BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 18 2010, pages 2305–2312
doi:10.1093/bioinformatics/btq421

Systems biology

Advance Access publication July 16, 2010

On reverse engineering of gene interaction networks using time course data with repeated measurements

E. R. Morrissey[1,†], M. A. Juárez[1,*,†], K. J. Denby[1,2] and N. J. Burroughs[1]

[1]Warwick Systems Biology Centre and [2] Warwick HRI, University of Warwick, Coventry CV4 7AL, UK

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Gene expression measurements are the most common data source for reverse engineering gene interaction networks. When dealing with destructive sampling in time course experiments, it is common to average any available measurements for each time point and to treat this as the actual time series data for fitting the network, neglecting the variability contained in the repeated measurements. Proceeding in such a way can affect the retrieved network topology.

Results: We propose a fully Bayesian method for reverse 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-Gaussian specification for dealing with measurement error. The network interactions are assumed linear and an autoregressive model is specified, augmented with indicator variables that allow inference on the topology of the network. We analyse two *in silico* and one *in vivo* experiments, the latter dealing with the circadian clock in *Arabidopsis thaliana*. A systematic attenuation of the estimated regulation strengths and a concomitant overestimation of their precision is demonstrated when measurement error is disregarded. Thus, a clear improvement in the inferred topology for the synthetic datasets is demonstrated when this is included. Also, the influence of outliers in the retrieved network is demonstrated when using the *in vivo* data.

Availability: Matlab code and data used in the article are available from http://go.warwick.ac.uk/majuarez/home/materials.

Contact: m.a.juarez@warwick.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 19, 2010; revised on June 23, 2010; accepted on July 14, 2010

1 INTRODUCTION

Identifying and understanding gene regulatory networks is of key importance in Systems Biology. Reverse engineering such networks is thus paramount and a plethora of literature dealing with the problem has developed in recent years (see Bansal *et al.*, 2007; Hache *et al.*, 2009 and references therein). Bayesian networks (BNs) have been used previously in gene network determination (Friedman *et al.*, 2000; Friedman, 2004; Hongqiang *et al.*, 2005). However, it is 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, $p(y^{t+1} \mid \mathbf{y}^t)$. In the case of a regression-based DBN, these relations can be written as

$$y_g^{t+1} = f_g(\mathbf{y}^t) + \varepsilon_g^{t+1}, \quad (1)$$

where y_g^t is the expression level of gene $g = 1, \ldots, G$, measured at time $t = 1, \ldots, T$, $\mathbf{y}^t = \{y_1^t, y_2^t, \ldots, y_G^t\}$ and ε_g^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 $f_i(\mathbf{y}^t)$, 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.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

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 network. As gene expression measurements frequently exhibit heavier-than-normal tail behaviour, we model the measurement process with Student-t errors. To this end, we present our approach in Section 2. To account for all sources of uncertainty in the model and 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.

2 APPROACH

Denote by y_g^t the expression level of gene $g=1,\ldots,G$, measured at time $t=1,\ldots,T$. We model the interaction network as a linear AR(1) process,

$$y_g^{t+1}=\mu_g+\sum_{j=1}^{G} y_j^t\tilde{\beta}_{jg}+\varepsilon_g^t, \tag{2}$$

where μ_g is the basal expression level of gene g; $\tilde{\beta}_{jg}=\gamma_{jg}\beta_{jg}$ measures the influence of gene j on gene g, with $\beta_{jg}\in\mathbb{R}$ and $\gamma_{jg}=1$, if j regulates g and $\gamma_{jg}=0$ otherwise. Finally, ε_g^t is an idiosyncratic error term, centred at zero and with precision parameter λ_g, typically assumed to be Gaussian. We augment the model with the parenthood (link) indicator variables $\Gamma=\{\gamma_{jg}\}$, which will be the basis for estimating the network topology.

Assume now that instead of measuring y_g^t directly, we are presented with R surrogate measurements $X=\{x_{gr}^t\}$, $r=1,\ldots,R$. As mentioned before, the common approach is to calculate

$$\bar{x}_g^t=\left[\sum_{r=1}^{R}\omega_{gr}\right]^{-1}\sum_{r=1}^{R}\omega_{gr}x_{gr}^t, \tag{3}$$

with $\omega_{gr}\geq 0$; and then replace y_g^t in (2) with $\bar{x}_g^t$. Influence of possible outliers will depend on the choice of $\boldsymbol{\omega}_g=\{\omega_{g1},\ldots,\omega_{gR}\}$. The pervasive choice in the literature is to set $\omega_{gr}=1$; but when outliers are suspected, sometimes a more robust alternative, such as the median, can be used. In the process of fitting (2), the probability of any given link being present depends on the relative magnitude of its associated coefficient, and hence failure of acknowledging measurement error can yield an attenuated network connectivity.

In Section 3, we provide a fully Bayesian approach to accommodating measurement error when estimating the topology of a gene interaction network within a linear AR(1) specification.

3 METHODS

To account for the additional uncertainty when repeated measurements are available, we assume that the regulation process can be captured by (2), but instead of actually observing y_g^t, we have noisy measurements, x_{gr}^t, such that

$$x_{gr}^t=y_g^t+\eta_{gr}^t, \qquad r=1,\ldots,R, \tag{4}$$

with η_g^t a zero mean measurement error term, with precision parameter τ_g, independent for all g,t,r. This error term is frequently assumed Gaussian; however, given that the measurement process can potentially produce outliers, we will use a Student-t specification, $\mathrm{St}(\eta_g^t\mid 0,\tau_g,\nu)$, such that $\mathrm{Var}[\eta_g^t]=\nu\tau_g^{-1}/(\nu-2)$ provided the degrees of freedom, $\nu>2$.

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

$$\begin{aligned}\ell(\Theta;X)=\prod_{g=1}^{G}\prod_{t=1}^{T}\prod_{r=1}^{R}&N(y_g^{t+1}\mid \mu_g+\boldsymbol{y}^t\tilde{\boldsymbol{\beta}}_g,\lambda_g)\\ &\times N(x_{gr}^t\mid y_g^t,\omega_{gr}^t\tau_g)\mathrm{Ga}(\omega_{gr}^t\mid \nu/2,\nu/2).\end{aligned} \tag{5}$$

where $Y=\{y_g^t\}$ are the unobserved expression levels, $X=\{x_{gr}^t\}$ denote their surrogate measurements and $\Theta=\{\boldsymbol{\mu},B,\Gamma,\boldsymbol{\lambda},\boldsymbol{\tau},\nu\}$ collects the model parameters, with $\boldsymbol{\mu}=\{\mu_1,\ldots,\mu_G\}$; $B=\{\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_G'\}\in\mathbb{R}^{G\times G}$ and $\boldsymbol{\beta}_g=\{\beta_{1g},\ldots,\beta_{Gg}\}$; $\Gamma=\{\gamma_{ij}\}$; $\boldsymbol{\lambda}=\{\lambda_1,\ldots,\lambda_G\}$; and $\boldsymbol{\tau}=\{\tau_1,\ldots,\tau_G\}$.

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

$$\pi(\Theta)=\pi(\rho)\pi(\nu)\prod_{g=1}^{G}\left[\pi(\mu_g)\pi(\boldsymbol{\beta}_g)\pi(\lambda_g)\pi(\tau_g)\pi(\boldsymbol{\gamma}_g)\right], \tag{6}$$

and specify componentwise conditionally conjugate priors where suitable. Thus,

$$\pi(\mu_g)=N(\mu_g\mid 0,k), \tag{7}$$

$$\pi(\boldsymbol{\beta}_g)=N_G(\boldsymbol{\beta}_g\mid \mathbf{0},k_\beta I),\ g=1,\ldots,G, \tag{8}$$

$$\pi(\lambda_g)=\mathrm{Ga}(\lambda_g\mid a_\lambda,b_\lambda), \tag{9}$$

$$\pi(\tau_g)=\mathrm{Ga}(\tau_g\mid a_\tau,b_\tau), \tag{10}$$

$$\pi(\boldsymbol{\gamma}_g\mid\rho)=\prod_{j=1}^{G}\mathrm{Ber}(\gamma_{jg}\mid\rho),\ g=1,\ldots,G, \tag{11}$$

$$\pi(\rho)=\mathrm{Be}(\rho\mid a_\rho,b_\rho), \tag{12}$$

$$\pi(\nu)=\mathrm{Ga}(\nu\mid a_\nu,b_\nu). \tag{13}$$

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 gene j 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,

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

Γ supply the means for performing an automated link selection (see Smith and Kohn, 1996). The overall connectivity of the network is controlled by ρ and any relevant information about this key aspect can be fed into the model through its prior.

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 ($p = \dim\Theta \geq G(G+1)$) is typically much greater than the available data points ($n = T \times G$), i.e. the so-called $p \gg n$ 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 $\boldsymbol{\tau}$ and use a rather flat prior on this parameter, while carefully eliciting the parameters for $\pi(\boldsymbol{\lambda})$. To the extent of our knowledge, there is no conventional prior for the degrees of freedom, ν. We decided to use a gamma distribution such that $\mathrm{P}[\nu \leq 30] \approx 0.6$ and with mode at 15; thus giving roughly prior odds of 3 to 2 for the measurement error distribution being fat-tailed. A detailed specification of the prior parameters is given in Supplementary Section 2S.

There is no closed analytic expression for the posterior distribution, $\pi(\Theta \mid X) \propto \ell(\Theta; X)\pi(\Theta)$, and numerical methods to explore it are needed. To this end, we constructed a Markov chain Monte Carlo (MCMC) algorithm. We use Gibbs sampling for all of the parameters except B and ν. For the former, (Morrissey *et al.*, submitted for publication) showed that a Metropolis-within-Gibbs strategy improves mixing and, therefore, faster convergence of the chain, and we follow their suggestion. We use a Metropolis step for ν with a Gamma proposal centred at the previous draw and tune its coefficient of variation to control for the acceptance rate.

In passing, noteworthy is the Gibbs step used for drawing a new non-observable expression level y_g^t. These are drawn from a Gaussian distribution, $N(y_g^t \mid m_g^t, p_g^t)$, with location

$$m_g^t = \frac{\lambda_g m_{\mathrm{AR}} + \tau_g m_{\mathrm{meas}}}{p_g^t}$$

and precision

$$p_g^t = \lambda_g(1+\widetilde{\beta}_{gg}^2) + \tau_g \sum_{r=1}^{R} \omega_{gr}^t,$$

where

$$m_{\mathrm{AR}} = \sum_{i \neq g} \widetilde{\beta}_{ig}(y_i^{t-1} - \widetilde{\beta}_{gg} y_i^t) + \widetilde{\beta}_{gg}(y_g^{t-1} + y_g^{t+1})$$

and

$$m_{\mathrm{meas}} = \sum_{r=1}^{R} \omega_{gr}^t x_{gr}^t.$$

It is apparent from the expression above that draws of y_g^t depend on the weighted average of the observed measures, with the weights determined by the degrees of freedom through ω_{gr}^t. These averages are then combined with the AR(1) component of the model, thus effectively entertaining all sources of uncertainty. A detailed description of sampler is presented in Supplementary Section 3S.

Our estimation method is computationally intensive. Runtime for any of the measurement error models is not significantly longer than for the plain AR(1), though. Sampler's 2×10^5 iterations with the 16 gene linear data used in Section 4.1.1 took 2.10, 2.14 and 2.24 h for the AR(1), the Gaussian and the Student error models, respectively. In terms of scalability, the algorithms show the usual problems associated with network inference. Large datasets (i.e. thousands of genes) can take unrealistically long times to run. Runtime can be reduced by allowing only transcription factors to 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, ρ 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 η_g^t either Gaussian (GD) or Student (SD) distributed. For each distribution, we generated two datasets: one with few ($R = 4$) and a second with many ($R = 20$) replicates. In both cases, we consider rather noisy scenarios by setting $\tau_g^{-1/2}$ at 50% of the maximum absolute expression value of each gene (note that in the GD case, τ_g corresponds to the measurement precision). We fixed $\nu = 5$, 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 (M×E). 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

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

Table 1. Performance comparison using a synthetic linear network

	AUC				M×E			
	$R=4$		$R=20$		$R=4$		$R=20$	
	GD	SD	GD	SD	GD	SD	GD	SD
MM	0.78	0.68	0.91	0.82	0.40	0.59	0.24	0.33
GM	**0.86**	0.77	0.92	0.89	**0.32**	0.37	0.21	0.28
SM	0.85	**0.81**	**0.93**	**0.92**	**0.32**	**0.36**	**0.20**	**0.24**

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.

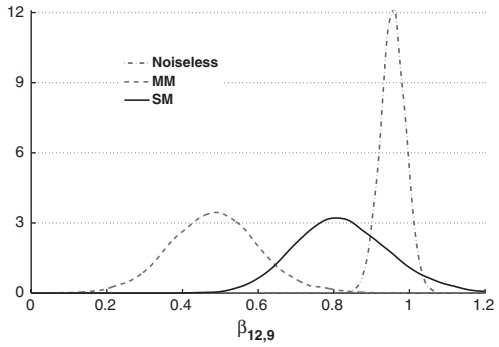

Fig. 1. Marginal posterior distributions of the coefficient, $\beta_{12,9}$ corresponding to link $\text{gene}_{12} \rightarrow \text{gene}_9$ of the *in silico* linear data. The noiseless data (dot-dashed) is the data without noise, the MM (dashed) is the mean of four replicates, neglecting measurement error and SM (solid) uses the four replicates and assumes Student measurement error.

divergence from the link structure of the true network to the posterior link probabilities, over all possible links (detailed in Supplementary Section 4.1S). As a baseline for comparison, the M×E of a perfectly inferred network is 0 and that of one predicted totally at random (i.e. probability of 1/2 for each link) is $-\log(1/2) \approx 0.7$. In the case of the AUC, this corresponds to values of 1 and 1/2, respectively. For instance, we fitted the AR(1) model with the ideal, noiseless data, resulting in an AUC of 0.99 and a M×E of 0.05.

When using GD, GM and SM perform equally well under both criteria—see the corresponding columns of Table 1. This is to be expected, since a Student distribution with large degrees of freedom approaches a Gaussian. Using a small number of replicates ($R=4$ in Table 1) GM and SM outperform MM in either criteria. When a large number of replicates is considered ($R=20$ in Table 1), the AUC for all three scenarios are quite close, indicating a similar ordering of the estimated link probabilities for all models fitted. However, the M×E scores are better for SM and GM, highlighting that the inferred probabilities with measurement error are comparatively higher for existent links and lower for non-existent.

Regarding inference on data with Student error, MM performs worse under both scores—first row in Table 1. In fact, the M×E score with few replicates is very close to that of random predictions. SM shows a small improvement over GM for both small and large R. This difference is slightly larger for $R=20$, illustrating that the degrees of freedom are hard to estimate; four replicates is barely enough to infer them, while $R=20$ allows for a more precise estimation.

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, $\text{gene}_{12} \rightarrow \text{gene}_9$. As expected in the ideal, noiseless case, the posterior distribution of the corresponding coefficient, $\beta_{12,9}$, 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. 1Sb). This is the 'noiseless' dataset. We generate noisy replicates using Student distributed errors. As before, we fix $\tau_g^{-1/2}$ 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,

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

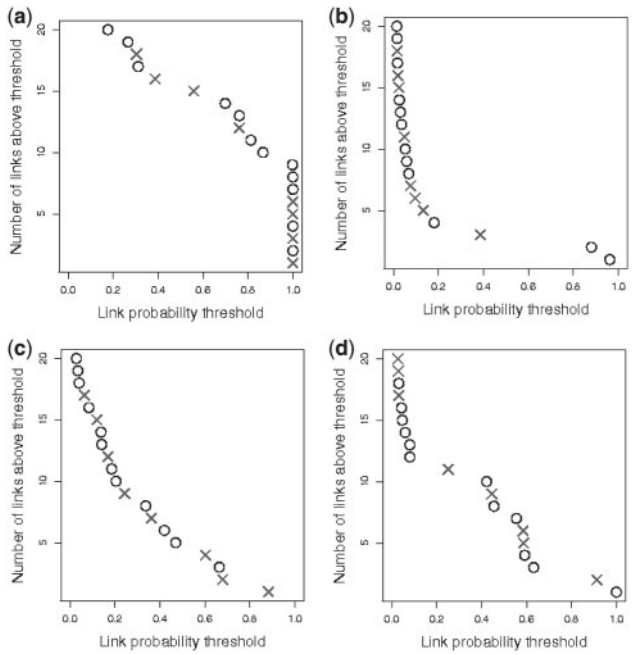

Fig. 2. Network topology retrieval for the ODE circadian clock *in silico* dataset with $R=4$ replicates. Links predicted to be present in the network versus posterior link probabilities estimated for the noiseless case and when using MM, GM and SM. A link (not) present in the ODE model is shown with a (circle) cross.

using a link probability threshold of 0.7, 14 out of 20 possible links are predicted to be present. The main reason for this is that the linear model is unable to explain the nonlinear interactions adequately, resulting in either a misprediction or a compensation by the inclusion of spurious parents (see Morrissey *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 towards 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 *A.thaliana* (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 (Pruneda-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 PRR9 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→GI→TOC1 loop, but also predicts two incorrect links (GI→CCA1 and TOC1→PRR7). The SM model also correctly predicts TOC1→GI but incorrectly TOC1→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→CCA1, CCA1→TOC1 and two incorrect: PRR7→GI and PRR7→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 τ_g, these are treated as interchangeable in the update of y_g^t (in this case, $\omega_{gr}^t \equiv 1$). In contrast, the weights in SM depend on the degrees of freedom and will be more variable for smaller values of ν, allowing for some of the measurements to dominate the average. To verify that this is

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

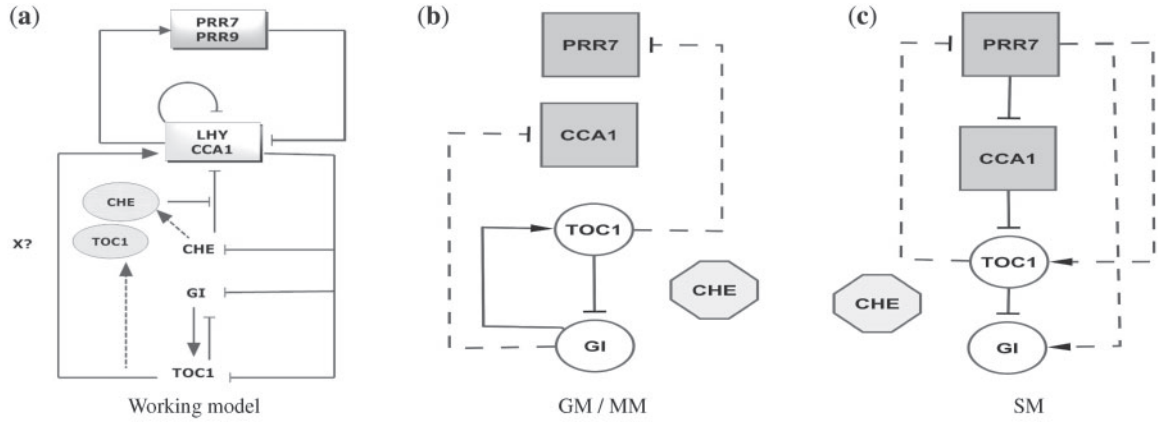

Fig. 3. *Arabidopsis* circadian clock. (**a**) Current working model (redrawn from Robertson and Webb, 2009): dotted lines represent protein production and oval shapes (binding) proteins. (**b**) and (**c**) Depict the network topologies inferred with the measurement error models with a threshold of 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.

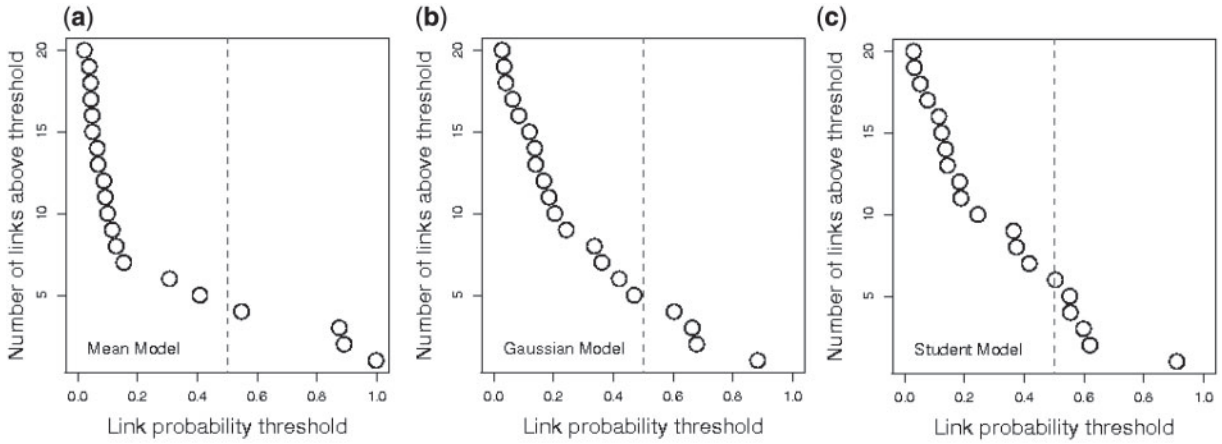

Fig. 4. Gene network link prediction for the *Arabidopsis* circadian clock microarray data. Posterior link probabilities from each model are depicted as circles. The vertical dotted line represents a threshold of 1/2.

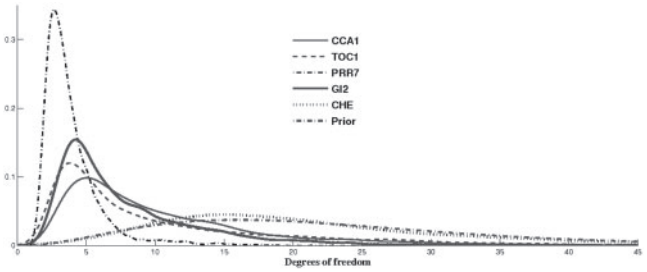

Fig. 5. Marginal posterior distributions of the degrees of freedom, ν, for each gene when fitting the Student measurement error model to the *Arabidopsis* circadian clock microarray data.

indeed the case, we fitted the AR(1) model using the median value of the replicates as the sole source of information. The predicted network topology (not shown) also has six links with only one of them differing from those inferred by SM (incorrect in both cases); nevertheless, the posterior distribution of the link probabilities was tighter, illustrating the sub-estimation of the coefficients' precision.

These comparisons are dependant 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

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

bind TOC1 and also binds the promoter region of CCA1. This suggests that we can expect non-additive, nonlinear effects that may be difficult to capture with an additive linear model. For the regulation of CHE, we do not expect these problems. From Pruneda-Paz *et al.* (2009), we know that CHE is regulated by CCA1. Analysing the SM posterior link probabilities, we found that the other four genes appear as regulators of CHE with posterior probabilities of around 0.13, suggesting that either the regulation is highly nonlinear, or there are other regulatory mechanisms such as post-translational modifications. The latter would not be unlikely as it is well known that targeted protein degradation and sequestration plays a very important role in the circadian clock.

We further explored the effect of having less replicates available. To this end, we sub-sampled the original four replicate dataset and generated 60 datasets, 30 with three replicates and 30 with one. The datasets were sampled in a controlled manner to ensure they were not too similar to the original four replicate dataset (see Supplementary Section 4.3S).

We summarize the information by setting the same link probability threshold as in the four replicate case ($\hat{\gamma}_{ij} \geq 0.5$), and then counted the times a link was predicted to be present. These counts are shown in Table 2 for the 1-replicate case and Table 3 for the 3-replicate. As expected, the 'no replicates' scenario (Table 2) shows high variability in the inferred networks. No single link was predicted as present 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

	CCA1	TOC1	PRR7	GI	CHE
CCA1	0	3	3	4	1
TOC1	6	0	3	14	0
PRR7	2	14	0	2	1
GI	0	13	6	0	0
CHE	1	3	0	0	0

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 overestimation of the precision of these estimates. This in turn can 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, $\boldsymbol{\lambda}$, and the measurement precision, $\boldsymbol{\tau}$. When $\boldsymbol{\tau}$ 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 $\boldsymbol{\lambda}$ must thence 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

	MM					GM					SM				
	CCA1	TOC1	PRR7	GI	CHE	CCA1	TOC1	PRR7	GI	CHE	CCA1	TOC1	PRR7	GI	CHE
CCA1	0	0	1	**17**	0	0	0	2	**24**	0	0	0	**21**	7	0
TOC1	1	0	1	**29**	0	4	0	1	**22**	0	**15**	0	7	5	0
PRR7	1	**29**	0	3	0	0	**30**	0	5	0	0	**30**	0	3	0
GI	1	**29**	1	0	0	0	11	**20**	0	0	0	9	**24**	0	0
CHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

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.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

are only point estimates of the network topology. Thus, the final effect on the inferred network topology is case dependant. 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

Bansal,M. *et al.* (2007) How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, **3**, 78.

Brody,J.P. *et al*. (2002) Significance and statistical errors in the analysis of DNA microarray data. *Proc. Natl Acad. Sci. USA*, **99**, 12975–12978.

Cao,J. and Zhao,H. (2008) Estimating dynamic models for gene regulation networks. *Bioinformatics*, **24**, 1619–1624.

Carroll,R.J. *et al.* (2006) *Measurement error in nonlinear models: A modern perspective*, 2nd edn. Chapman & Hall/CRC, Boca Raton.

Friedman,N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Fujita,A. *et al.* (2009) The impact of measurement errors in the identification of regulatory networks. *BMC Bioinformatics*, **10**, 412.

Fuller,W.A. (1987) *Measurement error models*. Wiley, New York.

Gustafson,P. (2004) *Measurement error and Misclassification in Statistics and Epidemiology. Impacts and Bayesian adjustments*. Chapman & Hall/CRC, Boca Raton.

Gustafson,P. (2005) On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Stat. Sci.*, **20**, 111–140.

Hache,H. *et al.* (2009) Reverse engineering of gene regulatory networks: a comparative study. *EURASIP J. Bioinform. Syst. Biol.*, **2009**, 1–12.

Hongqiang,L. *et al.* (2005) Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Human Mol. Genet.,*, **14**, 1119–1125.

Hoops,S. *et al.* (2006) COPASI: a COmplex PAthway SImulator. *Bioinformatics*, **22**, 3067–3074.

Lewin,A. *et al.* (2007) Fully Bayesian mixture model for differential gene expression: simulations and model checks. *Stat. Appl. Genet. Mol. Biol.*, **6**, 36.

Locke,J.C.W. *et al.* (2006) Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Mol. Syst. Biol.*, **2**, 59.

Murphy,K. and Mian,S. (1999) Modelling gene expression data using dynamic Bayesian networks. *Technical report*, Computer Science Division, University of California, Berkeley.

Perrin,B. *et al.* (2003) Gene network inference using dynamic Bayesian networks. *Bioinformatics*, **19**, ii138–ii148.

Poirier,D.J. (1998) Revising beliefs in nonidentified models. *Econometric Theory*, **14**, 483–509.

Pruneda-Paz,J.L. *et al.* (2009) A functional genomics approach reveals CHE as a component of the *Arabidopsis* circadian clock. *Science*, **323**, 1481–1485.

Robertson,F.C. and Webb,A.A.R. (2009) Revolutionary functional genomics liberates CHE. *Nat. Chem. Biol.*, **5**, 276 – 277.

Schmid,C.H. *et al.* (1994) Incorporating measurement error in the estimation of autoregressive models for longitudinal data. *J. Stat. Plan. Inference*, **42**, 1–18.

Smith,M. and Kohn,R. (1996) Nonparametric regression using Bayesian variable selection. *J. Econom.*, **75**, 317–343.

Yu,J. *et al.* (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, **20**, 3594–3603.

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

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016