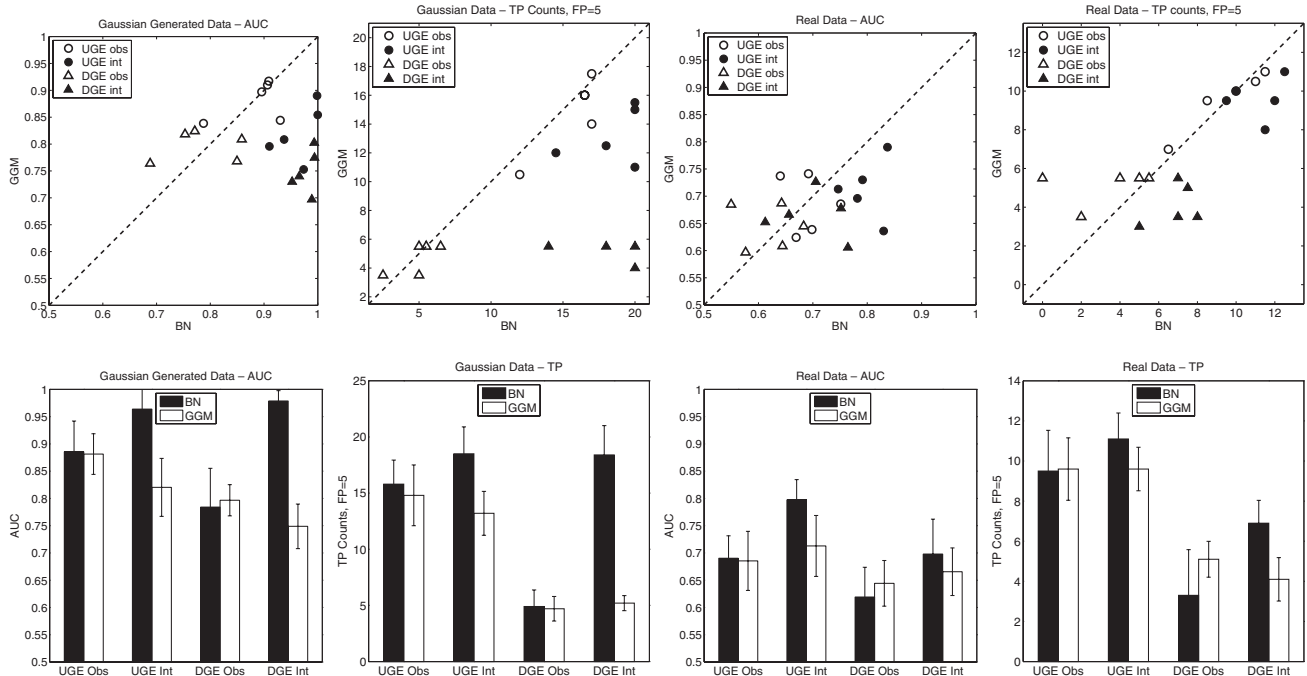


Fig. 3. GGMs versus BNs on Gaussian and cytometry data. Top row: Scatterplots comparing the performance of GGMs (vertical axis) with BNs (horizontal axis). The diagonal line represents equal performance. Symbols above that line indicate that GGMs outperform BNs. Conversely, symbols below that line point to a better performance of BNs over GGMs. Each subfigure compares the results obtained from two different data types, using only passive observations (empty symbols) and including active interventions (filled symbols). Two different evaluation criteria have been applied, based on directed graphs (DGE, represented by triangles) and their undirected skeletons (UGE, represented by circles). Bottom row: Histograms showing the average AUC scores and TP counts for BNs (filled bars) and GGMs (empty bars). The codes under the histograms indicate the type of evaluation (UGE versus DGE) and whether observational (Obs) or interventional (Int) data have been used. Columns: The four columns refer to different data and scoring criteria. Left: Gaussian data, AUC score. Centre left: Gaussian data, TP counts. Centre right: Cytometry data, AUC score. Right: Cytometry data, TP counts.

and DGE-AUC, for all noise levels, with $p < 0.002$. The improvement is most pronounced for the medium noise level, where the number of correctly predicted edges increases from 7.2 to 17.3 ($p \ll 10^{-4}$). A comparison between GGMs and BNs reveals that with interventions, BNs consistently outperform GGMs when taking the edge direction into account; all differences are significant with ($p < 0.005$).

Figure 2 of the Supplementary Material and Figure 5 compare the performance of BNs and GGMs with RNs. On the Gaussian observational data, both GGMs and BNs consistently outperform RNs. However, there is no significant difference in the performance of the methods on the nonlinear simulated data (Netbuilder) and the cytoflow protein concentrations when no interventions are used; in fact, the DGE-TP scores for BNs are actually worse than those obtained with RNs (see the next section for a discussion). With interventions, GGMs outperform RNs on the cytometry data (UGE: $p = 0.001$, DGE: $p = 0.001$), and they obtain higher TP counts than RNs on the non-linear simulated data ($p < 0.0002$ for both UGE and DGE). BNs consistently outperform RNs on all datasets with respect to all scoring schemes when interventions are used ($p < 0.001$).

7 DISCUSSION

Dependence on the noise level. When varying the noise level on the non-linear simulated data (Fig. 4) we observe that when increasing

the noise level, the performance with BNs first increases, and then decreases. For instance, the average number of predicted true undirected edges increases from TP = 11 for $\sigma = 0.01$ to TP = 18 for $\sigma = 0.1$, and then decreases again to TP = 15.5 for $\sigma = 0.3$. To understand this behaviour, consider a parent node that regulates several children, where the children do not have any direct interactions; see Figure 1, centre left. Without noise, the response of each child is a deterministic function of the parent. However, this implies a deterministic functional relationship between the children. Consequently, the true network cannot be distinguished from a network in which all children are connected by edges, and it is intrinsically impossible to learn the true network. The deterministic relationship between the children is destroyed by the addition of noise, which renders, on average, the signal of a child more similar to that of its parent than that of a sibling. Consequently, some noise is useful and forms the basis for learning gene regulatory networks from data. However, when the noise level becomes so large that it hides the regular signal, successful learning will no longer be feasible. Hence, we would expect the accuracy of reconstructing regulatory networks to first increase and then decrease with increasing noise level, and this trend is confirmed in our simulations.

GGMs versus RNs. To better understand the different performance of GGMs and RNs, we computed the average posterior probabilities of the true and false edges from the (partial) correlation coefficients according to the scheme described in Schäfer

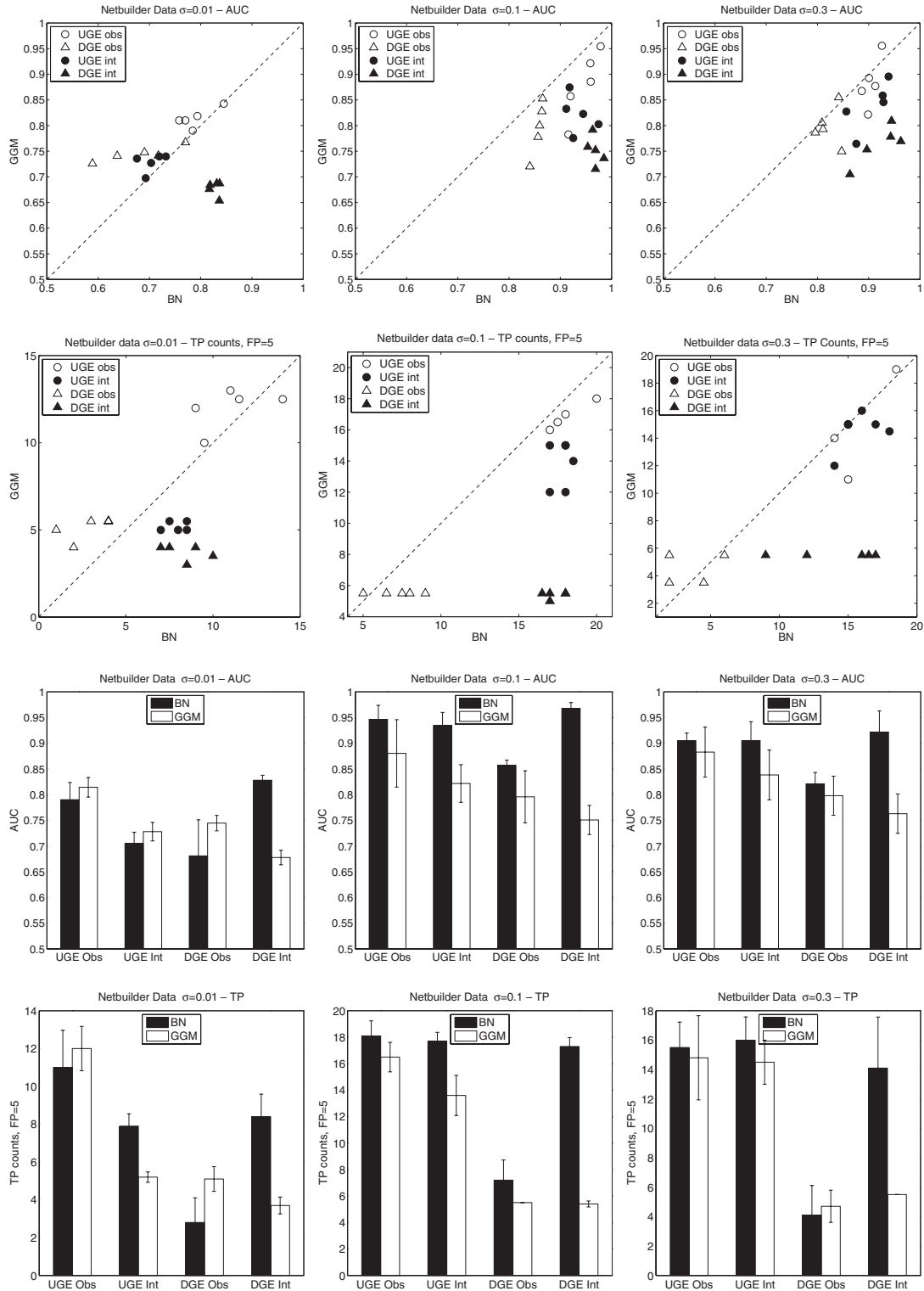


Fig. 4. GGMs versus BNs on data simulated with Netbuilder. This figure compares the performance of GGMs and BNs on the synthetic data generated with Netbuilder. The columns refer to different standard deviations of the additive Gaussian noise. Left column: $\sigma = 0.01$. Centre column: $\sigma = 0.1$. Right column: $\sigma = 0.3$. The top panel (top two rows) shows scatterplots of GGM scores plotted against BN scores; a detailed explanation of the symbols is given in the caption of Figure 3. The bottom panel (bottom two rows) shows histograms with average AUC scores and TP counts for BNs (filled bars) and GGMs (empty bars); see the caption of Figure 3 for further explanations. In each panel, the two rows refer to different scoring criteria, discussed in Section 5. Top row: AUC score. Bottom row: TP count.

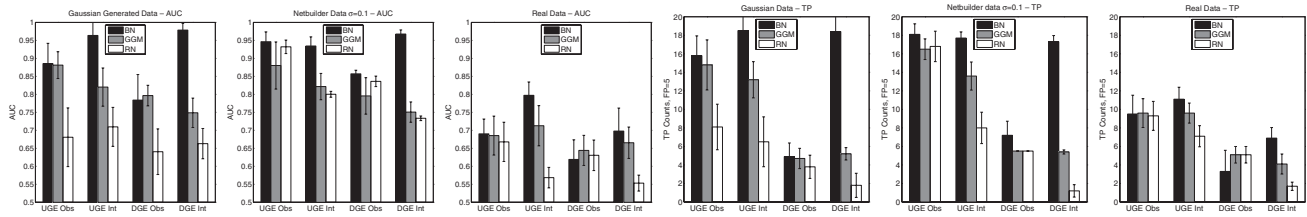


Fig. 5. Cross-data comparison between BNs, GGMs and RNs. The histograms show the average AUC scores and TP counts for BNs (black bars), GGMs (grey bars) and RNs (white bars). The codes under the histograms indicate the type of evaluation (UGE versus DGE) and whether observational (Obs) or interventional (Int) data have been used. The six panels refer to different data and scoring criteria. From left to right: (1) Gaussian data, AUC; (2) Netbuilder data, AUC; (3) Cytometry data, AUC; (4) Gaussian data, TP; (5) Netbuilder data, TP; (6) Cytometry data, TP.

and Strimmer (2005a). The results are shown in Table 2 of the Supplementary Material and suggest that GGMs show a clearer separation of the true and false edges than RNs. This difference has not translated itself into an improved performance of GGMs over RNs in terms of AUC and TP scores for the un-intervened non-Gaussian data. The reason is that although the separation between the scores is poorer for RNs than for GGMs, it has not affected the ranking of the edges. However, this finding suggests that inference with RNs is less stable than with GGMs. In fact, for interventions, RNs show a more substantial degradation in their performance than GGMs; GGMs consistently outperform RNs on the intervened cytoflow data ($p < 0.021$), and obtain significantly higher TP counts on the non-linear simulated data ($p < 10^{-4}$).

Interventions for low noise level. The left column of Figure 4 reveals a curious finding for the low-noise scenario: on interventions, the UGE score for BNs deteriorates. As discussed above, the ability to suppress spurious associations between unconnected nodes deteriorates for low noise levels. Interventions reduce the average noise level; so if the noise is already very low, this further reduction in the noise may lead to the prediction of spurious associations. The deterioration of the UGE (as opposed to the DGE) score can be explained by the fact that a spurious undirected edge is equivalent to two spurious directed edges (since there are twice as many directed as undirected edges in the graph), and that the UGE score does not benefit from any corrections of edge directions that result from the interventions.

Dependence on the network topology. We investigated the influence of the network topology as follows. We removed four edges from the graph to create four v-structures, and reran the whole analysis. Since undirected graphs intrinsically cannot represent v-structures, as discussed in the Supplementary Material, we would expect an increase in the performance of BNs relative to GGMs. Owing to space restrictions we have relegated the details of this study to the Supplementary Material. The findings were, overall, similar to the results obtained on the original network. On the observational linear-Gaussian data, the comparison of BNs versus GGMs showed a significant shift in favour of BNs, with $p < 0.05$ for all performance scores; this confirms our hypothesis. There was no significant difference between the performance scores of BNs and GGMs on the non-linear data generated with Netbuilder, though.

Learning directed graphs from the cytometry data. BNs obtain poorer DGE scores on the un-intervened cytometry data than RNs and GGMs, while there is no significant difference in the UGE scores. This suggests that while BNs learn the skeleton of the network as accurately as GGMs and RNs, some of the edge directions

are systematically inverted. A possible explanation are errors in the gold standard network. In fact, a recent publication (Dougherty *et al.*, 2005) reports evidence for negative feedback loops, which are not included in the gold standard network of Sachs *et al.* (2005). Such feedback could explain systematic deviations between the predicted and the ‘gold standard’ network. Negative feedback is also known to have a stabilizing effect with respect to interventions; this might explain why the improvement in the DGE scores for BNs is less pronounced than for the simulated data. 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.

8 CONCLUSION

Our main findings can be summarized as follows. BNs and GGMs tend to outperform RNs, but the difference is less pronounced for the non-linear simulated data (Netbuilder) and the measured protein concentrations (cytometry experiments) than for Gaussian data. Also, there is insufficient evidence for any significant difference between BNs and GGMs on observational data. These findings are different from those reported in Pournara (2005), which seems to result from the improved inference algorithm for GGMs (Schäfer and Strimmer, 2005b). However, for interventional data, BNs outperform GGMs and RNs when taking the edge directions into account. This suggests that the higher computational costs of inference with BNs over GGMs and RNs are not justified for passive observations, but that active interventions in the form of gene knockouts and over-expressions are required to exploit the full potential of BNs.

ACKNOWLEDGEMENTS

The authors are grateful to Karen Sachs, Douglas Armstrong, Jane Hillston and Peter Ghazal for helpful discussions. A.W. is supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). M.G. is supported by the Bioforschungsbund Dortmund. D.H. is supported by the Scottish Executive Environmental and Rural Affairs Department (SEERAD).

Conflict of Interest: none declared

REFERENCES

- Atkins, P.W. (1986) *Physical Chemistry*, 3rd edn. Oxford University Press, Oxford.
- Butte, A.S. and Kohane, I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, 418–429.

- Butte,A.S. and Kohane,I.S. (2003) Relevance networks: a first step toward finding genetic regulatory networks within microarray data. In Parmigiani,G., Garrett,E.S., Irizarry,R.A. and Zeger,S.L. (eds), *The Analysis of Gene Expression Data*, Springer, New York, pp. 428–446.
- Chickering,D.M. (1995) A transformational characterization of equivalent Bayesian network structures. *Int. Conf. Uncertain. Artif. Intell.*, **11**, 87–98.
- Dougherty,M.K. *et al.* (2005) Regulation of raf-1 by direct feedback phosphorylation. *Mol. Cell*, **17**, 215–224.
- Edwards,D.M. (2000) *Introduction to Graphical Modelling*. Springer Verlag, New York.
- Friedman,N. and Koller,D. (2003) Being Bayesian about network structure. *Mach. Learn.*, **50**, 95–126.
- Friedman,N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.
- Geiger,D. and Heckerman,D. (1994) Learning Gaussian networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA, Morgan Kaufmann, pp. 235–243.
- Heckerman,D. (1999) A tutorial on learning with Bayesian networks. In Jordan,M.I. (ed.), *Learning in Graphical Models, Adaptive Computation and Machine Learning*. MIT Press, Cambridge Massachusetts, pp. 301–354.
- Heckerman,D. *et al.* (1995) Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.*, **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.
- Imoto,S. *et al.* (2003a) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proc. IEEE Comput. Soc. Bioinform. Conf.*, 104–113.
- Imoto,S. *et al.* (2003b) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. Bioinform. Comput. Biol.*, **1**, 231–252.
- Madigan,D. and York,J. (1995) Bayesian graphical models for discrete data. *Int. Stat. Rev.*, **63**, 215–232.
- Nariai,N. *et al.* (2005) Estimating gene regulatory networks and protein-protein interactions of *Saccharomyces cerevisiae* from multiple genome-wide data. *Bioinformatics*, **21** (Suppl 2), ii206–ii212.
- Pe’er,D. *et al.* (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17**, S215–S224.
- Pournara,I.V. (2005) Reconstructing gene networks by passive and active Bayesian learning. PhD thesis. Birbeck College, University of London, UK.
- Pournara,I.V. and Wernisch,L. (2004) Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinformatics*, **20**, 2934–2942.
- Sachs,K. *et al.* (2005) Protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.
- Schäfer,J. and Strimmer,K. (2005a) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.
- Schäfer,J. and Strimmer,K. (2005b) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article 32.
- Smith,V.A. *et al.* (2002) Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics*, **18**, S216–S224.
- Spirtes,P. *et al.* (2001) *Causation, Prediction, and Search*, Springer Verlag, New York.
- Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomwide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
- Yuh,C.H. *et al.* (1998) Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**, 1896–1902.
- Yuh,C.H. *et al.* (2001) *Cis*-regulatory logic in the *endo16* gene: switching from a specification to a differentiation mode of control. *Development*, **128**, 617–629.
- Zak,D.E. *et al.* (2001) Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data. In *Proceedings of the Second International Conference on Systems Biology*, Pasadena, CA, pp. 231–238.