We analyse two network interactions are assumed linear and an autoregressive model Gaussian specification for dealing with measurement error. The as these measurements often contain outliers, we use a non-
engineering a gene interaction network, based on time course data with repeated measurements. The observations are treated as surrogate measurements of the underlying gene expression.

As these measurements often contain outliers, we use a non-

Identifying and understanding gene regulatory networks is of key

well known that when followed through time, biological processes have feedback loops and thus the validity of BNS is questionable when modelling such systems. Dynamic BNS (DBNs) have been proposed for modelling time course (longitudinal) gene expression data (Cao and Zhao, 2008; Murphy and Mian, 1999; Perrin et al., 2003; Yu et al., 2004; Zou and Conzen, 2005). These can be thought of as ‘unfolding’ a BN for every time point and when folding it back self-regulation and cliques may be obtained.

Formally, a DBN is characterized by a set of conditional relations, \(p(y_{t+1} | y_t)\). In the case of a regression-based DBN, these relations can be written as

\[
y_{tg}^{t+1} = f(y_{tg}^t) + \epsilon_{tg}^{t+1},
\]

where \(y_{tg}^t\) is the expression level of gene \(g = 1, \ldots, G\), measured at time \(t = 1, \ldots, T\), \(y_t = [y_{t1}, y_{t2}, \ldots, y_{tG}]\) and \(\epsilon_{tg}^t\) is an idiosyncratic error term.

The approaches above assume one observed time series for each gene. However, gene expression measurement normally requires destruction of the sample, e.g. microarrays, and, therefore, the idea of a longitudinal time series becomes problematic. This is because a single individual is not followed throughout the experiment, but rather a population of cells or individuals are sampled and their gene expression measured. The phenomenon is particularly acute in experiments with multicellular organisms, where not even the same population of cells can be followed through time. Thus, rather than ‘real’ gene expression measurements, we are faced with a set of surrogate measures. In addition to the uncertainty involved in the sampling process, it is well known that gene expression measurement technologies, such as microarrays, render noisy data and frequently exhibit outliers (Brody et al., 2002; Lewin et al., 2007).

When repeated measurements are available, time course data used for reverse engineering gene interaction networks are commonly obtained as a (weighted) average of these replicates and, therefore, these sources of uncertainty are ignored. Neglecting the variability within the replicates can have severe effects when fitting a linear model, with perhaps the most important being attenuation of the coefficient estimates (see Carroll et al., 2006; Fuller, 1987).

Working within a univariate first-order linear autoregressive setting, Schmid et al. (1994) demonstrated that neglecting measurement error yields severe attenuation, of the autoregressive coefficient and the variability of this estimate. Interpreting this result within the framework of (1) with a linear specification of \(f(y_t)\), this suggests that an averaged time series will yield attenuated estimates of the interactions within the network, with a spurious sense of security given the concomitant underestimation of the variability of these estimates.
In the context of regulatory networks, Fujita et al. (2009) examine the effect of including measurement error in a linear model, both in static and dynamic contexts. They verify that the OLS estimator of the regression coefficients is attenuated and that its corresponding standard error is sub-estimated; then they provide a correction based on the asymptotic behaviour of the OLS. However, they do not assume sparsity in the network and, therefore, require more observations than the number of parameters in the model, making practical applications of their method to biological network inference difficult. Further, when dealing with time series data, they are forced to assume the measurement error precision known, omitting this source of uncertainty, which in turn will affect the estimation of the autoregression coefficients’ precision. The main aim of this article is to show that by neglecting the uncertainty within the measurement process some biases can be passed on to the estimated network and, more importantly, that the confidence in the retrieved network will be spuriously large.

In the sequel, we present a model that takes into account repeated measurements of time course gene expression data for estimating the topology of a gene interaction network. We model the network topology explicitly, enabling us to control for the effective number of parameters to be estimated and, therefore, able to apply the methodology to commonly available time course gene expression datasets, where the number of parameters exceeds the number of observations. To ease the presentation, we assume an underlying first-order linear autoregressive process, AR(1), for the interaction observations. To ease the presentation, we assume an underlying AR(1) model for the interaction process, with our prior specification and estimation measurement process, modelling and estimation is carried out from a Bayesian perspective, with our prior specification and estimation procedure explained in Section 3. Illustrations of model fitting and comparisons with the standard approach are conducted in Section 4.

A final discussion is provided in Section 5.

### 3 METHODS

When combining (2) and (4), and using the Student-t representation as a Gamma scale mixture of normals, one can write the likelihood as

\[ l((\theta), X) = \prod_{g=1}^{G} \prod_{t=1}^{T} N(y_{gt}^{i} | \mu_{gt}, \tau_{gt}^{2}) \times N(x_{gt}^{i} | 0, \lambda_{gt}^{2}) \]

where \( Y = \{y_{gt}^{i}\} \) are the unobserved expression levels, \( X = \{x_{gt}^{i}\} \) denote their surrogate measurements and \( \theta = (\mu, \beta, \Gamma, \tau, \nu) \) collects the model parameters, with \( \mu = \{\mu_{1}, \ldots, \mu_{G}\}, \beta = \{\beta_{g1}, \ldots, \beta_{gG}\} \in \mathbb{R}^{G^{2}} \) and \( \tau_{gt}^{2} = \{\tau_{g1}, \ldots, \tau_{gG}\}; \tau^{2} = \{\tau_{1}, \ldots, \tau_{G}\} \).

The Bayesian model is completed by specifying a prior for all the unknowns. We use a product structure

\[ \pi((\theta), \nu) \sim \prod_{g=1}^{G} \prod_{t=1}^{T} \text{N}(\mu_{gt} | 0, k) \times \text{Ga}(\beta_{gt} | a, b) \times \text{Ga}(\tau_{gt} | 1, 1/\nu) \]

and specify componentwise conditionally conjugate priors where suitable. Thus,

\[ \pi(\mu_{gt}) \sim \text{N}(0, k), \]

\[ \pi(\beta_{gt}) \sim \text{N}(0, b), \]

\[ \pi(\tau_{gt}) \sim \text{Ga}(a, b), \]

\[ \pi(\nu) \sim \text{Ga}(1, 1/\nu) \]

Of paramount importance in our modelling is the inclusion of the link indicator variables, \( \Gamma = \{\gamma_{ij}\} \), and their probabilistic structure, controlled by (11) and (12). By augmenting the model in this way, we are able to switch the regulation of gene \( i \) on or off, controlling for the effective number of parameters to be estimated. Moreover, the posterior mean of the link indicators, \( \hat{\gamma}_{ij} \), are interpreted as the posterior link probabilities, the building blocks of the network topology retrieval. From a formal viewpoint,
Identifiability is always a potential issue when dealing with measurement error models, particularly when neither gold-standard measurements or pure replicate measurements can be obtained (Gustafson, 2004, Section 6.3). This problem is further compounded in gene network determination where the number of potential parameters to be estimated \(p = \dim(\Theta) + (G + 1)\) is typically much greater than the available data points \(n = T \times G\), i.e. the so-called \(p \geq n\) case. From a Bayesian perspective, models are always identifiable as long as a proper prior is specified. However, this formal identifiability may imply not learning from the data (Poirier, 1998) and thus calls for a careful elicitation of the prior (Gustafson, 2005). Indeed this is the case with the model and measurement precision parameters, when repeated measurements are absent \((R = 1)\). We take advantage of the information contained in the measurements \(X\) about \(\tau\) and use a rather flat prior on this parameter, while carefully eliciting the parameters for \(\pi(\cdot)\).

There is no closed analytic expression for the posterior distribution, \(\pi(\theta|X)\), and numerical methods to explore it are needed. To this end, we constructed a Markov chain Monte Carlo (N) within-Gibbs strategy improves mixing and, therefore, faster convergence of the chain, and we follow their suggestion. We use a Metropolis step for \(\theta\) with a Gamma proposal centred at the previous draw and tune its coefficient \(\nu\) to give roughly prior odds of 3 to 2 for the measurement error models. This coefficient can be regulators (and possibly other genes that could affect regulation, such as kinases), encoding such information in the prior through \((11)\). Further, the algorithm is straightforwardly parallelized, as the parameters for each gene can be computed independently, the CPU-nodes needing to communicate only for updating the overall connectivity, \(\rho\) and for collecting the draws, thus reducing the runtime roughly proportionally to the number of available CPU-nodes.

4 RESULTS

We analyse two in silico and one in vivo datasets. The simulated experiments allow us to isolate the effect of explicitly modelling measurement error in controlled situations and highlight the attenuation effect. We then turn to a real experiment dealing with the circadian clock in Arabidopsis thaliana. In all cases, we fit the model with and without the measurement error component, using the same prior structure \(((6)-(13))\), deleting the relevant terms when not accounting for measurement error or when assuming it is Gaussian distributed.

4.1 In silico networks

The first synthetic network is linear, and thus serves as a baseline for comparisons. The second is nonlinear and will allow comparisons within a more realistic, yet still controlled setting. For each synthetic dataset, we generate a rather large number of time points, 41 and 50, respectively, so we can highlight the effect of measurement error in network reconstruction.

4.1.1 Linear interactions

Here, we use \((2)\) and \((4)\) to generate a synthetic data set with \(G = 16\) and \(T = 41\). We set the network connectivity \(\rho \approx 0.13\) and produced a layered network: a hub gene is perturbed by an (unmeasured) external input, the signal is then propagated to a second layer of genes with another hub which, in turn, propagates the signal to a third layer. A small amount of links feeding forward and backward between layers are also included. Expression profiles can be found in Supplementary Figure S1a.

This dataset is regarded as the ‘noiseless’ case. Using it as basis, we generated noisy replicates according to \((4)\) with \(\nu\) either Gaussian (GD) or Student (SD) distributed. For each distribution, we generated two datasets: one with \(R = 4\) and a second with many \((R = 20)\) replicates. In both cases, we consider rather noisy scenarios by setting \(\nu = 4\), 50% of the maximum absolute expression value of each gene (note that in the GD case, \(\nu\) corresponds to the measurement precision). We fixed \(v = 5\), for the Student-\(t\) case.

We fitted three models to each dataset: one where the measurement error is assumed to be Student distributed (SM), the second assumes normal errors (GM) and the third disregards measurement error by taking the mean of the replicates as the true time series (MM). To summarize the results of the inference on the network topology, we use two threshold-independent scores: the area under the ROC curve (AUC) and mean cross entropy (MCE). The AUC provides an overall accuracy measure of network retrieval, using the link predictions sorted according to their magnitude. It thus fails to account for the strength in the predictions; in our case, the estimated link probabilities. These are key as, when performing inference on an unknown network, we will normally set a threshold above which links will be predicted as being present. For this reason, we also calculate the M×E defined as the average Kullback-Leibler...
Table 1. Performance comparison using a synthetic linear network

<table>
<thead>
<tr>
<th></th>
<th>$R = 4$</th>
<th>$R = 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MM</td>
<td>GM</td>
</tr>
<tr>
<td></td>
<td>AUC</td>
<td>AUC</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>SD</td>
</tr>
<tr>
<td></td>
<td>0.78</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>0.68</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>0.82</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>0.59</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>0.24</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0.28</td>
</tr>
</tbody>
</table>

AUC and M+E scores obtained by fitting the model without measurement error (MM) and those with Gaussian (GM) and Student (SM) errors to in silico data with Gaussian (GD) and Student (SD) distributed errors. Bold values are the best scores for each case. The smaller the M+E the better. The larger the AUC the better.

Fig. 1. Marginal posterior distributions of the coefficient, $\beta_{12,3}$ corresponding to link gene12 $\rightarrow$ gene3 of the in silico linear data. The noiseless data (dot-dashed) is the data without noise, the MM (dashed) is the mean of four replicates, neglecting measurement error and SM (solid) uses the four replicates and assumes Student measurement error.

The effect of measurement error can also be highlighted when concentrating on the data with several replicates (columns with $R = 20$ in Table 1), while relative differences in the AUC between MM and SM are reduced to <2% for GD and 11% for SD, relative differences in M+E are 17% and 28%, respectively. This can be understood since attenuation reduces with increased number of replicates and, therefore, the point estimates of the coefficients from either model will be closer to each other, resulting in a similar AUC. However, the variance of these estimates will still be underestimated when using MM and as a result the M+E, which takes into account the actual value of the estimated probabilities, will capture these differences.

Attenuation and underestimation of the variability on the coefficients estimates are illustrated in Figure 1 for a specific link, gene12 $\rightarrow$ gene3. As expected in the ideal, noiseless case, the posterior distribution of the corresponding coefficient, $\beta_{12,3}$, has its mode close to the true value of the coefficient (1.0) and has a rather large precision. With noisy measurements MM renders an attenuated coefficient, with its posterior distribution shifted towards the origin (dashed curve). Moreover, the corresponding posterior has negligible mass near the true value. In contrast, the posterior from SM overlaps nicely (solid). Also, the MM estimate has a posterior precision of about 72, larger than the SM, 62, illustrating the underestimation of the uncertainty in the coefficient estimate. This effect was observed more or less markedly in all predicted links (not shown).

4.1.2 Nonlinear interactions A dataset with nonlinear interactions was generated using a gene network model built with ordinary differential equations (ODEs). The network is a mathematical model of the A.thaliana circadian clock and consists of G=5 genes with eight links, and also includes protein production and transport, as well as daylight (Locke et al., 2006). We generated data from this model using COPASI (Hoops et al., 2006) with the light source fixed permanently on. To mimic realistic sampling regimes, we sub-sampled the data so as to have a time spacing of an hour and then took logs. The resulting time series has a total of $T=50$ time points (plots of the profiles are shown in Supplementary Fig. 1B). This is the ’noiseless’ dataset. We generate noisy replicates using Student distributed errors. As before, we fix $\tau_{4i}$ at 50% of the maximum value of the noiseless gene expression, use $\nu=5$ and produced two datasets: one with 4 replicates and the other with 20 replicates.

Due to the small size and high connectivity of the network ($\rho \approx 0.40$), the scores used in the previous example become quite sensitive. This is because the ROC is a piecewise constant function and with a small number of genes (and therefore links), the size of the steps becomes larger and thus comparisons get more sensitive. For improved interpretability, we plot in Figure 2 the links included in the predicted network against the posterior link probability when fitting the data with $R=4$ replicates. We use a circle (cross) for an incorrect (correct) link; for instance, in the noiseless case the predicted network using a threshold of 0.95 would have nine links (circles and crosses with link probability threshold above 0.95 in Fig. 2a), four out of which (crosses) are correct.

Inspection of the inferred link probabilities using the noiseless data (Fig. 2a), shows that there is some overfitting. For instance,
on these estimates (the tight grouping). In contrast, the posterior link in part to the combined effect of attenuated coefficients (and thus a comparing Figure 2b to Figure 2c and d. Most posterior link concentrated when using MM (see Supplementary Fig. S3). However, the spread of posterior link probabilities is still more available, the three models yield similar network reconstructions; 4 out of 10 and SM 5 out of 11. When a large number of replicates are for MM (one cross and two circles to the right of 0.2), while GM has measurement error: 4 out of 10 for GM and 3 out of 9 with SM. i.e. it has a high false negative rate, compared to the models with present in the network with a threshold of 0.2 (i.e. those crosses we compare model fit for low thresholds, MM predicts 17 links not resulting in either a misprediction or a compensation by the inclusion of spurious parents (see Morrissy et al., submitted for publication, for a detailed discussion).

For the case where we have few replicates, GM and SM outperform MM (Fig. 2b–d). MM predicts only two links with high probability ($\hat{\gamma}_{ij} \geq 0.8$), both incorrect. For a link probability threshold of 0.7, GM performs marginally better than SM (even though the true noise is Student). Using a threshold of 0.5, the inferred network would be the same for both models, though. If we compare model fit for low thresholds, MM predicts 17 links not present in the network with a threshold of 0.2 (i.e. those crosses and circles to the left of 0.2). Seven out of these are real (crosses); i.e. it has a high false negative rate, compared to the models with measurement error: 4 out of 10 for GM and 3 out of 9 with SM. Moreover, the true positive rate for this (low) threshold is 1 out of 3 for MM (one cross and two circles to the right of 0.2), while GM has 4 out of 10 and SM 5 out of 11. When a large number of replicates are available, the three models yield similar network reconstructions; however, the spread of posterior link probabilities is still more concentrated when using MM (see Supplementary Fig. S3).

The effect of attenuation in network retrieval can be seen by comparing Figure 2b to Figure 2c and d. Most posterior link probabilities of MM are tightly clustered toward zero; this is due in part to the combined effect of attenuated coefficients (and thus a lower overall connectivity) and the underestimation of the variability on these estimates (the tight grouping). In contrast, the posterior link probabilities in either GM or SM have a wider spread reflecting both the larger estimates of the coefficients and the increased variability in the estimates when considering measurement error.

4.2 In vivo data
We used a microarray time series of gene expression profiles from Arabidopsis (Denby, K.J. unpublished data). Sampling is destructive, with a different plant used for each time-replicate. To reduce variability, the same leaf was used for each sample. A total of 96 plants were grown under a 16 h:8 h light:dark cycle and the seventh leaf to emerge from each of $R=4$ plants was sampled every 2 h over a 48-h period, i.e. $T=24$.

To select those genes to be included in the analysis, we referred to the current working model of the circadian clock in Arabidopsis (Robertson and Webb, 2009), sketched in Figure 3a. Recently, a new gene (CHE) was identified as a member of the core circadian clock (Prumeda-Paz et al., 2009), so we include this gene to entertain the most up-to-date version of the clock. Two nodes in Figure 3a (LHY/CCA1 and PRR7/PRR9) represent pairs of genes that perform the same role and have very similar expression profiles. To avoid collinearity, a single gene to represent each pair was selected. PRR7 is chosen over PRR0 as it shows a higher signal to noise ratio and CCA1 over LHY given that CHE is predicted to regulate CCA1 and not LHY (traces of the expression profiles are shown in Supplementary Fig. 2S).

Figure 4 depicts the distribution of the inferred link probabilities for each of the three models. As there are few links predicted with high probability, we set a link probability threshold of 0.5. Figure 3 shows the inferred networks for this threshold. GM and MM infer the same network topology, whereas SM infers a network with six links, only two of them in common with the GM/MM network. This large difference is explained by the posterior probability distribution of the degrees of freedom (Fig. 5), where four out of the five genes are inferred to have a fat-tailed measurement error distribution, with a mode of $\nu=3$ for PRR7. The effect of attenuation in the estimation of the link probabilities is illustrated by comparing the three pictures on Figure 4. Again, the majority of probabilities estimated by MM are smaller than those estimated by either GM or SM and are more tightly grouped. On the other hand, GM and SM reflect the additional uncertainty in the measurement process by dragging these probabilities towards the centre of the plot.

The GM/MM network correctly predicts the TOC1 $\rightarrow$ GI $\rightarrow$ TOC1 loop, but also predicts two incorrect links (GI $\rightarrow$ CCA1 and TOC1 $\rightarrow$ PRR7). The GM model also correctly predicts TOC1 $\rightarrow$ GI but incorrectly TOC1 $\rightarrow$ PRR7. The other four links that are absent in the GM/MM predicted network involve either PRR7 (three links) or CCA1 (one) as a regulator. Out of these, there are two correct predictions: PRR7 $\rightarrow$ CCA1, CCA1 $\rightarrow$ TOC1 and two incorrect: PRR7 $\rightarrow$ GI and PRR7 $\rightarrow$ TOC1.

Noteworthy is the discrepancy between the SM and MM/GM inferred networks (Fig. 3). The fact that for the given threshold the MM and GM inferred networks are the same illustrates the effect of outliers in the estimation: when using GM, despite taking into account the uncertainty in the replicates through $t_\nu$, these are treated as interchangeable in the update of $\gamma_{ij}$ (in this case, $\alpha_{ij} = 1$). In contrast, the weights in SM depend on the degrees of freedom and will be more variable for smaller values of $\nu$, allowing for some of the measurements to dominate the average. To verify that this is

![Network topology retrieval for the ODE circadian clock in silico dataset with $R=4$ replicates. Links predicted to be present in the network versus posterior link probabilities estimated for the noiseless case and when using MM, GM and SM. A link (not) present in the ODE model is shown with a (circle) cross.](https://academic.oup.com/bioinformatics/article-abstract/26/18/2305/208626/2305-2312)
Fig. 3. *Arabidopsis* circadian clock. (a) Current working model (redrawn from Robertson and Webb, 2009): dotted lines represent protein production and oval shapes (binding) proteins. (b) and (c) Depict the network topologies inferred with the measurement error models with a threshold of 1/2, using the microarray data. Solid edges represent predicted links that are present in the working model and dashed represent links predicted by either model and not present in the working model. (b) The retrieved topology with Mean and Gaussian models and (c) with the Student model.

Fig. 4. Gene network link prediction for the *Arabidopsis* circadian clock microarray data. Posterior link probabilities from each model are depicted as circles. The vertical dotted line represents a threshold of 1/2. Nevertheless, the posterior distribution of the link probabilities was tighter, illustrating the sub-estimation of the coefficients' precision. These comparisons are dependent on the threshold selected and as such are only a point estimate of the network. As described in the Section 1, the fundamental effect of not accounting for repeated measurements is the overconfidence in the estimation of the coefficients and its effect on the estimates of link probabilities. Even though the topology retrieved from MM and GM, for a fixed threshold, is similar, the distribution of the link probabilities is extremely different. MM predicts three links with very high probability (>0.8) and 14 with very low probability (<0.2), with only three in the central region of Figure 4a. In contrast, GM predicts only 1 link with high probability and 10 with low, leaving 9 possible links in the central zone of Figure 4b. The conclusions obtained from either model can thus be quite distinct, while MM suggests a high confidence in its predictions, GM correctly warns about the high levels of uncertainty in the recovery of the network.

Interestingly, no parents or children are predicted for the new clock gene CHE. Pruneda-Paz et al. (2009) showed that CHE can...
As expected, the ‘no replicates’ scenario (Table 2) shows high variability in the inferred networks. No single link was predicted in more than half the datasets. A third of the predicted networks had no more than one predicted link and a further third of the networks had from four to six predicted links (not shown). Three links appeared more frequently than the rest: the two link loop TOC1→GI→TOC1 and the PRR7→TOC1 link. The two links in the loop were predicted together only in five of the datasets, reflecting the loss in estimation precision when no replicates are available.

Table 2. Circadian clock experimental data. Link prediction counts using the 30 sub-sampled datasets with one replicate and a threshold of 1/2.

<table>
<thead>
<tr>
<th></th>
<th>CCA1</th>
<th>TOC1</th>
<th>PRR7</th>
<th>GI</th>
<th>CHE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA1</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>TOC1</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>PRR7</td>
<td>2</td>
<td>14</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>GI</td>
<td>0</td>
<td>13</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHE</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

As there is only one replicate per gene, the model without measurement error is used. Genes in columns are regulators and rows are regulatees.

Table 3 illustrates the benefits of including repeated measurements: there is a clear separation in the link prediction frequency, with a few links being predicted quite frequently and the rest barely appearing. Those links predicted in more than half of the datasets (highlighted in Table 3) are consistent with those predicted using the full dataset. Links predicted by MM are exactly the same as those with the full data (see Fig. 3). Compared to SM, GM incorrectly swaps TOC1→GI for PRR7→GI while SM misses two links (TOC1→GI and PRR7→TOC1), but is still able to correctly predict two further links (CCA1→TOC1 and PRR7→CCA1). This shows that even in the case when less data is available, SM is still able to infer the degrees of freedom and outperforms MM/GM. In a scenario where few replicates and time points are available, it may be advisable to modify SM making the degrees of freedom common to all genes.

5 DISCUSSION

We demonstrate here that the uncertainty conveyed in repeated measurements of time course gene expression data can have a strong effect when estimating a gene interaction network. In the case of a linear autoregressive network specification, not accounting for this uncertainty leads to attenuation of the autoregressive coefficients and over- or underestimation of the precision of these estimates. This in turn may affect the network topology retrieval. To address this issue, we propose a model that explicitly includes this variability.

Our modelling is fully Bayesian, with the true gene expression unobserved and thus inferred. Inference of these expression values draws information from both the surrogate measurements and the linear AR(1) process assumed for the gene network interaction, with the influence of each source weighted by the relative value of the AR(1) precision, \( \lambda \), and the measurement precision, \( \tau \). When \( \tau \) is relatively small, inference on the expression values will be predominately determined by the AR(1) part of the model. In the case where there is little information in the data about the regulatory process, the prior on \( \lambda \) must there be carefully elicited. We provide a benchmark prior for the kinds of datasets arising from microarrays. Our model accommodates simultaneous inference of the network topology along with the interaction coefficients. We showed that attenuation of the network coefficients as well as the underestimation of the variability of these estimates is systematic. Such behaviour is then passed on to the estimated link probabilities, yielding a more concentrated distribution of link probabilities towards either one or zero. The retrieved networks are obtained by setting an (arbitrary) threshold on the posterior link probabilities, and thus

Table 3. Link prediction counts using the Arabidopsis circadian clock microarray sub-sampled datasets with three replicates.

<table>
<thead>
<tr>
<th></th>
<th>MM</th>
<th>GM</th>
<th>SM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TOC1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>PRR7</td>
<td>1</td>
<td>29</td>
<td>20</td>
</tr>
<tr>
<td>GI</td>
<td>1</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>CHE</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Bold values are those links that were predicted present in more than one half of the 30 subsets, using a threshold of 1/2. Genes in columns are regulators and rows are regulatees.
are only point estimates of the network topology. Thus, the final effect on the inferred network topology is case dependent. This is because, while searching for regulatory dependencies, the estimation process implicitly compares alternative parenthood configurations and the inclusion/exclusion of a link depends on the specific dynamics of each gene and its relative variability. However, the distribution of these probabilities can be quite different, providing the experimentalists with a more accurate description of the uncertainty contained in the model fitting, and thus will be better informed when designing further experiments.

High-throughput technologies yield noisy measurements, with the noise distribution typically exhibiting heavier than Gaussian tails. Not accounting for this behaviour can also have a negative impact when performing inference on the interaction coefficients of the network. In our examples, we showed that the inferred topology with the synthetic datasets improved when using SM over GM and MM. Even though there is not a definite network as yet for the in vivo dataset, the inferred topology showed no difference between MM and GM for the selected threshold, while the posterior distributions of the degrees of freedom indicate heavy tails for all but one of the genes, indicating significant outliers in the data and thus suggesting the MM/GM predictions are questionable.

**Funding**: Warwick Systems Biology Doctoral Training Centre (to E.R.M.); BBSRC grant BB/F003498/1 (to M.A.J.); Experimental data was provided by KJ Denby through the PRESTA Project, grant number BB/F005806/1.

**Conflict of Interest**: none declared.

**REFERENCES**


