Inference of Circadian Regulatory Networks

Marco Grzegorczyk[1], Andrej Aderhold[2], V. Anne Smith[2], and Dirk Husmeier[3]

[1] Johann Bernoulli Institute, University of Groningen, The Netherlands
m.a.grzegorczyk@rug.nl,
[2] School of Biology, University of St Andrews, UK
[3] School of Mathematics and Statistics, University of Glasgow, UK

Abstract. We assess the accuracy of various state-of-the-art methods for reconstructing gene and protein regulatory networks in the context of circadian regulation. Gene expression and protein concentration time series are simulated from a recently published regulatory network of the circadian clock in *A. thaliana*, which is mathematically described by a Markov jump process based on Michaelis-Menten kinetics. Our study provides relative network reconstruction accuracy scores for a critical comparative performance evaluation, quantifies the influence of systematically missing values related to unknown protein concentrations and mRNA transcription rates, and investigates the dependence of the performance on the network topology and the degree of recurrency. An application to recent gene expression time series from qPCR experiments suggests new hypotheses about the structure of the central circadian gene regulatory network in *A. thaliana*.

Keywords: Regulatory network inference, Circadian clock, ANOVA

1 INTRODUCTION

The ultimate objective of systems biology is the elucidation of the regulatory networks and signalling pathways of the cell. The ideal approach would be the deduction of a detailed mathematical description of the entire system in terms of coupled non-linear differential equations. In fact, in the last few years, substantial progress has been made to model the central processes of a variety of regulatory networks [1, 2, 3, 4, 5]. However, due to the complexity of the underlying systems of coupled differential equations, proper statistical inference is extremely challenging. Consequently, there are substantial differences and inconsistencies among published gene regulatory networks of biological model systems, as demonstrated e.g. in [6]. This lack of knowledge is aggravated when trying to understand densely connected networks with strong feedback mechanisms. What is needed is reliable statistical inference of the structure of the molecular regulatory networks and signalling pathways by utilization and systematic integration of transcriptomic, proteomic and metabolic concentration proles.

In fact, statistical inference of molecular regulatory networks from post-genomic data has been a central topic in computational systems biology for over a decade. Following up on the seminal paper of [7], a variety of abstract

models from machine learning and computational multivariate statistics have been proposed as an alternative to mechanistic models based on differential equations (see Section 2), and their comparative assessment is an active area of ongoing research [8]. Our article complements this work in several ways. We have used a more realistic simulation process that allows for the intrinsic stochasticity of transcription initiation and translation and is specically targeted at densely connected feedback systems (Section 3.1). We have quantied the dependence of the network reconstruction accuracy on the network connectivity and the degree of recurrency. We have compared different alternative ways to approximate unknown de novo mRNA transcription rates. We have quantied the effect of missing data related to unknown protein concentrations. And we have carried out a comparative evaluation of various state-of-the-art methods from machine learning and computational statistics with a multivariate ANOVA scheme. Following on from this systematic comparative assessment, we have applied the best method identied to recent gene expression time series for the genes in the central clock of *Arabidopsis thaliana*, and have compared the predicted network with various networks proposed in the molecular systems biology literature.

2 METHOD OVERVIEW

Sparse Regression (Lasso and Elastic Nets): An efficient and widely applied penalized linear regression method for sparse network reconstruction is the Least Absolute Shrinkage and Selection Operator (Lasso), introduced in [9, 10] and first applied to systems biology in [11]. The Lasso optimizes the regression parameters of a linear model based on the residual sum of squares subject to an $L1$-norm regularization term that simultaneously shrinks and selects non-zero regression parameters. The Elastic Net method, proposed in [12], combines the $L1$ penalty with an $L2$ penalty to address two problems of the Lasso, related to saturation effects and the selection of correlated variables.
Tesla: A time-varying generalization of sparse regression, called Tesla, was proposed in [13]. The idea is to divide a time series into segments and perform sparse regression for each time series segment separately. Each segment is associated with a different set of regression parameters. To prevent over-complexity and avoid overfitting, an additional $L1$-norm penalty is imposed on the parameter differences for adjacent time series segments. We selected light as the primary segmentation criterion, and grouped measurements obtained under the same light condition (light versus darkness) together. The original formulation of Tesla in [13] is for logistic regression and binary data. The modification to linear regression is straightforward and more appropriate for our application.
Hierarchical Bayesian regression ('HBR'): The HBR model can be regarded as a Bayesian generalization of Tesla, where parameters are sampled from the posterior distribution with MCMC, and inference borrows strength from the systematic coupling of related variables. We implemented the method as described in [14], with the same conjugate priors, and we used a fixed data segmentation to reflect the light phase (light versus dark).

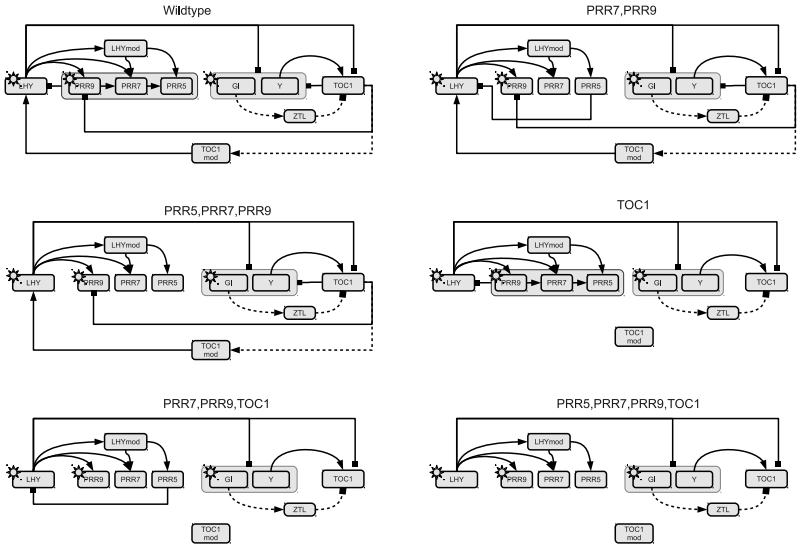


Fig. 1. Model network of the circadian clock in *Arabidopsis thaliana* and network modifications. Each graph shows interactions among core circadian clock genes. Solid lines show protein-gene interactions; dashed lines show protein modifications; and the regulatory influence of light is symbolized by a sun symbol. The top left panel ("Wildtype") shows the network structure published in [2]. The remaining panels show modified network structures, corresponding to constant knockouts of the proteins shown above the corresponding network structure. Grey boxes group sets of regulators or regulated components.

Gaussian Processes: Gaussian processes (GPs) are a powerful nonlinear regression method from nonparametric Bayesian statistics. We have applied the approach described in [15], where the GP framework was specifically adapted to equation (1) and the problem of learning gene regulatory networks, inferring parameters by marginal likelihood maximization.

3 DATA

3.1 Realistic data

Various mathematical models have been developed to describe the molecular interactions and signal transduction processes in the central circadian clock of *Arabidopsis thaliana* [1, 3, 5]. They are based on systems of ordinary differential equations (ODEs) that describe the chemical kinetics of transcription initiation, translation, and post-translational modification, using mass action kinetics and/or Michaelis-Menten kinetics. A limitation of ODEs is that they typically converge to limit cycles with regular oscillations and constant amplitude, which fail to capture the stochastic amplitude variation observed in real qPCR experiments. For a more realistic approach, which captures the intrinsic fluctuations of

molecular processes in the cell, we modelled the individual molecular processes of transcription, translation, degradation, dimerization etc. as individual discrete events. Statistical mechanics arguments then lead to a Markov jump process in continuous time whose instantaneous reaction rates are directly proportional to the number of molecules of each reacting component [16, 17]. We followed [4] and adopted the Bio-PEPA framework [18] to simulate gene expression profiles for the core circadian clock of *Arabidopsis thaliana*, using the Bio-PEPA Eclipse Plug-in[4]. This framework is built on a stochastic process algebra implementation of chemical kinetics, and the stochastic simulations are run with the Gillespie algorithm [19].

We simulated mRNA and protein concentration profiles over time from the circadian clock regulatory network published in [4] and [2], shown in the top left panel of Figure 1, named 'Wildtype'. This involves genetic regulatory reactions for mRNA transcription, protein translation, and mRNA and protein degradation for 9 genes. A full list of reactions and their corresponding mathematical descriptions is available from the supplementary material of [4]. In addition, we simulated mRNA and protein concentration time series from a series of modified network structures, see Figure 1. For each of these network types we created 11 interventions, in emulation of the biological protocols of [20] and [21]. These interventions include knock-outs of the proteins GI, LHY, TOC1, and the double knock-out PRR7/PRR9. The knock-outs were simulated by setting, in each step of the Markov jump process, the concentrations of the targeted proteins to zero. Following these simulations, their values were replaced by random noise, drawn from a truncated normal distribution (to ensure non-negativity of the concentrations). Again, in emulation of the biological protocols of [20] and [21], we simulated varying photo-periods of 4, 6, 8, 12, and 18 hours as well as a full dark (DD) and a full light (LL) cycle, each following a 12h-12h light-dark cycle entrainment phase over 5 days. For each type of intervention, concentration time series were generated to encompass a simulated epoch of 6 days, of which the first 5 days were used for entrainment. After entrainment, molecule counts of mRNA and proteins were recorded in 2 hour intervals of simulated time, for 24 hours, giving a total of 13 'observations'. Combining these 13 observations for each intervention type yields 143 observations in total for each of the regulatory network structures shown in Figure 1. For each intervention type and sampling interval length, five independent data sets were generated. To standardize the data, we followed the widely established procedure to rescale all molecule concentrations to zero mean and unit standard deviation. Two different data types were used in our evaluation procedures: complete data, which include both the mRNA and the protein concentrations, and incomplete data, in which the protein concentrations are missing and regulatory network structures have to be inferred on the basis of mRNA concentrations alone. In summary, we generated data for six different network structures, shown in Figure 1, repeating each data generation 5 times independently (i.e. starting from different random number

[4] http://www.biopepa.org

generator seeds), and using complete observations (mRNAs and proteins) versus mRNA concentrations only.

3.2 Real application

We also used real transcription profiles for the key circadian regulatory genes in the model plant *A. thaliana*. The data used in our study come from the EU project TiMet [22], whose objective is the elucidation of the interaction between circadian regulation and metabolism in plants. The data consist of transcription profiles for the core clock genes from the leaves of various genetic variants of *A. thaliana*, measured with qPCR. The study encompasses two wildtypes of the strains Columbia (Col-0) and Wasilewski (WS) and 4 clock mutants, namely a double knock-out 'LHY/CCA1' in the WS strain, a single knock-out of 'GI' and 'TOC1' in the strain Col-0, and a double-knockout 'PRR7/PRR9' in strain Col-0. The plants were grown in the following 3 light conditions: a diurnal cycle with 12 hours light and 12 hour darkness (12L/12D), an extended night with full darkness for 24 hours (DD), and an extended light with constant light (LL) for 24 hours. Samples were taken every 2 hours to measure mRNA concentrations; see [20]. We focus on the genes that are included in the models from the literature: LHY, CCA1, PRR5, PRR7, PRR9, TOC1, ELF3, ELF4 with a total of 288 samples per gene. We used the log mean copy number of mRNA per cell and applied a gene wise Z-score transformation for data standardization. An additional binary light indicator variable was included to indicate the status of the light condition.

4 METHODOLOGICAL DETAILS

4.1 Rate estimation

Based on the fundamental equation of transcriptional regulation

$$\frac{dy_i}{dt} \;=\; c_i + f_i(\mathbf{x}_i(t), \boldsymbol{\theta}) - \lambda_i y_i(t) \tag{1}$$

introduced in [23], where i refers to a gene in the biopathway, $y_i(t)$ denotes its mRNA concentration at time t, c_i is a baseline production rate, λ_i is a decay rate, $\mathbf{x}_i(t)$ is a vector of concentrations of potential regulators that control the concentration of mRNA i, and $\boldsymbol{\theta}$ is a vector of regulation parameters, all methods described in Section 2 aim to predict the time derivatives of the target mRNA concentrations from the (mRNA or protein) concentrations of the putative regulators. In the absence of de novo mRNA data, we approximated these derivatives with finite difference quotients, $\frac{dy}{dt} \approx \frac{y(t+\delta t)-y(t-\delta t)}{2\delta t}$, trying two different time intervals: $\delta t = 2$ hours (coarse gradient), and $\delta t = 24$ minutes (fine gradient). Alternatively, we used an approach based on smooth interpolation with Gaussian processes, exploiting the fact that the derivative of a Gaussian process is also a Gaussian process; hence analytic expressions for the mean and the standard deviation of the derivative are available [24].

4.2 Implementation details

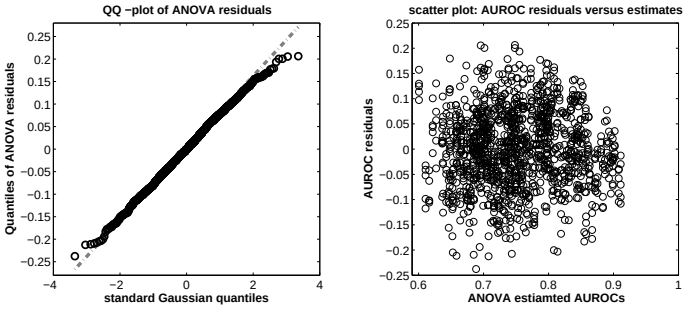
For the sparse regression models a 10-fold cross-validation scheme was applied to optimize the regularization parameters. For Lasso and the Elastic Net we optimized the regression parameters with cyclical coordinate descent, using the R-package *GLMNET* from CRAN. Tesla was run with a linear regression implementation, and the regression parameters were optimized with convex programming, using the software from [13]. Absolute values of non-zero regression coefficients were used for ranking molecular interactions. For Tesla we have two segments (light versus darkness) with potentially different regression parameters, and we used the average absolute values of the non-zero regression coefficients for ranking the molecular interactions. The MCMC simulations for hierarchical Bayesian regression (HBR) were run for 20,000 iterations each, with a burn-in period of 10,000 iterations discarded. This choice gave satisfactory convergence diagnostics, based on correlation scatter plots and Gelman-Rubin potential scale reduction factors [25]. Marginal posterior probabilities of molecular interactions were obtained from the MCMC trajectories, estimated from the relative frequency of inclusion of the corresponding edges in the sampled models. For the Gaussian process we used the implementation in the *GP4GRN* software package, developed in [15].

4.3 Network Inference Scoring Scheme

The methods under comparison provide means by which interactions between genes and proteins can be ranked in terms of their significance or influence. If the true network is known, this ranking defines the Receiver Operating Characteristic (ROC) curve, where for all possible threshold values, the sensitivity or recall is plotted against the complementary specificity. By numerical integration we then obtain the area under the curve (AUROC) as a global measure of network reconstruction accuracy, where larger values indicate a better performance, starting from AUROC=0.5 to indicate random expectation, to AUROC=1 for perfect network reconstruction. We note that for networks of the size considered here a performance evaluation based on AUROC and AUPRC (area under the precision-recall curve) gives very similar results [26], and AUROC scores have the advantage of a clearer statistical interpretation [27].

4.4 ANOVA

For our evaluation, we were running hundreds of simulations for a variety of different settings, related to the observation status of the molecular components (mRNA only versus mRNAs and proteins), the method for derivative (rate) estimation, the regulatory network structure (shown in Figure 1), and the method applied for learning this structure from data (see Section 2). The AUROC results are complex and elude clearly discernible patterns and trends. In order to disentangle the different factors, and in particular distinguish the effect of the model from the other confounding factors, we adopted the DELVE evaluation procedure for comparative assessment of classification and regression methods



(a) Quantile Quantile (QQ) plot (b) Residual scatter plot

Fig. 2. Diagnostics for the ANOVA model, see equation (2). (a) In the QQ plot the quantiles of the residuals (vertical axis) are plotted against the quantiles of the Gaussian distribution (horizontal axis). The linear relation indicates a good agreement with the Gaussian distribution; the deviations for very low and high values point to slightly longer tails. **(b)** Scatter plot of the residuals (vertical axis) against the AUROC values fitted with the ANOVA model (horizontal axis): For high values, the spread of the residuals seems to become slightly tighter, but this effect is weak, and overall there is no clearly discernible pattern of any dependence between the residuals and the fitted values.

[28] and set up a multi-way analysis of variance (ANOVA) scheme (see, e.g.,[29]). Let Y_{ognmk} denote the AUROC score obtained for observability status o, gradient computation g, network topology n, network reconstruction method m, and data instantiation k. The range of these index parameters is as follows: $o \in \{0, 1\}$, where $o = 0$ indicates partial (mRNAs only) and $o = 1$ complete (mRNAs and proteins) observation; $g \in \{0, 1, 2\}$, where $g = 0$ denotes coarse gradient, $g = 1$ fine gradient, and $g = 2$ gradient from a smooth GP interpolant; $m \in \{0, 1, 2, 3, 4, 5\}$, where $m = 0$ represents "wildtype" (the published network topology), and $m \neq 0$ the five network modifications shown in Figure 1; $n \in \{0, 1, 2, 3, 4\}$, for Lasso (0), Elastic Net (1), Tesla (2), GP (3), and hierarchical Bayesian regression (4); and $k \in \{0, 1, 2, 3, 4\}$ for five different data instantiations. We model the AUROC scores with the following ANOVA approach:

$$y_{ognmk} \;=\; O_o + G_g + N_n + M_m + \varepsilon_{ognmk} \tag{2}$$

where $\varepsilon_{ognmk} \sim N(0, \sigma^2)$ is zero-mean white additive Gaussian noise, and O_o, G_g, N_n, and M_m are main effects associated with observation status, gradient computation, network topology, and network reconstruction method, respectively. Figures 2 and 3 show the results of a standard residual analysis. Since the results do not indicate any violation of the model assumptions, the ANOVA analysis provides an adequate mechanism for extracting trends and patterns from our simulations studies.

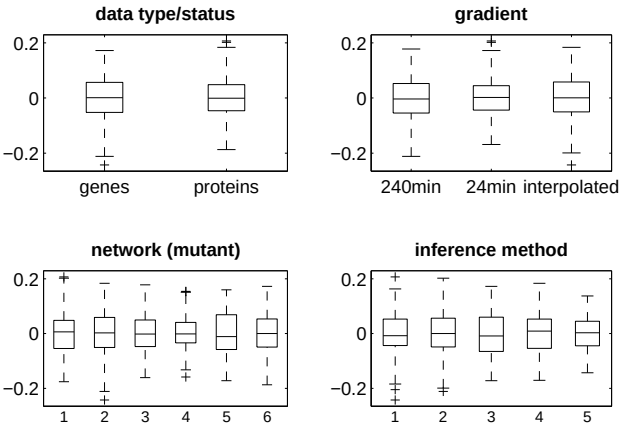


Fig. 3. Diagnostics for the ANOVA model, see equation (2), continued Boxplot representations of the distribution of the residuals for all possible values of the four main effects. There are no obvious deviations from a uniform pattern, and the results are consistent with the assumption that the distributions of the residuals are identical and independent of the main effects.

5 RESULTS

5.1 Comparative evaluation study

Comparison between the methods: A main objective of our study is a systematic comparative performance evaluation of the models reviewed in Section 2. These models were applied to the different data described in Section 3.1, different observabilities (proteins and mRNAs versus mRNAs only), different gradient computations (Section 4.1), and different network topologies (as shown in Figure 1). The AUROC scores vary considerably, depending on the different factors. To enable a clearer interpretation we adopted the ANOVA method described in Section 4.4. The quantity of interest is M_m - the main effect of the network reconstruction method, which is plotted in Figure 4(a). Our study suggests that the hierarchical Bayesian regression model performs significantly better than the Gaussian Process method of [15] and sparse regression methods (Lasso, Elastic Net, and Tesla), justifying the higher computational costs required for the inference scheme (MCMC).[5] **Influence of rate estimation:** In the present study we estimated the time derivatives of mRNA concentrations directly from the mRNA concentration time courses. We approximated the time derivatives by finite difference quotients from the low frequency time series, where observations were taken every $\delta t = 2h$ hours ('coarse gradient'). Alternatively, we tried a finer resolution of $\delta_t = 24$ minutes around the main time points ('fine gradient'). As a

[5] The approximate average run times per data set on an Intel(R) Xeon(R) CPU X5570 with 2.93GHz were $<$ 1min for Lasso and Elastic, 2.5min (50min) for Tesla without (with) 10-fold cross validation, 8.5min for GP, and 21 min for HBR.

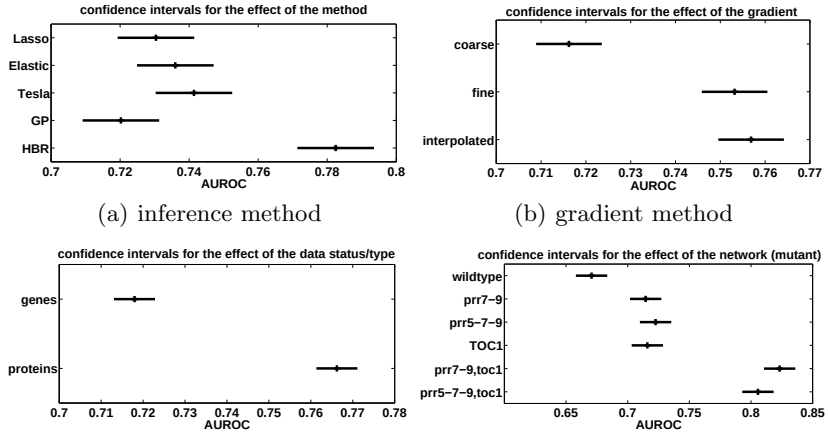
further alternative, we applied a Gaussian process smoothing approach. The results are shown in Figure 4(b). The fine gradient achieves an improvement on the coarse gradient, which is consistent with expectation. However, our study also allows a quantification of the improvement, which is in the order of Δ AUROC $= 0.035$ on average. Interestingly, our study suggests that gradient computation in combination with smooth interpolation using Gaussian processes achieves an even more substantial improvement of about Δ AUROC $= 0.041$. This indicates that intelligent data preprocessing leads to a better boost in predictive performance than blindly carrying out additional experiments.

Influence of missing protein concentrations: We have carried out the simulations for two types of data: complete observation, where both protein and mRNA concentrations are available, and partial observation, where protein concentrations are missing. The results are shown in Figure 4(c) and confirm the expected trend that the network reconstruction accuracy becomes worse with missing data. The important new contribution of our study is to objectively assess the difference in performance, profiled over different network topologies, different ways of preprocessing the data, and different statistics and machine learning methods. This has been effected with the ANOVA approach described in Section 4.4, which quantifies the effect of missing protein concentrations as leading to a deterioration of $\Delta AUROC = 0.05 \pm 0.01$.

Influence of network topology and feedback loops: An important aspect of our study is the investigation of how the network reconstruction accuracy depends on the connectivity of the network topology and the proportion of recurrent connections. To this end we have successively pruned feedback interactions, as shown in Figure 1. Figure 4(d) suggests that there is a noticeable pattern, with less recurrent network structures appearing to be easier to learn and leading to higher AUROC scores. While this confirms a known and intuitively plausible trend, our study allows an objective quantification of the difference in performance, which has been found to amount to $\Delta AUROC = 0.14$ between the most and least recurrent structures.

5.2 Circadian regulation network in *Arabidopsis thaliana*

Figure 5 shows the network learned from the TiMet data, and four hypothetical networks published in [1] and [2, 3, 5]. Solid lines show transcriptional regulation, dashed lines represent protein complex formation. The latter cannot be learned from transcriptional data and are thus systematically missing. This explains, for instance, why ZTL and TOC1 are detached from the remaining network. The same applies to the modified proteins TOC1-mod and LHY-mod. Various features of the published networks are reproduced, though, like the acute light response in the transcription of LHY and CCA1, the activation of PRR7 by PRR9, the inhibition of GI by LHY/CCA1, and the inhibition of ELF4 by TOC1, which can be found in the network P2013. Various features are similar to the published networks. In the reconstructed network, PRR5 is directly activated by PRR9, while in the published networks, the activation is indirect, via PRR7. The positive feedback loop from the so-called evening genes to the morning genes



(a) inference method (b) gradient method

(c) data type/status (d) network structure

Fig. 4. Confidence intervals for the group means associated with the four main effects from the ANOVA analysis. (a) Effect of the inference method: Lasso, Elastic Net, Tesla, Gaussian Processes (GP), and hierarchical Bayesian regression (HBR). **(b)** Effect of the three rate estimation method: coarse gradient (top), fine gradient (middle), and gradient from a smooth GP interpolant (bottom). **(c)** Effect of the observation status: complete observations of both protein and mRNA concentrations ('proteins') versus the observation of mRNA concentrations only ('genes'). **(d)** Effect of the network structure: the wildtype network and the five modified structures shown in Figure 1. As one descends from the top to the bottom, the network structures become sparser, with feedback loops increasingly being pruned.

consists of an activation of LHY/CCA1 by GI. The nature of this feedback loop (activation) is consistent with [2]. In these publications, the regulatory influence is caused by TOC1 rather than GI, but these two genes are "neighbours" in the published networks (meaning: regulating each other, and exhibiting similar expression profiles). One of the PRR-genes (PRR5) is predicted to be inhibited by ELF3. This is consistent with [3, 5], although in these publications, the interaction is indirect (via EC) and affects a neighbouring target gene (PRR9). As mentioned above, it is intrinsically unfeasible to learn post-transcriptional processes, like protein complex formation, from transcriptional data alone; so it is no surprise to see that the protein complex EC is detached from the remaining network. It is particularly interesting to note that a key network motif repeatedly found in the reconstructed network concurs with the published networks. This is the two-node feedback motif in which a gene is the activator of its own inhibitor. This structure is particularly clearly seen in Locke2006 [1], where it occurs three times: within the group of morning genes (LHY/CCA1 activating PRR7/PRR9, PRR7/PRR9 inhibiting LHY/CCA1), within the group of evening genes (GI activating TOC1, TOC1 inhibiting GI), and between the morning and evening genes (LHY/CCA1 inhibiting TOC1, TOC1 activating LHY/CCA1). These three feedback mechanisms exist in the reconstructed net-

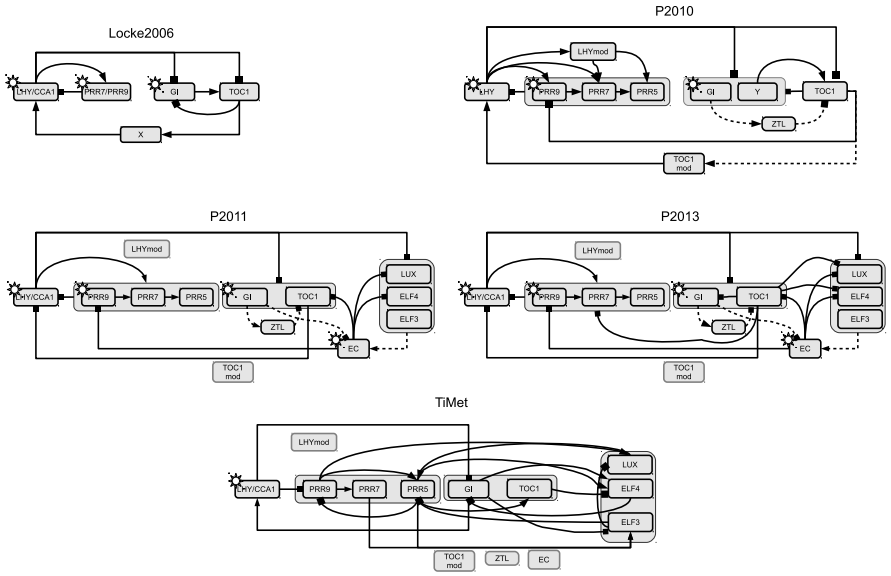


Fig. 5. Hypothetical circadian clock networks from the literature, and inferred from the TiMet gene expression data. All panels except for the bottom show hypothetical networks from the literature: Locke2006 [1], P2010 [2], P2011 [3], and P2013 [5]. The bottom panel (TiMet) displays the reconstructed network from the TiMet data, described in Section 3.2, using the hierarchical Bayesian regression model. Gene interactions are shown by solid lines, protein interactions are shown by dashed lines, and regulation by light is represented by a sun symbol. Arrow head: activation; vertical bar: inhibition. The interactions in the reconstructed network were obtained from their estimated posterior probabilities. Those above the selected threshold of 0.95 were included in the interaction network; those below were discarded.

work also, involving neighbouring nodes in the same three gene groups: morning genes (PRR9 activating PRR5, PRR5 inhibiting PRR9), evening genes (GI activating ELF4, ELF4 inhibiting GI), and between morning and evening genes (GI activating LHY/CCA1, LHY/CCA1 inhibiting GI). This suggests that, despite deviations in the detailed mechanisms, the key topological features of the published networks have been successfully reconstructed.

6 DISCUSSION

We have carried out a comparative evaluation of five state-of-the-art methods for regulatory network reconstruction, using the central gene regulatory network of the circadian clock in *A. thaliana*. Our results confirm various intuitively plausible trends: that the difficulty of network reconstruction increases with increasing network connectivity, that network reconstruction deteriorates with incomplete observation (missing protein concentrations), and that for estimating de novo

mRNA transcription rates, data smoothing has a beneficial effect. The novel contribution of our study consists in objectively quantifying these effects, in terms of average AUROC score differences associated with the respective main effects in the ANOVA scheme. For the model comparison, we have shown that hierarchical Bayesian regression outperforms penalized linear regression and Gaussian processes, again objectively quantifying the performance gain. We have applied the best network reconstruction method from the comparative assessment to the mRNA concentration profiles from the TiMet project. The reconstructed network contains several topological features that are consistent with recently published regulatory networks of the circadian clock in *A. thaliana*. However, the detailed structure differs in various aspects. This difference is a consequence of the different nature of the methods. For the networks published in the literature, the processes of transcriptional regulation were modelled with ordinary differential equations. The network structures were not selected with rigorous statistical inference; doing that e.g. with the procedure proposed in [30] is computationally prohibitive. The consequence is an intrinsic uncertainty about the true network structure, as discussed in [6] and evidenced by repeated recent network modifications in the literature (see Figure 5). The methods applied in the present article are based on more abstract models of molecular regulatory interactions, which render objective statistical inference viable. Hence, our understanding of circadian regulation at the molecular level will potentially improve as a consequence of a synthesis of both approaches, which will suggest novel avenues for model adjustment. The evaluated network reconstruction methods are particularly useful for linking circadian regulation in plants to metabolism, due to the current absence of detailed hypotheses and reliable mechanistic models.
The aim of our future work is to treat the interventional data more realistically. When a gene is knocked out, we currently set the concentration of the corresponding protein to zero. This corresponds to post-transcriptional gene silencing, in which the translation of mRNA into functional protein is inhibited. What is actually happening in the experiments that motivated our study (reported in Section 5.1) is that certain plant genes are knocked out by mutagenesis. This can be simulated more realistically by setting the corresponding mRNA rather than protein concentrations to zero, which requires a modification of our current simulation set-up. Due to space and time constraints, our present study has focused on regression-type models. Various other methods have been proposed in the literature, including state-space models, Bayesian networks and approaches based on mutual information. We aim to expand our current work to include them in our comparative evaluation.

Acknowledgments. The work described in the present article is part of the TiMet project on linking the circadian clock to metabolism in plants. TiMet is a collaborative project (Grant Agreement 245143) funded by the European Commission FP7, in response to call FP7-KBBE-2009-3. Parts of the work were done while M.G. was supported by the German Research Foundation (DFG), research grant GR3853/1-1. A.A. is supported by the BBSRC. We are grateful to Andrew Millar and Alexander Pokhilko for helpful discussions.

Bibliography

[1] Locke, J., Kozma-Bognár, L., Gould, P., Fehér, B., Kevei, E., Nagy, F., Turner, M., Hall, A., Millar, A.: Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana. Molecular systems biology **2**(1) (2006)

[2] Pokhilko, A., Hodge, S., Stratford, K., Knox, K., Edwards, K., Thomson, A., Mizuno, T., Millar, A.: Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. Molecular systems biology **6**(1) (2010)

[3] Pokhilko, A., Fernández, A., Edwards, K., Southern, M., Halliday, K., Millar, A.: The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. Molecular systems biology **8** (2012) 574

[4] Guerriero, M., Pokhilko, A., Fernández, A., Halliday, K., Millar, A., Hillston, J.: Stochastic properties of the plant circadian clock. Journal of The Royal Society Interface **9**(69) (2012) 744–756

[5] Pokhilko, A., Mas, P., Millar, A., et al.: Modelling the widespread effects of TOC1 signalling on the plant circadian clock and its outputs. BMC systems biology **7**(1) (2013) 1–12

[6] Bujdoso, N., Davis, S.: Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of *Arabidopsis thaliana*. Frontiers in Plant Science **4** (2013) Article 3

[7] Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. Journal of Computational Biology **7** (2000) 601–620

[8] Marbach, D., Costello, J., Küffner, R., N.M., V., R.J., P., Camacho, D., Allison, K., Consortium, D., Kellis, M., J.J., C., G., S.: Wisdom of crowds for robust gene network inference. Nature Methods **9** (2012) 796–804

[9] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) **58**(1) (1995) 267–288

[10] Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective (with comments). Journal of the Royal Statistical Society, Series B **73**(3) (2011) 273–282

[11] van Someren, E., Vaes, V., Steegenga, W., Sijbers, A., Dechering, K.J., Reinders, M.: Least absolute regression network analysis of the murine osteoblast differentiation network. Bioinformatics **22**(4) (2005) 477–484

[12] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2) (2005) 301–320

[13] Ahmed, A., Xing, E.: Recovering time-varying networks of dependencies in social and biological studies. Proceedings of the National Academy of Sciences **106** (2009) 11878–11883

[14] Grzegorczyk, M., Husmeier, D.: A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology. Statistical Applications in Genetics and Molecular Biology (SAGMB) **11**(4) (2012) Article 7.
[15] Äijö, T., Lähdesmäki, H.: Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. Bioinformatics **25**(22) (2009) 2937–2944
[16] Wilkinson, D.: Stochastic modelling for quantitative description of heterogeneous biological systems. Nature Reviews Genetics **10**(2) (2009) 122–133
[17] Wilkinson, D.: Stochastic modelling for systems biology. Volume 44. CRC press (2011)
[18] Ciocchetta, F., Hillston, J.: Bio-PEPA: A framework for the modelling and analysis of biological systems. Theoretical Computer Science **410**(33) (2009) 3065–3084
[19] Gillespie, D.: Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry **81**(25) (1977) 2340–2361
[20] Flis, A., Fernandez, P., Zielinski, T., Sulpice, R., Pokhilko, A., McWatters, H., Millar, A., Stitt, M., Halliday, K.: Biological regulation identified by sharing timeseries data outside the 'omics. Submitted (2013)
[21] Edwards, K., Akman, O., Knox, K., Lumsden, P., Thomson, A., Brown, P., Pokhilko, A., Kozma-Bognar, L., Nagy, F., Rand, D., et al.: Quantitative analysis of regulatory flexibility under changing environmental conditions. Molecular systems biology **6**(1) (2010)
[22] Consortium: The TiMet Project - Linking the clock to metabolism: http://timing-metabolism.eu (October 2012)
[23] Barenco, M., Tomescu, D., Brewer, D., Callard, R., Stark, J., Hubank, M.: Ranked prediction of p53 targets using hidden variable dynamic modeling. Genome Biology **7**(R25) (2006)
[24] Solak, E., Murray-Smith, R., Leithead, W., Leith, D., Rasmussen, C.: Derivative observations in Gaussian process models of dynamic systems. In: Proceedings of Neural Information Processing Systems, MIT Press (2002)
[25] Gelman, A., Rubin, D.: Inference from iterative simulation using multiple sequences. Statistical Science **7** (1992) 457–472
[26] Grzegorczyk, M., Husmeier, D.: Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models. Machine Learning **91**(1) (2013) 105–154
[27] Hanley, J., McNeil, B.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**(1) (1982) 29–36
[28] Rasmussen, C., Neal, R., Hinton, G., van Camp, D., Revow, M., Ghahramani, Z., Kustra, R., Tibshirani, R.: The Delve manual. URL http://www. cs. toronto. edu/~ delve (1996)
[29] Brandt, S.: Data Analysis: Statistical and Computational Methods for Scientists and Engineers. Springer, New York, USA (1999)
[30] Vyshemirsky, V., Girolami, M.: Bayesian ranking of biochemical system models. Bioinformatics **24** (2008) 833–839