Euphytica manuscript No.
(will be inserted by the editor)

Dynamic Bayesian networks in molecular plant science: Inferring gene regulatory networks from multiple gene expression time series

Frank Dondelinger · Dirk Husmeier · Sophie Lèbre

Received: date / Accepted: date

Abstract To understand the processes of growth and biomass production in plants, we ultimately need to elucidate the structure of the underlying regulatory networks at the molecular level. The advent of high-throughput postgenomic technologies has spurred substantial interest in reverse engineering these networks from data, and several techniques from machine learning and multivariate statistics have recently been proposed. The present article discusses the problem of inferring gene regulatory networks from gene expression time series, and we focus our exposition on the methodology of Bayesian networks. We describe dynamic Bayesian networks and explain their advantages over other statistical methods. We introduce a novel information sharing scheme, which allows us to infer gene regulatory networks from multiple sources of gene expression data more accurately. We illustrate and test this method on a set of synthetic data, using three different measures to quantify the network reconstruction accuracy. The main application of our method is related to the problem of circadian regulation in plants, where we aim to reconstruct the regulatory networks of nine circadian genes

F. Dondelinger
Biomathematics and Statistics Scotland
JCMB, EH9 3JZ, Edinburgh, UK
Tel.: +44-(0)131 650 7536
Fax: +44-(0)131 650 4901
E-mail: frankd@bioss.ac.uk

D. Husmeier
Biomathematics and Statistics Scotland
JCMB, EH9 3JZ, Edinburgh, UK
Tel.: +44-(0)131 650 7547
Fax: +44-(0)131 650 4901
E-mail: dirk@bioss.ac.uk

S. Lèbre
Bioinformatique Théorique
FDBT, LSIIT (UMR CNRS-ULP 7005)
Université de Strasbourg
Pôle API, Boulevard Sébastien Brant - BP 10413
67412 Illkirch, France
Tel.: +33-(0) 390 244 889
E-mail: s.lebre@unistra.fr

in *Arabidopsis thaliana* from four gene expression time series obtained under different experimental conditions.

Keywords Gene regulatory networks · reverse engineering · dynamic Bayesian networks · data integration · information sharing · *Arabidopsis thaliana* · circadian regulation

1 Introduction

Reconstructing gene regulatory networks is a problem of great importance in plant biology. Rapid development of sequencing and computer technology has lead to the complete sequencing and annotation of many important model organisms. In order to understand the functioning of an organism, the next major step is to identify which genes are expressed, in which conditions and to what extent. As gene expression is a complex process regulated at several stages in the synthesis of proteins, the identification of genes whose products function together in the cell is a major task of post-genomic approaches. Genes that encode transcription factors, signalling proteins and proteins involved in the phosphorylation of other proteins can all have an effect on gene expression, and hence on the expression levels of other genes. A gene regulatory network is the graphical abstract representation of these interactions.

There is no physical interaction between the DNA molecules of the genes in a regulatory network; rather, the genes interact via the proteins they encode. In this paper, we say that two genes interact if the level of expressed mRNA of one gene depends on the level of expressed mRNA of the other via one of these mechanisms. For example, activation or inhibition via a transcription factor would have this effect. If we can understand the interactions among genes, then we can predict the effect that, for example, silencing a gene will have on other genes.

Knowing the gene regulatory networks gives us a deeper understanding of the working of plants at the molecular level and allows us to identify likely targets for genetic modifications. It also enables us to elicit gene functions by clustering the network into functional modules and looking at annotations for related genes. For instance, in Mochida et al (2011), the barley gene network was reconstructed by looking at co-expression of genes, which enabled the identification of functional modules for stress response and cell biogenesis, as well as finding modules that are specific to the *Triticeae* tribe, which contains domesticated crops such as rye, barley and wheat. A lot of research has centred on gene networks in Arabidopsis (Aoki et al, 2007; Ma et al, 2007; Morohashi and Grotewold, 2009), especially using the ATTED co-expression database (Obayashi et al, 2006; Okazaki et al, 2009). This research has translated to important crops such as barley (Sreenivasulu et al, 2008) and rice (Hamada et al, 2011; Jiao et al, 2009).

While many gene interactions can be elicited via experimental techniques such as DNA-level knockouts, or by extrapolating from the protein interactions found using yeast 2-hybrid, these methods are time-consuming and hampered by difficulties: yeast 2-hybrid experiments are noisy, while knockouts only look at one gene at a time. This means that traditional gene-by-gene approaches are not always sufficient. By analyzing the mRNA expression values obtained simultaneously for all genes from microarray experiments, we are able to point out possible transcriptional regulation in a gene regulatory network.

Early approaches to the analysis of microarray data focused mainly on correlations (Butte and Kohane, 2000; Moriyama et al, 2003). However, two genes that interact indirectly via a third gene will have a high correlation coefficient, despite the absence of any direct interaction. More sophisticated approaches use partial correlation coefficients (Schäfer and Strimmer, 2005) (which calculate the correlation between two genes conditional on all other genes) or sparse regression (Rogers and Girolami, 2005; van Someren et al, 2006).

In this paper, we apply dynamic Bayesian networks (DBNs), a probabilistic graphical model, to this problem. DBNs were first introduced for the analysis of gene expression time series by Friedman et al (1998) and Murphy and Mian (1999). DBNs can handle indirect interactions between genes by modeling conditional independence of nodes explicitly. Their probabilistic nature also allows them to capture the inherent uncertainty about which interactions are important. This uncertainty is not only a consequence of measurement errors during the microarray experiment. Transcription is not an on/off process; the presence of a transcription factor does not guarantee that transcription will take place. For example, the folding of the DNA may block access to the transcription factor binding site. Even if transcription is initiated, the process could be aborted, or result in an erroneous mRNA string which cannot be translated to a functional protein.

In the next section, we will present the model and explain how we can infer the structure of the gene regulatory network from gene expression data. We will also describe an extension of the model that can improve the inference if we are integrating data from different sources (e.g. from different labs or cell lines). We will then apply the methods to simulated data from a simple regression model, and to microarray data from Arabidopsis.

2 Materials and Methods

2.1 Dynamic Bayesian Network Model

As described in the previous section, our goal is to present a method that can:

– Infer gene interaction networks from gene expression data.
– Deal with uncertainty arising from measurement noise or intrinsic fluctuations.
– Share information between networks inferred from different datasets.

In this section we will describe the framework for dynamic Bayesian networks, and show how we can achieve all of these goals.

Framework To understand dynamic Bayesian networks, it is useful to first look at the more general framework of Bayesian networks. Bayesian networks are a probabilistic model for the joint distribution of a set of random variables: In our case, the genes, as represented by their gene expression values.

Let us say that we are looking at three genes involved in the circadian clock in *Arabidopsis thaliana* (the system that we are going to investigate in Section 3.2). Fig. 1 shows the putative network where *CCA1* regulates *TOC1* and *PRR3*. The directed interaction from gene *CCA1* to gene *TOC1* means that the expression level of *TOC1* (the child) directly depends on the expression level of gene *CCA1* (the parent). For

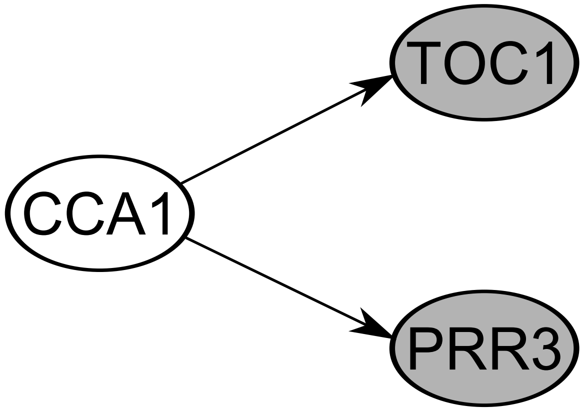


Fig. 1: Illustrative Example: Sub-network of the Arabidopsis circadian clock gene regulatory network. Gene *CCA1*, which is active in the morning, regulates the evening genes *TOC1* and *PRR3*.

example, *CCA1* could be a crucial transcription factor of *TOC1*. If there is no interaction between two genes, then their expression levels are independent given their parents. For example, the expression levels of *TOC1* and *PRR3* might be correlated or anti-correlated because they depend on the same transcription factor *CCA1*. However, conditional on the observation of the expression values of *CCA1* (and any other common regulators), the expression levels of *TOC1* and *PRR3* are no longer dependent: the nature of their apparent correlation has been explained away by the regulator. This example illustrates the practical benefit of BNs as a tool for representing direct regulations ($CCA1 \rightarrow TOC1$, $CCA1 \rightarrow PRR3$), and distinguishing them from indirect ones ($TOC1 \nrightarrow PRR3$).

A Bayesian network is defined by the joint probability distribution:

$$P(X_1, ..., X_N) = \prod_{i=1}^{N} P(X_i|\tau_i) \tag{1}$$

where each X_i is the expression value of one of the N genes in the network, and $\tau_i \subseteq \{X_1, ..., X_N\}$ denotes the set of parents of gene i[1]. Hence, a Bayesian network decomposes the joint probability of the expression values of all the genes in the network into the probability of the expression value of each gene, given its regulators. So the joint distribution of our example network in Fig. 1 would be expressed as $P(CCA1, TOC1, PRR3) = P(CCA1)P(TOC1|CCA1)P(PRR3|CCA1)$, allowing us to consider the conditional probability of each gene and its regulators separately.

The directed graph associated with this Bayesian network representation of the joint probability $P(X_1, ..., X_N)$ is a graph like Fig. 1, where the nodes are the genes X_i and where an interaction is drawn from each parent to its child in the conditional probability of equation (1). One important theoretical condition that needs to be satisfied is that there can be no directed cycles in a Bayesian network. So if *CCA1* regulates *TOC1* and *TOC1* regulates other genes, then none of these genes can in turn regulate *CCA1*. A BN

[1] For brevity, we will sometimes refer to X_i simply as a gene, even though strictly speaking it is the expression value of a gene.

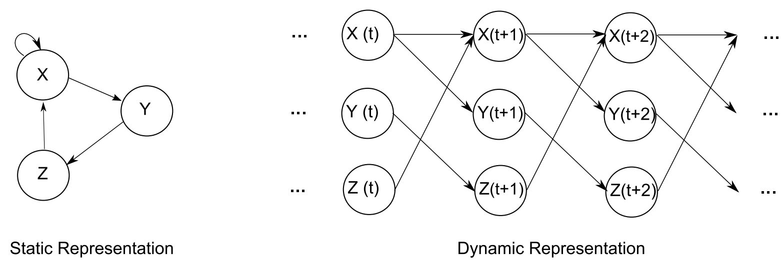


Fig. 2: Structure of a dynamic Bayesian network. The static representation shows the regulatory interactions between genes, while the dynamic representation shows the interactions over time, where X(t) represents the gene expression measurement of gene X at time t, and so on. Three genes are included in the network, and three time steps are shown in the dynamic representation.

is entirely defined by a directed acyclic graph (DAG) and the conditional probability of each node given its parents in the DAG (see e.g. Friedman et al (2000) for more details).

We are interested in learning the causal relationships between the interacting genes; e.g. the presence of transcription factor *CCA1* may inhibit the expression of *TOC1*. While a causal network of gene interactions would form a valid Bayesian network, the inverse relation may not hold: If we learn a network from the data, then the network we obtain does not necessarily represent the correct causal relationships. One reason could be the absence of a key regulator gene from the dataset: If measurements of *CCA1* are absent from the network in Fig. 1, then we may infer that there is a dependence between *TOC1* and *PRR3*, without being aware that it is the consequence of a common regulator. Even under the assumption of complete observation, finding causal relationships may be impeded by the fact that certain networks are equivalent. A simple example is two conditionally dependent genes, A and B, where the two networks A $\rightarrow$ B and A $\leftarrow$ B are equivalent. We can break these symmetries and determine the true causal relationships by using time series data. If gene A is always active before gene B becomes active, then A $\rightarrow$ B is likely to be the correct causal relationship. To deal with dynamic (time series) data, we use the framework of dynamic Bayesian networks (DBNs).

Dynamic Bayesian networks are based on the same principles as general Bayesian networks, but the genes in the network are organised in a special way. We assume that the network is subdivided into a sequence of time steps, each containing the same number of genes. The only connections that are allowed are those going from a gene in step t to a gene in step $t+1$, and we assume that the connections are homogeneous in time, meaning that the connections are the same between any two steps.

Dynamic Bayesian networks are well-suited for analysing time series data. Each time step will be a measurement in the time series, and interactions among genes link one time step to the next. This represents the fact that a change in the expression level of gene A will only show an effect on the expression level of interacting gene B after some time has elapsed. A consequence of introducing a time delay is that unlike general Bayesian networks, DBNs allow us to represent cycles. For example, in Fig. 2, the interaction between $X(t)$ and $X(t{+}1)$ corresponds to a self-loop on X (X regulating itself). $X \rightarrow Y \rightarrow Z \rightarrow X$ represents another regulatory cycle in the network.

Inference Inference aims to determine the structure of the dynamic Bayesian network directly from the data. As DBNs are a probabilistic framework, which allows for modelling uncertainty, we could do this by finding the model (or network) M with the highest probability given the data D, i.e. the network that maximises $P(M|D)$. However, it is usually preferable to look at the whole distribution of $P(M|D)$, because by restricting ourselves to the single 'best' network, we may neglect some interactions that occur less frequently, but still with regularity (for example, as a consequence of alternative pathways). To sample from the distribution, we use a process called Markov Chain Monte Carlo (MCMC), which creates a Markov chain of samples that is guaranteed to converge to a sample from the true distribution asymptotically. For more information on MCMC, see Madigan and York (1995); MacKay (1998). Once we have enough samples from $P(M|D)$, we can construct a network by looking at the proportion of occurrences of each interaction in the samples and retaining the interactions that are most likely.

To sample from $P(M|D)$ using MCMC, we need to be able to calculate a function proportional to $P(M|D)$ for any given network M. Bayes' rule allows us to express the posterior $P(M|D)$ using the likelihood of the data $P(D|M)$ and the prior probability of the network $P(M)$ as follows:

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

The prior $P(M)$ denotes the probability of the network before we have observed any data, and the likelihood $P(D|M)$ denotes how likely the data is given the model. The posterior $P(M|D)$ denotes the probability of the model after taking the observed data into account, by combining the likelihood and the prior.

The particular DBN model that we will consider here is based on Lèbre et al (2010) and assumes that the value $x_i(t)$ of each variable at time t is calculated as a weighted sum of the values of all variables at time $t-1$ (plus some noise). This is called a linear regression:

$$x_i(t) = \sum_{j=1}^{N} \{w_{ij} x_j(t-1)\} + \epsilon_i(t) \tag{3}$$

where w_{ij} is the weight given to the interaction between gene j and gene i if j is a parent of i in the current network and 0 otherwise. $\epsilon_i(t)$ denotes the value of a random variable drawn from $N(0, \sigma_i^2)$. This represents random Gaussian noise with variance σ_i^2, that will capture non-systematic measurement errors. The likelihood of observation x_i for gene i can then be expressed as:

$$P(x_i|M, \psi) = (2\pi(\sigma_i)^2)^{-T/2} exp\{-\frac{1}{2(\sigma_i)^2}(x_i - \hat{x}_i)^t (x_i - \hat{x}_i)\} \tag{4}$$

where T is the length of the time series, $\hat{x_i}$ is the estimated value of x_i using the regression model in (3) and M describes the network structure: The set of s_i parents τ_i for each gene i in the network. ψ describes the parameters of the regression model: the weights w_{ij} for the interactions between node i and its parents, and the noise level σ_i. Using the rules of probability, the prior $P(M, \psi)$ can be written as:

$$P(M, \psi) = P(M)P(\psi|M) \tag{5}$$

For more details, see Lèbre et al (2010). Lèbre et al. showed that it is possible to integrate out the regression parameters in ψ and obtain a closed-form solution that only depends on the network structure. This means that one does not have to worry about inferring the values of $w_i = (w_{ij})_{1 \leq j \leq N}$ and σ_i when inferring the other parameters.

Information Sharing So far, we have described how to infer a DBN from a single monolithic dataset. But what about the situation where we have different datasets reflecting different experimental conditions? Two possible approaches immediately suggest themselves: Either combine them into one dataset, ignoring their inherent differences, or learn separate networks for each dataset (and possibly use a majority voting scheme to combine them into one network).

The first approach is obviously sub-optimal, since it ignores the possibly detrimental effect of mixing more noisy datasets with less noisy ones. The second approach is preferable, but has the disadvantage that we are not using the whole of the available data for inferring the individual networks.

A better solution is to modify the second approach by introducing information sharing between the K individual networks, each network corresponding to a dataset obtained under specific conditions. This can be done most efficiently by assuming that the interactions are independent and the indicator variables for each interaction e_{ik}^n between genes i and k in network n follow a Bernoulli distribution with parameter θ_{ik}, so that:

$$P(e_{ik}^n|\theta_{ik}) = (\theta_{ik})^{e_{ik}^n}(1-\theta_{ik})^{1-e_{ik}^n} \tag{6}$$

which gives the probability of observing edge e_{ik}^n.

We can obtain an expression for $P(e_{ik}^n|\{e_{ik}^{\tilde{n}}\}_{\tilde{n}\neq n})$, the probability of e_{ik} in network n given the state of that interaction in all other networks $\tilde{n}$ with $\tilde{n} \neq n$, by integrating out θ_{ik}:

$$P(e_{ik}^n|\{e_{ik}^{\tilde{n}}\}_{\tilde{n}\neq n}) = \int P(e_{ik}^n|\theta_{ik})P(\theta_{ik}|\{e_{ik}^{\tilde{n}}\}_{\tilde{n}\neq n})d\theta_{ik} \tag{7}$$

where the rules of probability give that:

$$P(\theta_{ik}|\{e_{ik}^{\tilde{n}}\}_{\tilde{n}\neq n}) \propto P(\{e_{ik}^{\tilde{n}}\}_{\tilde{n}\neq n}|\theta_{ik})P(\theta_{ik}) \tag{8}$$

We now need to define $P(\{e_{ik}^{\tilde{n}}\}_{\tilde{n}\neq n}|\theta_{ik})$, that is to say, the probability of interaction e_{ik} in all networks except n. Note the difference between $P(\{e_{ik}^{\tilde{n}}\}_{\tilde{n}\neq n}|\theta_{ik})$ and $P(\theta_{ik}|\{e_{ik}^{\tilde{n}}\}_{\tilde{n}\neq n})$. The first describes the probability of interaction e_{ik} given a known parameter θ_{ik}; the second describes the probability of parameter θ_{ik} given the state of interaction e_{ik} in all networks except n.

Let us define B_{ik}^n as the number of networks (different from n) where edge e_{ik} is present, and $\overline{B_{ik}^n}$ is the number of networks where e_{ik} is absent. Because we have excluded network n, if we are learning K networks, then $B_{ik}^n + \overline{B_{ik}^n} = K - 1$. By assuming that the edges are independent, and using equation (6), we get a binomial distribution:

$$P(\{e_{ik}^{\tilde{n}}\}_{\tilde{n}\neq n}|\theta_{ik}) = \theta_{ik}^{B_{ik}^n}(1-\theta_{ik})^{\overline{B_{ik}^n}} \tag{9}$$

As for $P(\theta_{ik})$, the standard prior distribution for the parameter of a binomial distribution is the beta distribution. This is the conjugate prior of the binomial distribution, which means that if $P(\theta_{ik})$ is a beta distribution with parameters α_{ik} and $\overline{\alpha_{ik}}$, $P(\theta_{ik}|\{e_{ik}^{\tilde{n}}\}_{\tilde{n}\neq n})$ will also be a beta distribution with parameters $B_{ik}^n + \alpha_{ik}$ and

$\overline{B_{ik}^n} + \overline{\alpha_{ik}}$. By inserting (6) and (8) into (7) (and doing some simplifications), we can work out that:

$$P(e_{ik}^n = 1 | \{e_{ik}^{\tilde{n}}\}_{\tilde{n} \neq n}) = \frac{\alpha_{ik} + B_{ik}^n}{\alpha_{ik} + B_{ik}^n + \overline{\alpha_{ik}} + \overline{B_{ik}^n}} \tag{10}$$

$$P(e_{ik}^n = 0 | \{e_{ik}^{\tilde{n}}\}_{\tilde{n} \neq n}) = \frac{\overline{\alpha_{ik}} + \overline{B_{ik}^n}}{\alpha_{ik} + B_{ik}^n + \overline{\alpha_{ik}} + \overline{B_{ik}^n}} \tag{11}$$

We then simply replace the prior on the network structures $P(M)$ in equation (5), thus providing a way of sharing information between the networks learned from the different datasets. This approach is based on Ferrazzi et al (2008).

2.2 Simulation Model

In order to determine the suitability of the dynamic Bayesian network model, we want to compare the interactions that we infer from the data to the actual network of gene interactions. However, in most cases the true network will not be completely known. For this reason, we have implemented a simulation model to generate synthetic data from a known network.

The simulation model produces a time series of data points, each of which represents the normalised expression values of a gene. We start with a network M with the number of parents for each node drawn from a sparse Poisson prior (to keep the number of interactions low). Each directed interaction from gene A (the parent) to gene B (the child) has a weight that measures how much gene A will influence gene B. To ensure that the expression values stay at equilibrium, we test if the absolute value of all eigenvalues of the matrix of weights is less than 1, and remove interactions randomly until this condition is satisfied. The value $x_i(t)$ of each variable at time t is calculated using a linear regression, as in equation (3).

To simulate measurements under different experimental conditions, we applied two strategies: Changing the standard deviation σ_i of the noise in the regression in order to reflect the fact that the measurement noise may vary across time, and using a modified network that introduces a small number of changes with respect to the originally generated network (adding and removing interactions), reflecting the assumption that not all pathways are active all of the time. Fig. 3 shows an example network with four nodes and the modified networks that have been generated from it.

For our experiments, we generated two kinds of datasets: One with long time series to test how well our methods can reconstruct the network when a lot of data is available, and one which replicates the Arabidopsis data, to see how the performance changes when there are fewer observations available. For the long time series, we generated networks with 10 genes and time series with 50 time steps. We generated 10 different datasets under two conditions: In the first case we did not modify the original network structure, while in the second case, we introduced an average of two changes for the modified networks. We changed the noise level for each network; we used $\sigma = 1$ as the starting value and added a value in the range $[-0.5, 0.5]$, drawn from a uniform random distribution. For the Arabidopsis-like time series, we generated networks with 9 genes and time series with 13 time steps. We generated 4 different datasets under the same two conditions as before, and used the same procedure to vary the noise levels. Table 2 describes which datasets were used for the network reconstruction simulation study in Section 3.1.

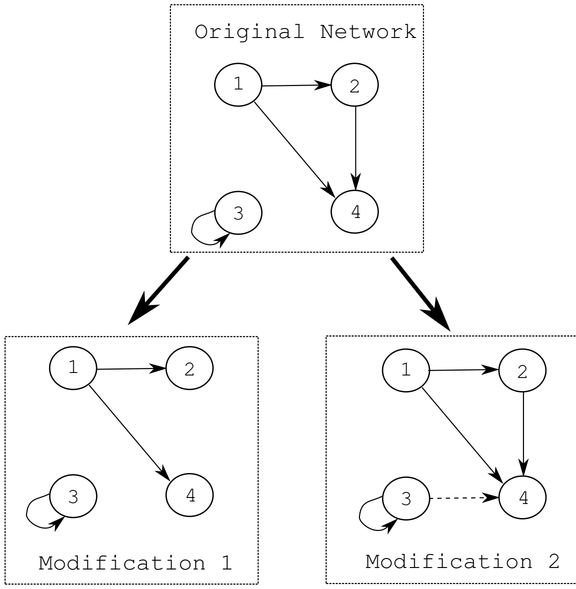


Fig. 3: Simulation model process: The original network is modified to obtain two new networks, each with a different change with respect to the original network.

2.3 Arabidopsis Data

Plants assimilate carbon via photosynthesis during the day, but have a negative carbon balance at night. They buffer these daily alternations in their carbon budget by storing some of the assimilated carbon as starch in their leaves in the light, and utilising it as a carbon supply during the night. In order to synchronize these processes with the external 24 hour photo period, plants possess a circadian clock that can potentially provide predictive, temporal regulation of metabolic processes over the day/night cycle. The proper working of this circadian regulation is paramount to biomass production and growth, and considerable research efforts are therefore underway to elucidate its underlying molecular mechanism. In the present article, we aim to reconstruct the regulatory network of nine circadian genes in the model plant *Arabidopsis thaliana*.

Our analysis is based on four independent gene expression profiling experiments described in Mockler et al (2007), Edwards et al (2006) and Grzegorczyk et al (2008). In these studies, wild-type Col-0 seedlings of *Arabidopsis thaliana* were grown for 7 days under artificially controlled light-dark cycles. On the 8th day the seedlings were placed in constant light. ¿From these seedlings, RNA was extracted and assayed on Affymetrix GeneChip oligonucleotide arrays at regular time intervals. The data were background-corrected and normalised according to standard procedures, using the GeneSpring software (Agilent Technologies). The experiments were carried out at different laboratories and under different pre-experiment entrainment conditions and for different time intervals of measurements. An overview is provided in Table 1. Table 2 describes how the datasets were used for the network reconstruction study in Section 3.2.

	Mockler et al.(2007)	Edwards et al. (2006)	Grzegorcyk et al. (2008) Data 1	Grzegorcyk et al. (2008) Data 2
Time points	12	13	13	13
Time point interval	4h	4h	2h	2h
Pretreatment entrainment	12h-light 12h-dark cycle	12h-light 12h-dark cycle	10h-light 10h-dark cycle	14h-light 14h-dark cycle
Measurement conditions	Constant light	Constant light	Constant light	Constant light
Laboratory	Kay Lab	Millar Lab	Millar Lab	Millar Lab

Table 1: Overview of the gene expression profiling experiments for *Arabidopsis thaliana*. Measurements were started after 7 days of growth of the seedlings and were repeated every 2 or 4 hours, depending on the dataset, for up to two days. Pretreatment entrainment specifies the light conditions before measurements were taken.

Figure	Type	Datasets	Genes	Samples	Notes
4(left)	Simulated	10	10	50	Same Network Structure
4(right)	Simulated	4	9	13	Same Network Structure
5(left)	Simulated	10	10	50	10% Network Structure Changes
5(right)	Simulated	4	9	13	10% Network Structure Changes
6	Simulated	4	9	13	0-20% Network Structure Changes
7	Arabidopsis	4	9	12-13	Reconstruction without information sharing
8	Arabidopsis	4	9	12-13	Reconstruction with information sharing
9	Arabidopsis	4	9	13	Comparison of confidence scores of interactions for the two Grzegorcyk et al. (2008) datasets with and without information sharing
10	Arabidopsis	4	9	12-13	Comparison of agreement of reconstructed networks with and without information sharing

Table 2: Overview of datasets used for the reconstruction of simulated and real gene regulatory networks in Section 3. For each figure we give the type of the data (simulated or real), the number of datasets (time series) that were used for the network reconstruction, the number of genes and samples (measurements) in each time series, and any other details of interest. Note that for the reconstructions of Arabidopsis gene networks, we used all four datasets from Table 1 unless otherwise specified.

3 Results

3.1 Recovering Simulated Networks

Setup We used the simulation model presented in Section 2.2 to generate time series from an underlying network under two different conditions:

1. All time series are generated using the structure of the underlying network, but varying the interaction weights and noise level.
2. Each time series is generated from a different network where we introduce a small number of changes (10%) with respect to the structure of the underlying network. We also vary the interaction weights and noise level.

The second condition should provide a more difficult inference problem than the first one, since there is less scope for information sharing. For both cases, we generate ten independent datasets, each with a different underlying network, to allow us to carry out paired t-tests for significance.

Results We used three measures to evaluate the performance of our methods: The area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate versus the false positive rate, the area under the precision-recall (PR) curve, which plots precision (fraction of true positives out of detected interactions) versus recall (fraction of true positives out of actual interactions; another name for the true positive rate) and the true positive rate at a false positive rate of 5% (TPFP5). We obtain the curves for the first two scores by varying a threshold on the marginal posterior probability of the interactions, and by only keeping those interactions that lie above the threshold at each point. The ROC curve will always be increasing from (0,0) to (1,1), while the precision-recall curve does not have to follow such a clear trend (although precision will generally decrease as recall increases). Taking the area under the curve allows us to reduce the curve to one number that indicates the overall performance[2]. A perfect score for all three methods is a score of 1, which means that we always retrieve all of the true positives, and don't retrieve any false positives at the highest threshold.

These measures are interesting for different reasons: The ROC curve describes the overall performance of the network reconstruction method over positives and negatives, while the precision-recall curve is of practical interest because it does not include the true negatives, and hence focuses on how well the true edges are reconstructed. The TPFP5 score gives the fraction of true edges that we could expect to retrieve at a reasonable fraction of false positives.

Looking at the comparison in Figs. 4-5, it is clear that the information sharing model outperforms the DBN approach without information sharing. In every case, there is a significant improvement in the score when we apply information sharing. The improvement is most drastic when the underlying network structure is unchanged (Fig. 4). Applying small changes to the network structure for each new simulated time-series (Fig. 5) leads to a smaller, but still significant improvement.

The relative performance between the model with and without information sharing is similar whether we use long time series (left column in Figs. 4-5) or time series that

[2] In order to calculate the area, we need to interpolate to find additional points of the curve. For the ROC curve, this is a straightforward linear interpolation, while for the precision-recall curve, we follow Davis and Goadrich (2006).

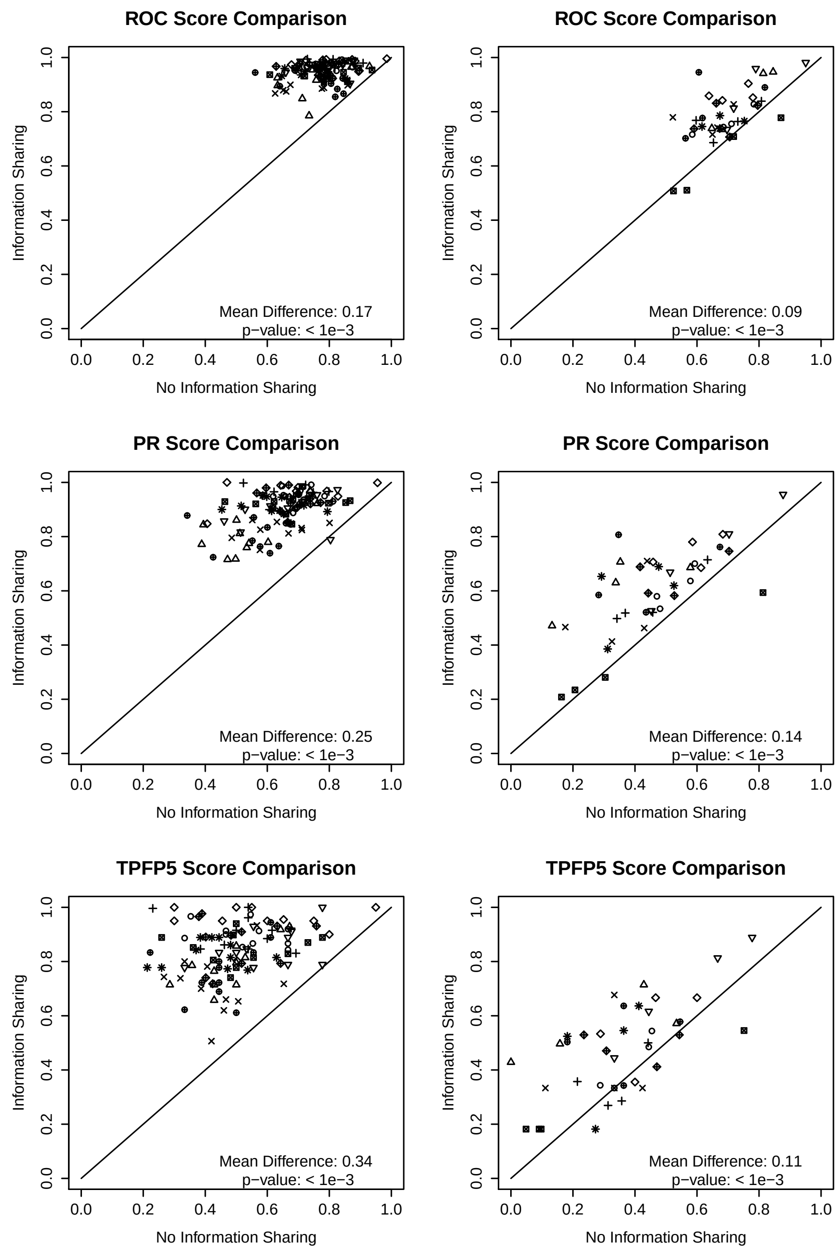


Fig. 4: **Same Structures**: Comparison of network reconstruction performance using the DBN model with and without information sharing. Left Column: For each underlying network, we generated 10 time-series of length 50 without changing the network structures. Right Column: For each underlying network, we generated 4 time-series of length 13 without changing the network structures. Top Row: Area under the ROC curve score. Middle Row: Area under the precision-recall curve. Bottom Row: True positive rate at 5% false positives. In each case, a score of 1 denotes perfect reconstruction of the network. Points with the same symbol are the scores of networks reconstructed from different time series associated with a single underlying network.

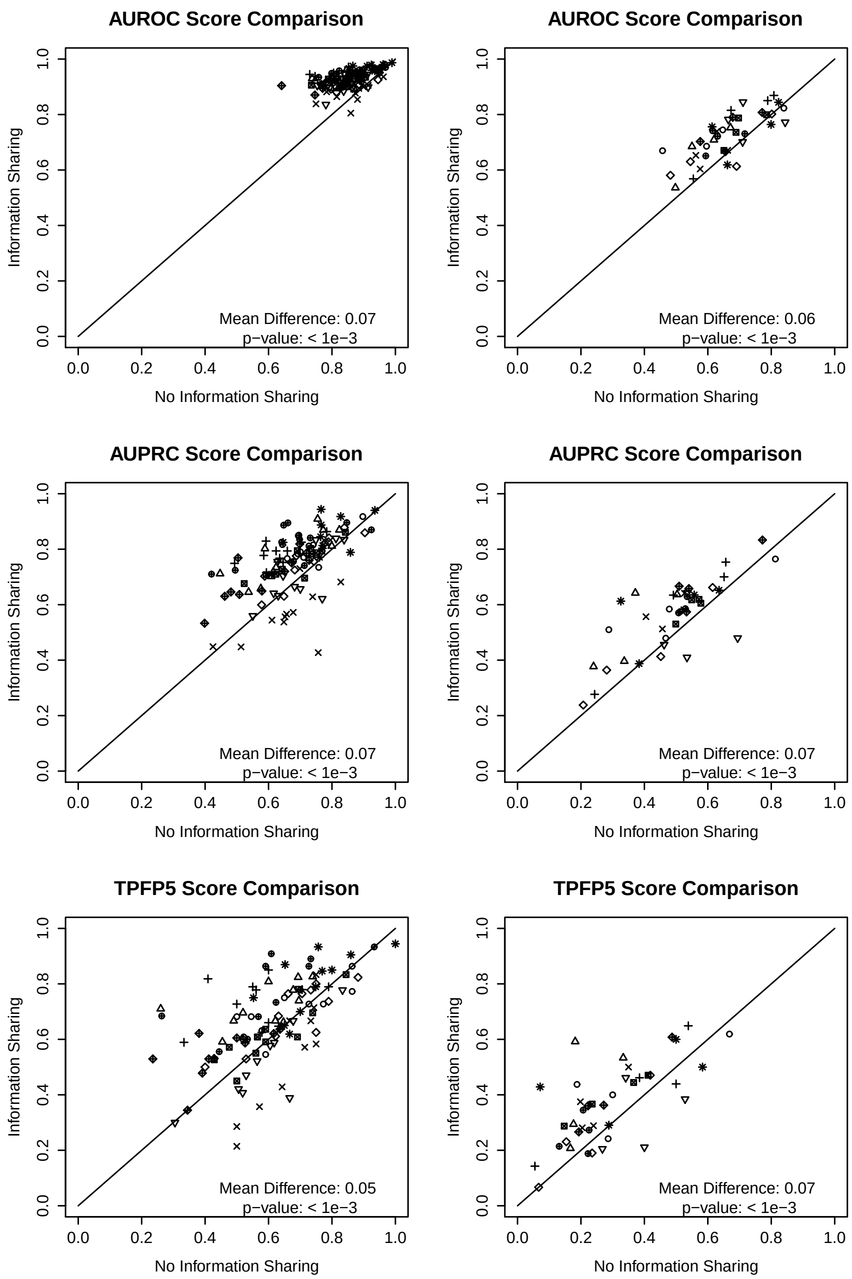


Fig. 5: **Different Structures**: Comparison of network reconstruction performance using the DBN model with and without information sharing. Left Column: For each underlying network, we generated 10 time-series of length 50, changing about 10% of the network structure each time. Right Column: For each underlying network, we generated 4 time-series of length 13, changing about 10% of the network structure each time. Top Row: Area under the ROC curve score. Middle Row: Area under the precision-recall curve. Bottom Row: True positive rate at 5% false positives. In each case, a score of 1 denotes perfect reconstruction of the network. Points with the same symbol are the scores of networks reconstructed from different time series associated with a single underlying network.

have the same length as the Arabidopsis data (right column). The improvement is larger for longer time series, however, most likely due to there being more datasets available that can benefit from the information sharing (10 rather than 4). In absolute terms, the performance increases when the time series are longer, which is reasonable because it means that more data is available. Nevertheless, the performance with simulated time series of the same length as the Arabidopsis data is still reasonable, and there is a definite increase in the accuracy of the reconstructed networks when using information sharing. This is an encouraging finding, which motivates the application of our method to the Arabidopsis gene expression time series. Note that for the latter, an objective evaluation is not feasible owing to the lack of a proper gold standard.

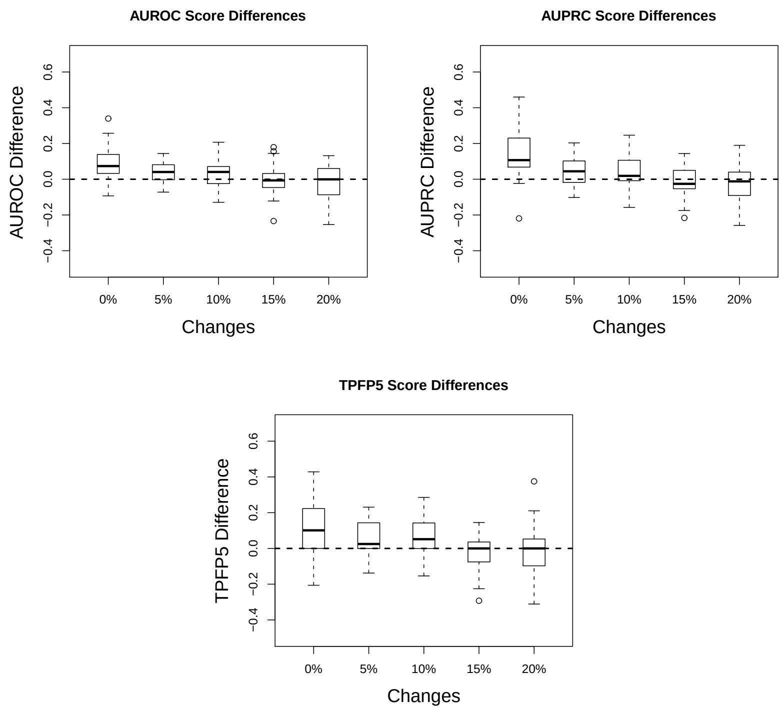


Fig. 6: **Influence of the Number of Changes**: We vary the number of changes applied to each network from 0% to 20%. Network reconstruction performance is measured using the area under the ROC curve (AUROC) score, Area under the precision-recall curve (AUPRC) score, and the true positive rate at 5% false positives (TPFP5).The boxplots show the difference of the network reconstruction scores with information sharing to those without; larger differences indicate better performance of information sharing, 0 means they perform equally well. The horizontal bar of each boxplot shows the median, the box margins show the 25th and 75th percentiles, the whiskers indicate data within 2 times the interquartile range, and circles are outliers.

We notice that overall the PR scores are less impressive than the ROC scores. This is a consequence of the sparseness in the model; we have more non-interactions than interactions in the simulated network, and our DBN model favours fewer interactions, which means that we are more likely to detect true negatives correctly. This improves the false positive rate, but has no effect on the precision, meaning that the PR curve is not going to reflect this. This makes the PR curve (and the TPFP5 score) a better measure if we are more interested in the retrieved interactions than in the non-interactions. The trend is the same, however, in that the improvement when using information sharing is smaller (but still significant) when we apply small changes to the underlying network before simulating the data.

Further simulations, where we increased the noise levels to $\sigma = 2$ and doubled the number of changes in the network structure, showed that the benefit obtained through information sharing is robust to noise, but does not persist when the number of changes becomes too large, as could be expected. It is reasonable to ask how much of a topology disturbance we can have while still getting a significant improvement with information sharing. Fig. 6 plots the difference in network reconstruction scores between no information sharing and information sharing as the number of changes varies between 0% of the network and 20%. Note that the sparseness of gene regulatory networks means that 20% of the network represents a sizeable portion of the gene interactions. For example, if the original network has 10 genes, then 20% represents 20 interactions that change, so on average each gene will change two of its regulators. The crossover point where information sharing no longer gives a significant improvement seems to be around 15% of the network changing.

3.2 Arabidopsis

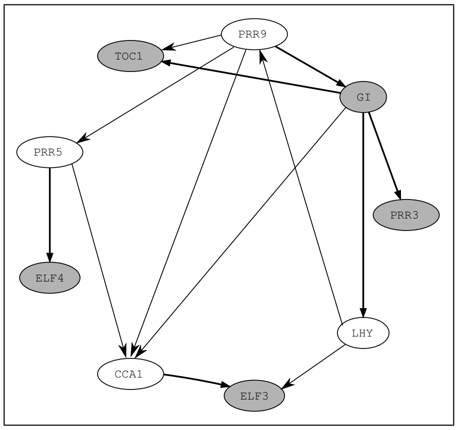
We applied DBNs to the Arabidopsis data described in Section 2.3. Fig. 7 shows the results when we did not use the information sharing approach, so that each network was only inferred from one dataset. To determine which interactions were relevant, we put a threshold on the marginal posterior probability of the interaction (the fraction of this interaction being present in the sampled networks).

We can observe a couple of effects of neglecting to use information sharing. First of all, the connectivity of the inferred networks varied widely between the four datasets. In fact, the network reconstructed from the Mockler et al. dataset had so few interactions with high posterior probability that we had to lower the threshold to obtain a network with a similar number of interactions compared to the other networks. Another effect is that some genes, such as *LHY*, vary from being regulated by just one or two genes (Fig. 7a-7c) to being regulated by no less than 5 genes (Fig. 7d).

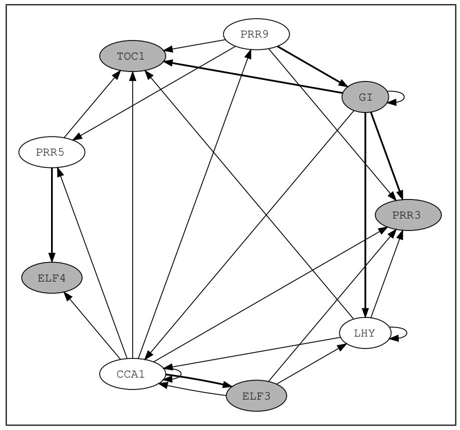
Fig. 8 shows the networks obtained when using the information sharing approach. The first thing to note is that the information sharing has a regularising effect on the network density, which allowed us to apply the same threshold to all four inferred networks. Overall, the sampled networks have fewer interactions, due to the penalising effect of the information sharing prior described in Section 2.1. This made it much easier to find an appropriate threshold on the posterior probability of the interactions.

Compared to the networks in Fig. 7, we notice that there is less variation in the connectivity, although the first network is still sparser. Also, the variation in the number of regulators for *LHY* is no longer as drastic as in Fig. 7.

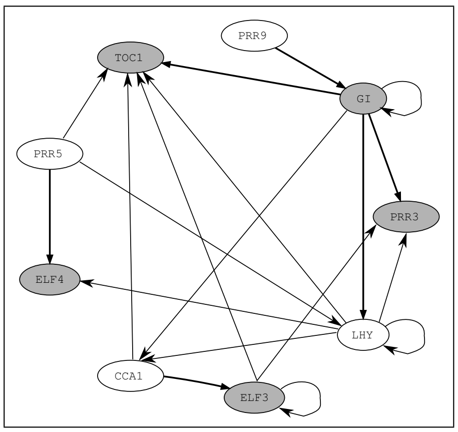
These networks reveal several gene interactions that can be found in the literature. For instance McClung (2006) shows *CCA1* and *LHY*, two genes that are active in the morning, as central regulators of genes that are active in the evening, such as *PRR9*, *TOC1* and *ELF3*. We recover these interactions in most (though not all) datasets. In



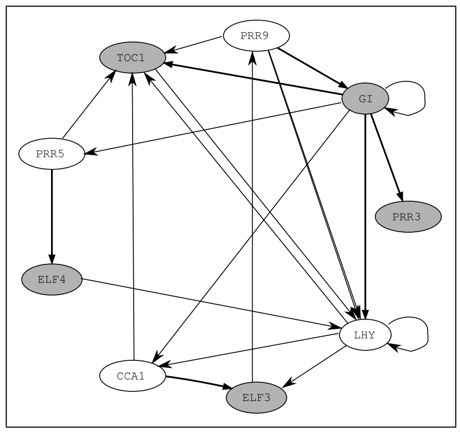
(a) Mockler et al. Dataset Network



(b) Edwards et al. Dataset Network



(c) Grzegorczyk et al. Dataset 1 Network

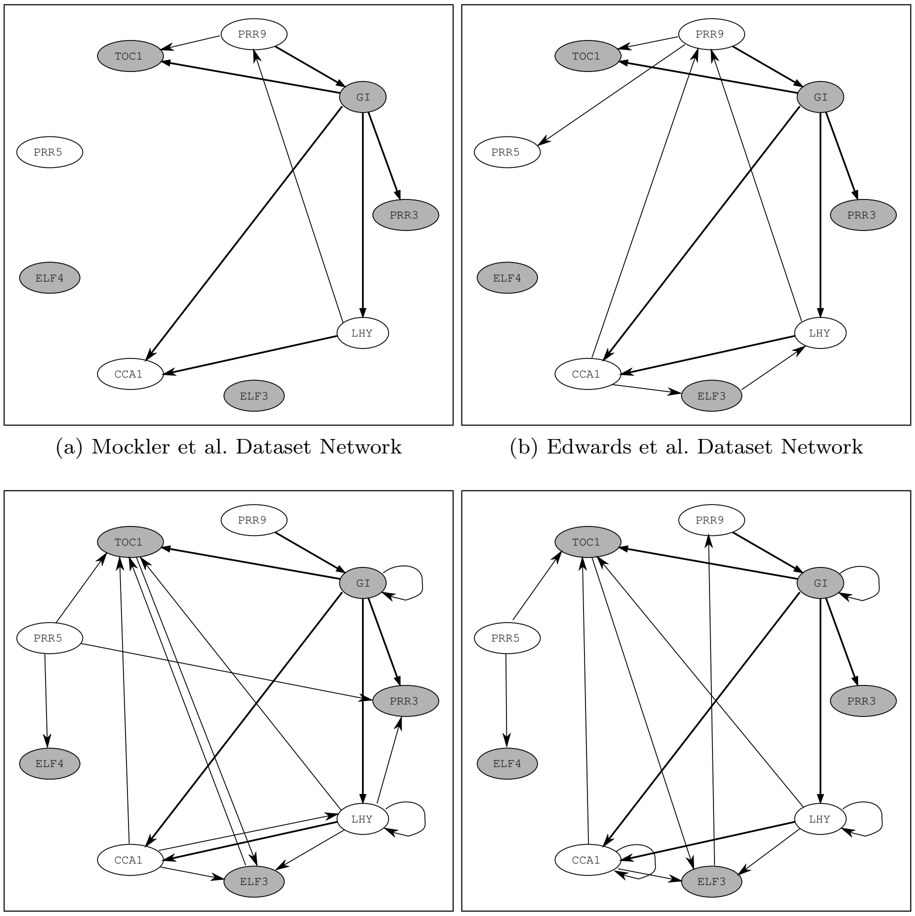


(d) Grzegorczyk et al. Dataset 2 Network

Fig. 7: Networks reconstructed from the four datasets, without information sharing (though with common hyperparameters and initialisation). Only interactions that were present in more than 35% of the sampled networks have been selected (except for the Mockler et al. dataset, where the sampled networks were sparse, and we lowered the threshold to 25% of the sampled networks). Interactions that were found in all datasets are marked in bold.

addition, it seems that *LHY* regulates *CCA1*; this interaction was discovered consistently in all datasets using our information sharing method.

Conversely, some of the evening genes are known or suspected to activate the morning genes. We discovered consistent interactions which identified *GI* (an evening gene) as a regulator of *CCA1* and *LHY*. In addition, we also found that *GI* regulates *TOC1*, an interaction which seems likely given results in Locke et al (2005). One interesting interaction that we found consistently was the regulation of *GI* by *PRR9*; McClung (2006) depicts *PRR9* as regulating *CCA1* and *LHY* directly, while our model seems to favour an indirect regulation via *GI*. Using the information sharing model helps us to identify these interactions more consistently, as comparing Fig. 7 and Fig. 8 shows:



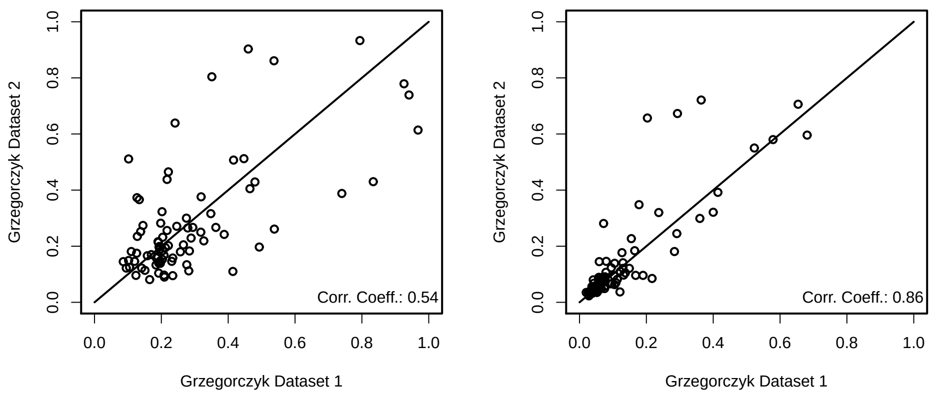
(a) Mockler et al. Dataset Network (b) Edwards et al. Dataset Network

(c) Grzegorczyk et al. Dataset 1 Network (d) Grzegorczyk et al. Dataset 2 Network

Fig. 8: Networks reconstructed from the four datasets, using the information sharing method described in Section 2.1. In all networks, only interactions that were present in more than 15% of the sampled networks have been selected. Interactions that were found in all datasets are marked in bold. The lower threshold compared to Fig. 7 can be explained by considering the information sharing as a penalisation factor. Even very strong edges will be penalised if they only occur in one or two of the four segments. This makes the procedure more selective and allows us to point out a restricted subset of interactions, i.e. interactions which are strong enough to be found in the data after penalisation.

Although one can find all of the interactions listed above in at least one of the networks in Fig. 7, they are found much more consistently across networks in Fig. 8.

We can also investigate the effect of information sharing more directly, by comparing whether the similarity of the marginal posterior probabilities of the gene interactions inferred from different datasets increases when we introduce information sharing. Fig. 9 shows scatterplots comparing the posterior probabilities obtained from Grzegorczyk et al. Dataset 1 and Grzegorczyk et al. Dataset 2. Originally, the posterior



(a) No Information Sharing

(b) Information Sharing

Fig. 9: Comparison of the marginal posterior probabilities of the gene interactions inferred from Grzegorczyk et al. Dataset 1 and Grzegorczyk et al. Dataset 2. (a) Without information sharing, (b) with information sharing. Correlation coefficients were calculated using Spearman rank correlation.

Table 3: Spearman rank correlations between the posterior probabilities for the gene interactions that were inferred for each dataset.

Dataset	Mockler	Edwards	Grzegorczyk 1	Grzegorczyk 2
Mockler	1	0.42	0.39	0.39
Edwards	0.42	1	0.33	0.40
Grzegorczyk 1	0.39	0.33	1	0.54
Grzegorczyk 2	0.39	0.40	0.54	1

(a) No Information Sharing

Dataset	Mockler	Edwards	Grzegorczyk 1	Grzegorczyk 2
Mockler	1	0.79	0.72	0.69
Edwards	0.79	1	0.74	0.73
Grzegorczyk 1	0.72	0.74	1	0.86
Grzegorczyk 2	0.69	0.73	0.86	1

(b) Information Sharing

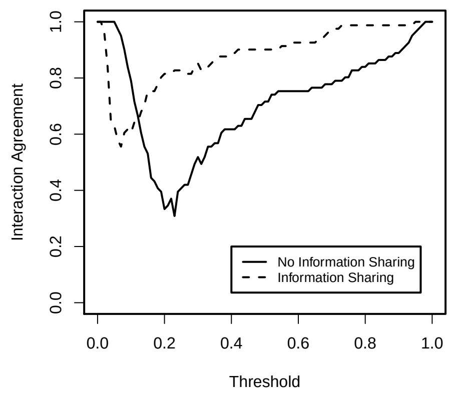


Fig. 10: Agreement between the networks inferred from the four datasets, plotted as the fraction of coinciding interactions (including coinciding non-interactions) in all four networks as the threshold on the posterior probability of the edges increases from 0 to 1. The solid line shows the agreement without information sharing, and the dotted line shows agreement with information sharing.

probabilities are quite scattered, with a Spearman rank correlation[3] of only 0.54. Using information sharing, the rank correlation increases to 0.86. For comparisons between other pairs of datasets, the increase in rank correlation was even bigger, as Table 3 shows.

Fig. 10 shows how the fraction of interactions and non-interactions that coincide in all four inferred networks changes as we increase the threshold on the posterior probabilities of the interactions from 0 to 1. For very low thresholds, all possible interactions will be included, so that the networks all coincide, while for high thresholds, no interactions will be included, again resulting in perfect agreement. The interesting part of the plot is the middle area, where we see that the agreement with information sharing increases much faster than the agreement without information sharing.

4 Conclusions

To understand the processes of growth and biomass production in plants, we aim to elucidate the structure of the underlying regulatory networks at the molecular level. In the present paper, we have proposed and assessed a machine learning method based on Bayesian networks, information sharing and Bayesian inference with Markov chain Monte Carlo to reconstruct gene regulatory networks from gene expression time series obtained under different experimental conditions.

We have discussed the concept of dynamic Bayesian networks for inferring gene regulatory interactions from gene expression time series. Unlike deterministic methods,

[3] Spearman rank correlation measures whether the order is similar, with values closer to 1 indicating probabilities that would produce the same ranking.

DBNs allow us to quantify the inherent uncertainty due to measurement noise and alternative pathways, by calculating the posterior probability of the interactions. One frequent problem in Systems Biology is the integration of data from different sources. Simply pooling the data can have undesired effects, such as producing spurious interactions if one of the datasets is contaminated by noise. To solve this problem, we have applied an information sharing method that can learn a gene network from each dataset, but takes advantage of information from the other datasets by encouraging interactions to be similar.

Using synthetic data, we have demonstrated that information sharing leads to a significant increase in the accuracy of the reconstructed gene networks. The increase was more pronounced when the underlying networks were exactly the same for each dataset, but even networks with a small number of modifications ($< 15\%$) were reconstructed more accurately under information sharing. This shows that our model is robust and can infer commonalities between networks more accurately than a method without information sharing.

Our application of the method to data from nine circadian clock genes in Arabidopsis has shown that it is possible to reconstruct known gene interactions. We discovered interactions between morning and evening genes that were reported in McClung (2006), as well as an interaction between *GI* and *TOC1* that seems likely given the results presented in Locke et al (2005). We also showed that the networks inferred using information sharing had more commonalities than those inferred without it, and were generally more interpretable. Finding gene interaction networks in this manner can help us find new interactions that can then be verified experimentally. For example, we consistently discovered an interaction between *PRR9* and *GI* that was not reported in McClung (2006).

We note that while our approach infers separate networks for each dataset, we are making the implicit assumption that the network structure never changes within one dataset, i.e. over the course of a time series. This assumption of homogeneity of the networks over time is not always appropriate; were we to consider longer time series which encompass several phases in the development of Arabidopsis, or which were measured in a changing environment, then it would be reasonable to assume that different parts of the gene regulatory network would be active at different times. If the timing of these events is unknown, then one would have to switch to a model that can infer when significant changes happen (e.g. Lèbre et al, 2010; Husmeier et al, 2010; Robinson and Hartemink, 2010) to reconstruct the changing networks.

In our specific scenario, all time series only covered up to two days within one stage in the development of *Arabidopsis thaliana* (immediately after the appearance of the first leaves) at constant light conditions, so the time-homogeneity assumption is likely to be a good approximation. In general, the approach presented in this paper is appropriate for data that can be split into separate datasets with time series measurements of gene expression where the time-homogeneity assumption holds, but heterogeneity is introduced through different experimental conditions or different treatments for each dataset. Figure 6 gives an indication of how much divergence from a common network structure is allowed for information sharing to give a significant improvement in network reconstruction performance.

Our method could be further extended, for example by implementing an information sharing scheme based on a latent network that represents the commonalities between the inferred network, as was done in Werhli and Husmeier (2008). One could also consider using prior biological information about gene interactions in the model (such as in Werhli and Husmeier (2007)), which may help identify new, previously unknown interactions.

There is great potential for probabilistic models, such as dynamic Bayesian networks, to help biological research by identifying plausible gene interactions and thus providing a tool for hypothesis generation. The interplay between modelling and experiments is vital; experimental data will inform the modelling, which in turn can prompt follow-up experiments. Future research is likely to make heavy use of network reconstruction and information sharing methods such as the ones we have described in this paper.

Acknowledgements

This work was supported by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD). Frank Dondelinger's PhD research is partly funded by the Engineering and Physical Sciences Research Council (EPSRC).

References

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant and cell physiology 48(3):381

Butte A, Kohane I (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, p 418

Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning, ACM, Pittsburgh, Pennsylvania, pp 233–240

Edwards K, Anderson P, Hall A, Salathia N, Locke J, Lynn J, Straume M, Smith J, Millar A (2006) FLOWERING LOCUS C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock. The Plant Cell Online 18(3):639

Ferrazzi F, Rinaldi S, Parikh A, Shaulsky G, Zupan B, Bellazzi R (2008) Population models to learn Bayesian networks from multiple gene expression experiments

Friedman N, Murphy K, Russell S (1998) Learning the structure of dynamic probabilistic networks. In: Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI98), Citeseer, pp 139–147

Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. Journal of computational biology 7(3-4):601–620

Grzegorczyk M, Husmeier D, Edwards K, Ghazal P, Millar A (2008) Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. Bioinformatics 24(18):2071

Hamada K, Hongo K, Suwabe K, Shimizu A, Nagayama T, Abe R, Kikuchi S, Yamamoto N, Fujii T, Yokoyama K, et al (2011) OryzaExpress: An integrated database of gene expression networks and omics annotations in rice. Plant and Cell Physiology 52(2):220

Husmeier D, Dondelinger F, Lebre S (2010) Inter-time segment information sharing for non-homogeneous dynamic Bayesian networks. Advances in Neural Information Processing Systems 23:901–909

Jiao Y, Tausta S, Gandotra N, Sun N, Liu T, Clay N, Ceserani T, Chen M, Ma L, Holford M, et al (2009) A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. Nature genetics 41(2):258–263

Lèbre S, Becq J, Devaux F, Lelandais G, Stumpf M (2010) Statistical inference of the time-varying structure of gene-regulation networks. Submitted

Locke JCW, Southern MM, Kozma-Bognr L, Hibberd V, Brown PE, Turner MS, Millar AJ (2005) Extension of a genetic network model by iterative experimentation and mathematical analysis. Molecular Systems Biology 1(1):E1–E9, DOI 10.1038/msb4100018

Ma S, Gong Q, Bohnert H (2007) An Arabidopsis gene network based on the graphical Gaussian model. Genome research 17(11):1614

MacKay DJC (1998) Introduction to Monte Carlo methods. In: Jordan MI (ed) Learning in Graphical Models, Kluwer Academic Publishers, The Netherlands, pp 301–354

Madigan D, York J (1995) Bayesian graphical models for discrete data. Int Stat Rev 63:215–232

McClung CR (2006) Plant circadian rhythms. Plant Cell 18(4):792–803

Mochida K, Uehara-Yamaguchi Y, Yoshida T, Sakurai T, Shinozaki K (2011) Global landscape of a co-expressed gene network in barley and its application to gene discovery in Triticeae crops. Plant and Cell Physiology

Mockler T, Michael T, Priest H, Shen R, Sullivan C, Givan S, McEntee C, Kay S, Chory J (2007) The diurnal project: Diurnal and circadian expression profiling, model-based pattern matching and promoter analysis. Cold Spring Harbor Symposia on Quantitative Biology 72:353–363

Moriyama M, Hoshida Y, Otsuka M, Nishimura S, Kato N, Goto T, Taniguchi H, Shiratori Y, Seki N, Omata M (2003) Relevance Network between Chemosensitivity and Transcriptome in Human Hepatoma Cells1. Molecular Cancer Therapeutics 2(2):199

Morohashi K, Grotewold E (2009) A systems approach reveals regulatory circuitry for Arabidopsis trichome initiation by the GL3 and GL1 selectors. PLoS genetics 5(2):e1000,396

Murphy K, Mian S (1999) Modelling gene expression data using dynamic Bayesian networks

Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H (2006) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic acids research 35(suppl 1):D863

Okazaki Y, Shimojima M, Sawada Y, Toyooka K, Narisawa T, Mochida K, Tanaka H, Matsuda F, Hirai A, Hirai M, et al (2009) A chloroplastic UDP-glucose pyrophosphorylase from Arabidopsis is the committed enzyme for the first step of sulfolipid biosynthesis. The Plant Cell Online 21(3):892

Robinson J, Hartemink A (2010) Learning non-stationary dynamic Bayesian networks. The Journal of Machine Learning Research 11:3647–3680

Rogers S, Girolami M (2005) A Bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics 21(14):3131–3137

Schäfer J, Strimmer K (2005) An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 21(6):754–764 van Someren EP, Vaes BLT, Steegenga WT, Sijbers AM, Dechering KJ, Reinders MJT (2006) Least absolute regression network analysis of the murine osterblast differentiation network. Bioinformatics 22(4):477–484

Sreenivasulu N, Usadel B, Winter A, Radchuk V, Scholz U, Stein N, Weschke W, Strickert M, Close T, Stitt M, et al (2008) Barley grain maturation and germination: metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. Plant Physiology 146(4):1738

Werhli AV, Husmeier D (2007) Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. Statistical Applications in Genetics and Molecular Biology 6(1), DOI 10.2202/1544-6115.1282

Werhli AV, Husmeier D (2008) Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions. Journal of Bioinformatics and Computational Biology 6(3):543–572