Using Higher-Order Dynamic Bayesian Networks to Model Periodic Data from the Circadian Clock of *Arabidopsis Thaliana*

Rónán Daly[1], Kieron D. Edwards[2], John S. O'Neill[3], Stuart Aitken[4],
Andrew J. Millar[4], and Mark Girolami[1]

[1] Inference Group, Department of Computing Science, University of Glasgow
`{rdaly,girolami}@dcs.gla.ac.uk`
[2] Advanced Technologies (Cambridge) Limited
`kieron.edwards@atcbiotech.com`
[3] Institute of Metabolic Science, Metabolic Research Laboratories,
University of Cambridge
`jso22@medschl.cam.ac.uk`
[4] Centre for Systems Biology at Edinburgh, The University of Edinburgh
`{s.aitken,andrew.millar}@ed.ac.uk`

Abstract. Modelling gene regulatory networks in organisms is an important task that has recently become possible due to large scale assays using technologies such as microarrays. In this paper, the circadian clock of *Arabidopsis thaliana* is modelled by fitting dynamic Bayesian networks to luminescence data gathered from experiments. This work differs from previous modelling attempts by using higher-order dynamic Bayesian networks to explicitly model the time lag between the various genes being expressed. In order to achieve this goal, new techniques in preprocessing the data and in evaluating a learned model are proposed. It is shown that it is possible, to some extent, to model these time delays using a higher-order dynamic Bayesian network.

Keywords: Dynamic Bayesian Network, Gene Regulatory Network, Gene Expression, Arabidopsis Thaliana.

1 Introduction

The analysis of data obtained from experiments looking at the expression of genes has become an important topic in the realm of bioinformatics. Whilst other bioinformatics tasks such as sequence alignment, gene finding, genome assembly and genome annotation have proven very amenable to computational analysis, there are tasks which are intrinsically difficult at the current level of knowledge. These include protein structure prediction and modelling gene regulatory networks. The latter will be examined in this paper.

It is well known that genes code for proteins and that certain proteins can affect how certain genes are expressed. Gene regulatory networks encode these dependencies between genes in an organism, abstracting away the influence of

proteins so that nodes correspond to genes and directed arcs correspond to causal gene-expression relations. In this paper, an attempt will be made to model these gene regulatory networks and in particular, the time lag between the expression of a gene and its effect on the expression of another gene. This modelling will proceed by using an algorithm based on ant colony optimisation known as ACO-E [1] to build dynamic Bayesian networks (DBNs). DBNs are an extension of Bayesian networks, graphical models that can be useful in modelling probabilistic relations between variables [2]. They consist of a prior network, used to model initial conditions, and a transition network, used to model the effect of interactions over time. Whilst DBNs have been used to model gene regulatory networks [3,4,5,6,7], the method presented here differs in how the data are treated and in how it can infer interactions at multiple time periods.

To test the methods proposed, a set of gene expression data obtained from experiments observing particular genes of the plant *Arabidopsis thaliana* will be used. These genes are known to behave in a clock-like fashion that regulates the function of the plant [8,9]. The data will be preprocessed and experiments will be conducted to learn the structure of DBNs using these preprocessed data. Using a new technique, these networks will be compared to a standard network developed by an expert in the field being analysed. Using these results, the ability of DBNs to reconstruct gene-regulatory networks will be investigated.

2   Background

The analysis of gene regulatory networks is a hard task. In order to see how various genes interact, experiments must be performed which measure the expression levels of these genes. Using this data, it is possible to develop models of genetic systems behaviour. One of the earliest and still commonly used techniques is to use differential equations that show how expression levels vary with respect to other expression levels.

However, recent advances in measurement of gene expression levels, using techniques such as DNA microarrays, have allowed the measurements of tens of thousands of genes to be performed simultaneously. With this amount of genes in an organism, all possible interactions between all the genes cannot be looked at. Therefore, methods involving graphical representations have become more widely used. These include Boolean networks, networks similar to neural networks, stochastic process calculi and Bayesian networks [10].

2.1   Data from *Arabidopsis Thaliana* Experiments

The experimental study in this chapter will involve luminescence data obtained from experiments on *Arabidopsis thaliana*. The luminescence is an indirect measure of the synthesis of new RNA transcripts from each gene, originating from a copy of the cognate promoter and $5'$ untranslated region that are fused to the firefly luciferase reporter gene and integrated into the plant genome at a random location. The particular experiments in question were designed to study

Table 1. Ratio of light to darkness for each experimental condition

Condition	AT0029	AT0030	AT0031b	AT0032	AT0033	AT0047
Ratio	6:18	9:15	12:12	15:9	18:6	3:21

the 'plant circadian rhythms' of this organism, i.e. the oscillating behaviour of plants as they respond to the change in sunlight [11]. Briefly, the expression levels of the genes in plants tend to synchronise with the rising and setting of the sun, oscillating between high and low levels. However, when the light source is removed, the expression levels continue to oscillate with the same frequency, with the behaviour decaying over time.

In the experiments, there were in general two phases. The first phase, called the entrainment, switched the light on and off at regular intervals and lasted three days. The second phase had constant light and also lasted three days. At intervals of 1.5 hours, readings were taken of the expression levels of the genes being investigated.

The experiments involved ten different genes: CAB, GI, CCA1, CCR2, LHY, CAT3, ELF3, PRR9, TOC1 and ELF4. Further information about these genes can be found elsewhere [11]. There were also six different conditions under which the experiments were conducted; each condition had a different light period for the entrainment phase of the experiment. Table 1 shows, for each experimental condition, the ratio of light to dark for that condition. E.g. for condition AT0029, there are 6 hours of light followed by 18 hours of darkness in the entrainment phase. Because the experiments lasted six days, and a sample was taken every 1.5 hours, there were a total of 96 samples taken for each condition. The first of these was taken as a reference datum and subtracted from each of the following samples, leading to 95 samples with which to work.

3   Modelling Gene Expression Data Using DBNs

It is the stochastic nature of the gene expression process that makes probabilistic models an appropriate choice in modelling them. One interesting approach is in using Bayesian networks as a probabilistic model. This enables large scale, non-linear interactions between many genes to be represented and simulated. E.g., recent studies have looked at learning Bayesian networks with thousands of genes [12]. The large amounts of data present in microarray studies can be used to learn both the structure and parameters of the network – missing values and noise can also be taken care of. Also, Bayesian networks can be given a causal interpretation. A problem arises however in that loops are not permitted in Bayesian networks. Therefore, feedback processes cannot be represented.

A solution to this problem is to use dynamic Bayesian networks (DBNs). With DBNs, it is possible to 'unroll' what would be a loop in a Bayesian network across
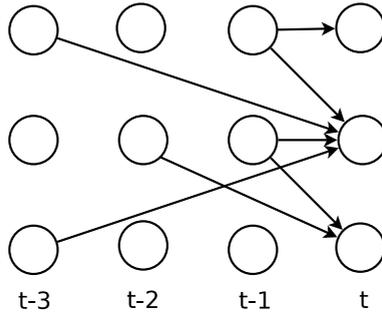
Fig. 1. A 4-layer dynamic Bayesian network

time. In doing so, a temporal order is brought in amongst the various variables. In this respect, DBNs are a natural way to model temporal data.

3.1   Learning a Higher-Order Dynamic Bayesian Network

Most Bayesian network structure learning algorithms are designed to learn a static structure. Whilst dynamic Bayesian networks are very similar to static networks, there are slight differences that mean extra care has to be taken when trying to learn them. Bearing in mind that a DBN has two parts, the prior and transition network, this section will focus on learning the transition network; the prior network is exactly the same as a normal Bayesian network.

In a transition network, the nodes are grouped into multiple layers that designate different timepoints. Arcs can never go back in time and the head of any arc can only be in a single layer, normally that layer at time t. It is very normal to have two layers, such that they correspond to variables at time t and $t-1$. In fact, in all the general literature on Bayesian networks surveyed, learning DBNs amounted to learning DBNs with two layers. However, in general, $n+1$ layers can be specified, from $t-n$ to t. An example of a 4-layer DBN is shown in Fig. 1.

When learning such a DBN, it is sufficient to add an extra constraint to the learning algorithm, i.e. the head of an arc must always be in the layer at time t. By following this, any Bayesian network structure learning algorithm that searches through the space of DAGs can be used and will produce a valid network.

4   Experimental Methodology

In order to test the ability of multi-layered DBNs to model a genetic regulatory network, experiments were conducted using the *Arabidopsis thaliana* data discussed in Sect. 2.1. This section will describe the steps taken to achieve this. This includes preprocessing the data, formulating prior knowledge, designing a methodology to run the algorithm and testing the results obtained against expert knowledge.

4.1   Preprocessing the Data

The data obtained from the experiments described in Sect. 2.1 are continuous in nature, i.e. they show the level of gene expression as measured. However, most scoring functions for use with structure learning are based on nominal data, i.e. data that is of a discrete nature. For the purposes of these experiments, the BDeu scoring function was used [13]. For this to be utilised, the data has to be discretised.

Whilst binary discretisation is an easy option, there is a problem with this as shown by Fig. 2a. It can be seen that as the expression level decays after entrainment, the discretised expression level flatlines, even though there is still oscillatory action occurring. One way to counter this is to discretise the data into multiple levels. However, in doing so, the amount of data at each level is reduced. This can lead to less support for dependencies when a model is fitted to the data, something that should be avoided when there is not much data to begin with.

In order to avoid this problem the first derivative of the data was taken and the data discretised as to whether this derivative was greater than or less than zero. The results of this procedure are shown in Fig. 2b. It can be seen that this discretisation captures the oscillations present in the original data. In a sense, instead of looking for correlations in the mRNA level, we are looking for correlations in the rising and falling of the mRNA level. To the authors' knowledge, this is the first time this has been done when learning Bayesian networks that model gene regulatory networks.

Finally due to the expert being able to provide knowledge on the behaviour of five of the genes and the light source, these were selected as the variables that would be used in the construction of the DBN. The genes in question were GI, CCA1, LHY, PRR9 and TOC1. Whilst the other genes could have been included in the analysis, any connections among them would be unverifiable by expert knowledge in this study.

(a) Expression level   (b) Expression level derivative
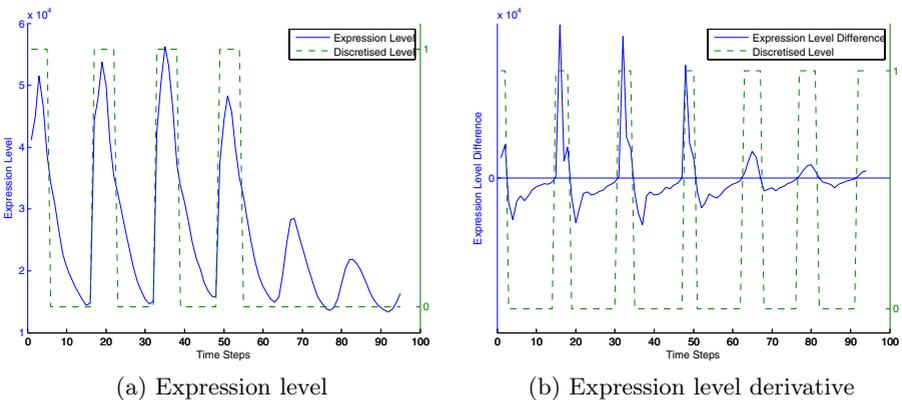
Fig. 2. Expression levels for PRR9 in AT0029 condition

4.2   Prior Knowledge

In order to provide more meaningful results and to compensate for the absence of much data, prior knowledge was incorporated into the learning algorithm in the form of constraints as to the allowable edges. These constraints are summarised below:

- Connections from a variable to the same variable at a later time point were disallowed. This type of information is trivial and does not add anything interesting to the model.
- No connection was allowed *to* the light variable, as this was causally independent of the gene expression values.
- Since interactions at this genetic level took time in the region of at least one time step, no connections were allowed among the variables at time t.
- A dynamic constraint was introduced that specified that if a connection was made from a variable X at time $t-i$ to a variable Y at time t, then no other connection could be made from X to Y. This is equivalent to saying that variables only affect each other temporally in a single manner, i.e. from X to Y only ever takes a set amount of time. Without a constraint such as this, in the absence of much data, multiple interactions between the same genes at different time points could easily be inferred.
- The number of slices of a possible dynamic Bayesian network was set to 9, i.e. the time slice at t and 8 slices back. Other authors show using DBNs with multiple layers to model gene regulatory networks with multiple time lags [14]. However, their methods consider arcs between arbitrary layers that are not at time t. This method is incorrect, as it means that arcs could be added between nodes without taking into account the other parents these nodes have. E.g. a local search could add an arc between X_{t-2} and X_{t-1} and between Y_{t-1} and X_t. This implies that X_t has parents X_{t-1} and Y_{t-1}, but both arcs were added without considering the other.

4.3   Testing Methodology

In order to test the ability of DBNs to recreate the expert's knowledge, a series of experiments were conducted. Each experiment used the data described in the preprocessing section, with the prior knowledge as described above. Also, the BDeu scoring criterion was used with an uniform structure prior. The parameters of the structure learning algorithm (ACO-E) were set as $\rho = 0.3$, $q_0 = 0.8$ and $\beta = 1$. These were selected as reasonable values that should perform well in most cases [1].

It has been known for some time that the value of the equivalent sample size parameter N' for the BDeu scoring function has a large effect on learning structure [15,16,17]. In a sense, the value of N' can be seen as a regularisation parameter - the larger the value of N', the more edges are supported in the learned graph [18]. This can be contrasted to scores like the BIC and AIC, where regularisation is implicit in the function and cannot be adjusted. Therefore, in

learning a Bayesian network structure from data, it is important to see how the N' parameter will affect the learned graph. Because of this, for each condition described in Table 1, experiments were performed that varied the equivalent sample size N' over 15 values ranging from 0.0001 to 500. These values were 0.0001, 0.001, 0.01, 0.1, 1, 3, 7, 10, 20, 30, 50, 80, 100, 300, 500. At each level of N', 10 experiments were performed. Also, the data from all the conditions was concatenated and experiments performed, varying N' over 20 values ranging from 0.01 to 8000. These values were 0.01, 0.1, 1, 3, 7, 10, 20, 30, 50, 80, 100, 300, 500, 750, 1000, 2000, 4000, 6000, 7500, 8000. Again, at each level of N', 10 experiments were performed.

For each experiment, the resulting DBN was saved for later comparison against the expert supplied knowledge.

4.4   Evaluation Criteria

Evaluating the learned DBNs was achieved by comparing them against the expert supplied knowledge. For this task, the knowledge was transformed into a DBN representing the state of the expert's knowledge on the domain. The DBN was validated by the expert and is shown in Fig. 3. Note that although the DBN looks different to published networks in the literature, it captures the same semantics. Please also note that nodes are not shown if they have no connection to any other node. Also, although the arcs in the network are definite, the knowledge of the domain expert was not as definite. Often the time points supplied were of the order of six hours long, i.e. instead of a definite time lag being given, the lag was bounded by two times roughly six hours apart. The effect of this would be to reduce the measured accuracy of any learned network.

To compare a learned network against the standard network, two measures were used: the true positive rate (TPR) and the false positive rate (FPR). These are defined as

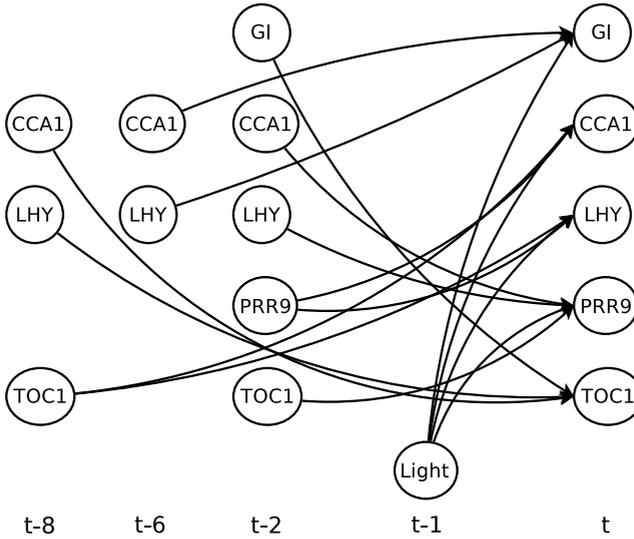$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

and

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

where

TP is the number of true positives, i.e. the number of arcs computed present, that are present. For the evaluation of the networks, this was calculated in two different ways:

Time-Independent True Positive. If there was an arc from gene X to gene Y in the standard network and also an arc from gene X to gene Y in the testing network, *irrespective of the time slice that X is in,* then this counts as a true positive.

Time-Dependent True Positive. If there was an arc from gene X to gene Y in the standard network and also an arc from gene X to gene Y in the testing network, then this counts as a *time-dependent* true positive. The

Fig. 3. Expert developed DBN for *Arabidopsis thaliana* circadian clock

degree to which it is a true positive depends on the difference of time slices that the tail of each arc comes from. A difference of 0 gives a true positive of 1 and a difference of 8 gives a true positive of 0. Other values are calculated linearly between these two values. E.g. with an arc in the standard network having the tail in slice 4 and an arc in the testing network having a tail in slice 2, this leads to a difference of 2, which gives a time dependent true positive of $1 - 2/8 = 6/8$.

FP is the number of false positives, i.e. the number of arcs computed present that are absent. If there exists an arc from gene X to gene Y in the testing network and there does not exist an arc in the standard network, then this counts as a false positive.

TN is the number of true negatives, i.e. the number of arcs computed absent that are absent. If there does not exist an arc from gene X to gene Y in the testing network and there does not exist an arc in the standard network, then this counts as a true negative.

FN is the number of false negatives, i.e. the number of arcs computed absent that are present. If there does not exist an arc from gene X to gene Y in the testing network and there does exist an arc in the standard network, then this counts as a false negative.

In previous work of a similar nature, evaluating the learned network proceeds in a different manner than above [14]. Some problems in their methods include:

– Firstly, as mentioned in Sect. 4.2, the method they use to add arcs can lead to incorrect conclusions about the structure of the dynamic network;

– Secondly, they don't take into account the time delay when evaluating the
learned network; and
– Thirdly, they don't take any false positive rate into account.

5 Results and Discussion

For each condition and each level of N', the mean of the TPR and that of
the FPR were taken over the 10 experiments. This was done for the Time-
Independent TP (TITP) and the Time-Dependent TP (TDTP). With these
results, two different types of graphs were plotted. The first was the receiver
operating characteristic (ROC) curve, which plots the FPR against the TPR.
The second plots the FPR and TPR as a function of N'. The results of these plots
for the concatenated data are shown in Fig. 4. Due to reasons of space it was
not possible to show the graphs for the individual conditions. To compensate,
the area under the curve (AUC) statistic is given in Table 2.

5.1 Discussion of Experimental Results

The value of the equivalent sample size N' has a large effect on learning structure
[15,16,17]. Therefore, in learning a Bayesian network structure from data, it is
important to see how this parameter will affect the learned graph. This is the
reason why the experiments described above were conducted over various values
of N'.

Looking first at Figs. 4a and 4b, it can be seen that the learned DBN structures
do a good job of identifying all of the connections between the genes as supplied
by the domain expert. In this case, the best value of N' is at 1, where the true
positive rate is 1 and the false positive rate at just over 0.3. With the time
dependent TPRs as seen in Figs. 4c and 4d, again the best value of N' is at 1,
with a time dependent TPR of around 0.8 and a false positive rate of 0.3. In this
case, 'best' is being taken as the absolute difference between the TPR and FPR.

These results show that the learned models are good at showing all the con-
nections between the genes, but not as proficient at keeping out bad connections.

Table 2. Area under the ROC curve for both the time-independent and time-dependent
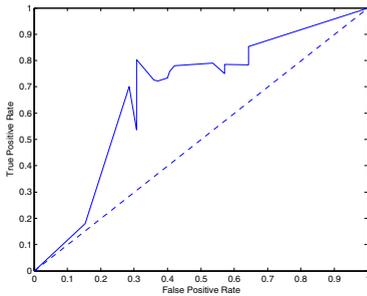true positive cases

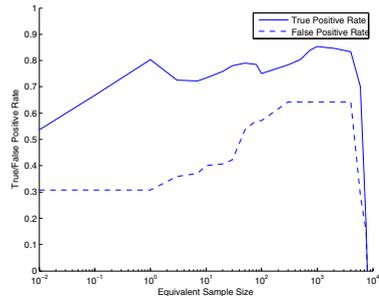| Condition	Time-Independent AUC	Time-Dependent AUC	Light Ratio
All	0.7913	0.6749	—
AT0029	0.6943	0.5749	6:18
AT0030	0.6597	0.5942	9:15
AT0031b	0.8234	0.7046	12:12
AT0032	0.7135	0.5696	15:9
AT0033	0.8156	0.7155	18:6
AT0047	0.5200	0.4326	3:21
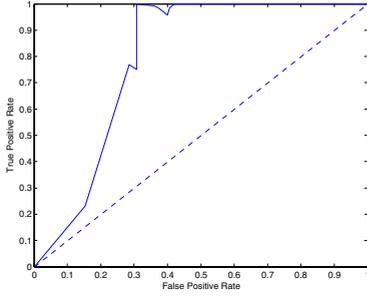
(a) ROC for the time-independent TPR and FPR
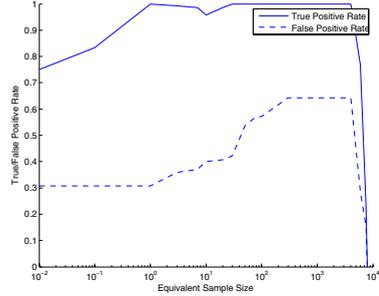
(b) The time-independent TPR and FPR as a function of N'

(c) ROC for the time dependent TPR and FPR

(d) The time dependent TPR and FPR as a function of N'

Fig. 4. Results for concatenated data

Indeed, when examining a trace of the algorithm it can be seen that spurious connections between highly synchronised genes are often inserted. E.g. if X causes Y and X causes Z, then a connection between Y and Z can easily appear as well as the connections from X to Y and Z. Problems such as these often appear with a small amount of data and prior knowledge becomes increasingly important in these situations.

With the results of the individual conditions, different behaviours can be observed. The value of N' for which best results are obtained differs widely depending on the condition used, ranging from 1 to 100. The different conditions also differ in how good the results are. E.g. the best performance came from the results with the larger light to dark ratio, i.e. AT0031b, AT0032 and AT0033. The worst performance came from the other conditions, i.e. AT0029, AT0030 and AT0047. Results such as these are plausible, as having more light in the entrainment phase equates to higher expression levels, which are less likely to be affected by noise.

It should be noted that not all the expert supplied knowledge is as accurate as may seem from the standard network. Whilst the time in that network was

given in time steps of 1.5 hours, the knowledge of the expert was often given in terms of much larger grained steps (e.g. 'the morning'), as opposed to a definite time period. This is an artifact of the domain in question, as it is not completely understood at the present time. In turn, it has an effect on the time dependent true positive values obtained from comparing two networks; with better prior knowledge a more accurate comparison could be used that bounds what the correct time lags might be for an arc. It should also be noted that the layers in the network impose a discretisation in time that might not be accurately reflected in the domain.

6   Conclusions

This paper asked the question on whether it is possible to use multi-layered dynamic Bayesian networks to model gene regulatory networks using gene expression data. To help answer this question new techniques were devised. Firstly, the gene expression data was preprocessed by a novel method in relation to learning gene regulatory networks; the first difference was obtained and this difference discretised into two bins – rising or falling. Secondly, because of the lack of much previous work on learning higher-order DBNs, new evaluation criteria were defined to judge how well the learned network reconstructed the expert supplied network. From the results it can be seen that it is possible, to some extent, to model both the connections between various genes and the time lags associated with these connections.

References

1. Daly, R., Shen, Q.: Learning Bayesian network equivalence classes with ant colony optimization. Journal of Artificial Intelligence Research 35, 391–447 (2009)
2. Daly, R., Shen, Q., Aitken, S.: Learning Bayesian networks: Approaches and issues. The Knowledge Engineering Review (in press, 2009)
3. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 19(17), 2271–2282 (2003)
4. Kim, S.Y., Imoto, S., Miyano, S.: Inferring gene networks from time series microarray data using dynamic Bayesian networks. Briefings in Bioinformatics 4(3), 228–235 (2003)

5. Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., d'Alché Buc, F.: Gene networks inference using dynamic Bayesian networks. Bioinformatics 19(suppl. 2), ii138–ii148 (2003)

6. Zou, M., Conzen, S.D.: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21(1), 71–79 (2005)

7. Geier, F., Timmer, J., Fleck, C.: Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. BMC Systems Biology 1, 11 (2007)

8. Edwards, K.D., Anderson, P.E., Hall, A., Salathia, N.S., Locke, J.C., Lynn, J.R., Straume, M., Smith, J.Q., Millar, A.J.: FLOWERING LOCUS C mediates natural variation in the high-temperature response of the *Arabidopsis* circadian clock. The Plant Cell 18, 639–650 (2006)

9. Locke, J.C.W., Kozma-Bognár, L., Gould, P.D., Fehér, B., Kevei, É., Nagy, F., Turner, M.S., Hall, A., Millar, A.J.: Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. Molecular Systems Biology 2(59) (2006)

10. Schlitt, T., Brazma, A.: Current approaches to gene regulatory network modelling. BMC Bioinformatics 8(suppl. 6), S9 (2007)

11. McClung, C.R.: Plant circadian rhythms. The Plant Cell 18, 792–803 (2006)

12. Huang, Z., Lib, J., Su, H., Watts, G.S., Chen, H.: Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining. Decision Support Systems 43(4), 1207–1225 (2007)

13. Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning 20(3), 197–243 (1995)

14. Xing, Z., Wu, D.: Modeling multiple time units delayed gene regulatory network using dynamic Bayesian network. In: Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops (ICDMW 2006), pp. 190–195 (2006)

15. Steck, H., Jaakkola, T.S.: On the Dirichlet prior and Bayesian regularization. In: Becker, S., Thrun, S., Obermayer, K. (eds.) Advances in Neural Information Processing Systems 15 (NIPS 2002), pp. 697–704. MIT Press, Cambridge (2003)

16. Silander, T., Kontkanen, P., Myllymaki, P.: On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In: Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2007 (2007)

17. Kayaalp, M., Cooper, G.F.: A Bayesian network scoring metric that is based on globally uniform parameter priors. In: Darwiche, A., Friedman, N. (eds.) Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI 2002), pp. 251–258. Morgan Kaufmann, San Francisco (2002)

18. Steck, H.: Learning the Bayesian network structure: Dirichlet prior vs data. In: McAllester, D.A., Myllymäki, P. (eds.) Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI 2008), pp. 511–518. AUAI Press (2008)