ASSESSMENT OF REGRESSION METHODS FOR INFERENCE OF REGULATORY NETWORKS INVOLVED IN CIRCADIAN REGULATION

Andrej Aderhold[1], *Dirk Husmeier*[2], *V. Anne Smith*[1], *Andrew J. Millar*[3], *and Marco Grzegorczyk*[4]

[1]School of Biology, University of St. Andrews, UK
[2] School of Mathematics and Statistics, University of Glasgow, UK
[3] SynthSys, University of Edinburgh, UK
[4] Department of Statistics, TU Dortmund University, Germany
Email: aa796@st-andrews.ac.uk, grzegorczyk@statistik.tu-dortmund.de

ABSTRACT

We assess the accuracy of three established regression methods for reconstructing gene and protein regulatory networks in the context of circadian regulation. Data are simulated from a recently published regulatory network of the circadian clock in *Arabidopsis thaliana*, in which protein and gene interactions are described by a Markov jump process based on Michaelis-Menten kinetics. We closely follow recent experimental protocols, including the entrainment of seedlings to different light-dark cycles and the knock-out of various key regulatory genes. Our study provides relative assessment scores for the comparison of state-of-the art regression methods, investigates the influence of systematically missing values related to unknown protein concentrations and mRNA transcription rates, and quantifies the dependence of the performance on the degree of recurrency.

1. INTRODUCTION

Plants have to carefully manage their resources. The process of photosynthesis allows them to utilize sunlight to produce essential carbohydrates during the day. However, the earth's rotation predictably removes sunlight, and hence the opportunity for photosynthesis, for a significant part of each day, and plants need to orchestrate the accumulation, utilisation and storage of photosynthetic products in the form of starch over the daily cycle to avoid periods of starvation, and thus optimise growth rates.

In the last few years, substantial progress has been made to model the central processes of circadian regulation, i.e. the mechanism of internal time-keeping that allows the plant to anticipate each new day, at the molecular level [1, 2]. Moreover, simple mechanistic models have been developed to describe the feedback between carbon metabolism and the circadian clock, by which the plant adjusts the rates of starch accumulation and consumption in response to changes in the light-dark cycle [3]. What is needed is the elucidation of the detailed structure of the molecular regulatory networks and signalling pathways of these processes, by utilization and integration of transcriptomic, proteomic and metabolic concentration profiles that become increasingly available from international research collaborations like Agrogenomics[1] and Timet[2].

The inference of molecular regulatory networks from postgenomic data has been a central topic in computational systems biology for over a decade. Following up on the seminal paper in [5], a variety of methods have been proposed [6], and several procedures have been pursued to objectively assess the network reconstruction accuracy [7, 8, 6]. The present study follows up on this work and extends it in four important respects. Firstly, to make the evaluation more targeted at the specific problem of inferring gene and protein interactions related to circadian regulation, we take the central circadian clock network in *Arabidopsis thaliana*, as published in [2] , as a ground truth for evaluation, and closely follow recent experimental protocols for data generation, including the entrainment of seedlings to different light-dark cycles, and the knock-out of various key regulatory genes. Secondly, to make the data generated from this network as realistic as possible, we model gene and protein interactions as a Markov jump process based on Michaelis-Menten kinetics. This is to be preferred over mechanistic models based on ordinary differential equations (used e.g. in [1]), as it captures the intrinsic stochasticity of molecular interactions. Thirdly, we assess the impact of missing values on the reconstruction task. Protein-gene interactions affect transcription rates, but both these rates as well as protein concentrations might not be available from the wetlab assays. In such situations, mRNA concentrations have to be taken as proxy for protein concentrations, and rates have to be approximated by finite difference quotients. For both approximations, we quantify the ensuing deterioration in network reconstruction accuracy. Fourthly and finally, we investigate the dependence of the network reconstruction accuracy on the degree of recurrency in the network. The central circadian clock network is densely connected with several tight feedback loops. However, we expect the regulatory network, via which the clock acts on carbon metabolism, to be sparser and with more feed-forward structures. In our study we therefore quan-

[1]https://agronomics.ethz.ch/
[2]http://timing-metabolism.eu/

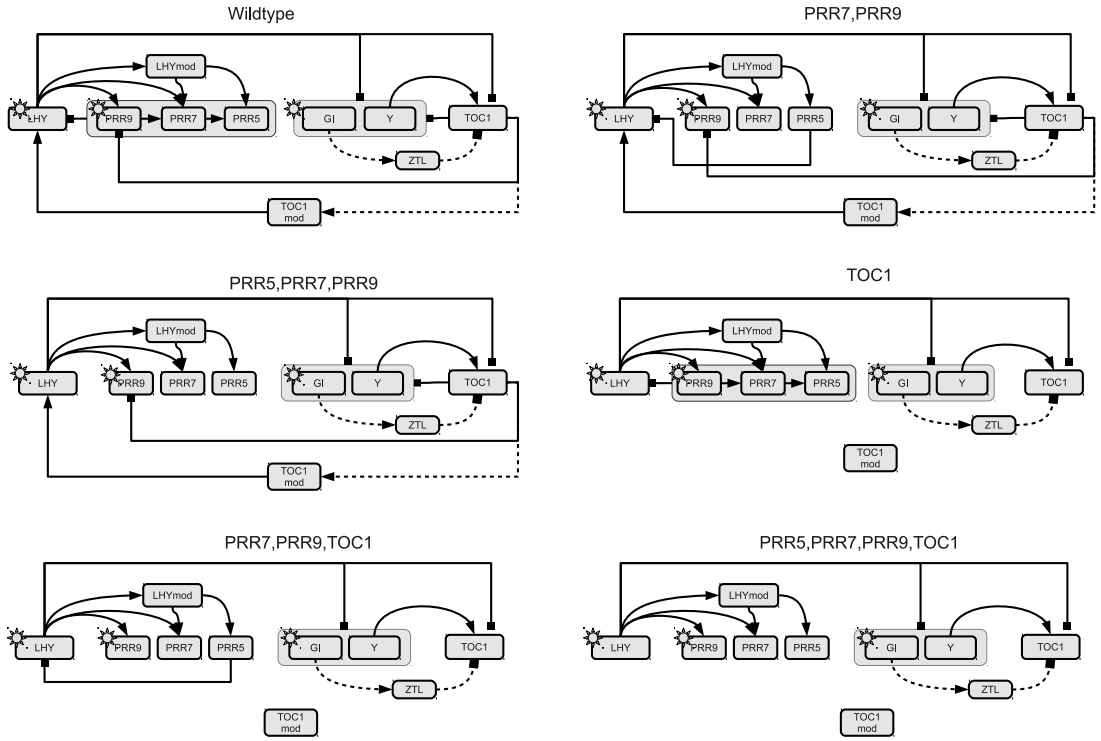

Figure 1. **Model network of the circadian clock in *Arabidopsis thaliana* based on [4] and subject to knock-outs.** Each graph shows interconnections of core circadian clock genes, where solid lines indicate protein influence on mRNA transcription, and dashed lines represent protein modifications. The top left panel shows the wildtye; the legends of the remaining panels indicate constant knock-outs of certain target proteins. Light influence is symbolized by a sun symbol, and grey boxes highlight set of regulators or regulated components.

tify how the network reconstruction depends on the degree of recurrency, and how the performance varies as critical feedback cycles are pruned.

2. METHOD OVERVIEW

2.1. Notation

Throughout the paper we use the following notation: For the regression models, which we will use to infer the network interactions, we have target variables y_g ($g = 1, \ldots, N$), each representing the temporal mRNA concentration gradient of a particular gene g. The realizations of each target variable y_g can then be written as a vector $\mathbf{y}_g = (y_{g,1}, \ldots, y_{g,T})^\mathsf{T}$, where $y_{g,t}$ is the realization of y_g in observation t. The potential covariates are either gene or protein concentrations, and the task is to infer a set of covariates $\boldsymbol{\pi}_g$ for each response variable y_g. The collective set of covariates $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N\}$ defines a regulatory interaction network, $\mathcal{M}$. In $\mathcal{M}$ the covariates and the target variables represent the nodes, and from each covariate in $\boldsymbol{\pi}_g$ a directed interaction (or "edge") is pointing to the target node g. The complete set of regulatory observations is contained in the design matrix $\mathbf{X}$. Realizations of the covariates in the set $\boldsymbol{\pi}_g$ are collected in $\mathbf{X}_{\pi_g}$, where the columns of $\mathbf{X}_{\pi_g}$ are the realizations of the covariates π_g. Design matrix $\mathbf{X}$ and $\mathbf{X}_{\pi_g}$ are extended by a constant element equal to 1 for the intercept.

2.2. Sparse regression

A widely applied linear regression method that encourages network sparsity is the Least Absolute Shrinkage and Selection Operator (Lasso) introduced in [9]. The Lasso optimizes the parameters of a linear model based on the residual sum of squares subject to an $L1$-norm penalty constraint on the regression parameters, $\|\mathbf{w}_g\|_1$, which excludes the intercept [10]:

$$\hat{\mathbf{w}}_g = \operatorname{argmin}\left\{ ||\mathbf{y}_g - \mathbf{X}^T \mathbf{w}_g||_2^2 + \lambda_1 \|\mathbf{w}_g\|_1 \right\} \quad (1)$$

where λ_1 is a regularisation parameter controlling the strength of shrinkage. Equation (1) constitutes a convex optimization problem, with a solution that tends to be sparse. Two disadvantages of the Lasso are arbitrary selection of single predictors from a group of highly correlation variables, and saturation at T predictor variables. To avoid these problems, the Elastic Net method was proposed in [11], which combines the Lasso penalty with a ridge regression penalty of the standard squared $L2$-norm $\|\mathbf{w}_g\|_2^2$ exluding the intercept:

$$\hat{\mathbf{w}}_g = \operatorname{argmin}\left\{ ||\mathbf{y}_g - \mathbf{X}^T \mathbf{w}_g||_2^2 + \lambda_1 \|\mathbf{w}_g\|_1 + \lambda_2 \|\mathbf{w}_g\|_2^2 \right\} \quad (2)$$

Like Equation (1), Equation (2) constitutes a convex optimization problem, which we solve with cyclical coordinate descent [10] implemented in the R software package *glmnet* . The regularization parameters λ_1 and λ_2 were optimized by 10-fold cross-validation.

2.3. Bayesian regression

In Bayesian regression we assume a linear regression model for the targets:

$$\mathbf{y}_g|(\mathbf{w}_g, \sigma_g, \pi_g) \sim \mathcal{N}(\mathbf{X}_{\pi_g}^{\mathsf{T}}\mathbf{w}_g, \sigma_g^2\mathbf{I}) \quad (3)$$

where σ_g^2 is the noise variance, and $\mathbf{w}_g$ is the vector of regression parameters, for which we impose a Gaussian prior:

$$\mathbf{w}_g|(\sigma_g, \delta_g, \pi_g) \sim \mathcal{N}(\mathbf{0}, \delta_g\sigma_g^2\mathbf{I}) \quad (4)$$

δ_g can be interpreted as a "signal-to-noise" hyperparameter [12]. For the posterior distribution we get:

$$\mathbf{w}_g|(\sigma_g, \delta_g, \pi_g, \mathbf{y}_g) \sim \mathcal{N}(\mathbf{\Sigma}_g\mathbf{X}_{\pi_g}\mathbf{y}_g, \sigma_g^2\mathbf{\Sigma}_g) \quad (5)$$

where $\mathbf{\Sigma}_g^{-1} = \delta_g^{-1}\mathbf{I} + \mathbf{X}_{\pi_g}\mathbf{X}_{\pi_g}^{\mathsf{T}}$, and the marginal likelihood can be obtained by application of standard results for Gaussian integrals [13]:

$$\mathbf{y}_g|(\sigma_g, \delta_g, \pi_g) \sim \mathcal{N}(\mathbf{0}, \sigma_g^2(\mathbf{I} + \delta_g\mathbf{X}_{\pi_g}^{\mathsf{T}}\mathbf{X}_{\pi_g})) \quad (6)$$

For σ_g^{-2} and δ_g^{-2} we impose conjugate gamma priors, $\sigma_g^{-2} \sim \mathrm{Gam}(\nu/2, \nu/2)$, and $\delta_g^{-1} \sim \mathrm{Gam}(\alpha_\delta, \beta_\delta)$.[3] The integral resulting from the marginalization over σ_g^{-2},

$$P(\mathbf{y}_g|\pi_g, \delta_g) = \int_0^\infty P(\mathbf{y}_g|\sigma_g, \delta_g, \pi_g)P(\sigma_g^{-2}|\nu)d\sigma_g^{-2}$$

is then a multivariate Student t-distribution with a closed-from solution (e.g. [13, 12]). Given the data for the potential covariates of y_g, symbolically $\mathcal{D}$, the objective is to infer the set of covariates π_g from the marginal posterior distribution:

$$P(\pi_g|\mathcal{D}, \mathbf{y}_g, \delta_g) = \frac{P(\pi_g)P(\mathbf{y}_g|\pi_g, \delta_g)}{\sum_{\pi_g^\star} P(\pi_g^\star)P(\mathbf{y}_g|\pi_g^\star, \delta_g)} \quad (7)$$

where the sum is over all valid covariate sets $\pi_g^\star$, $P(\pi_g)$ is a uniform distribution over all covariate sets subject to a maximal cardinality, $|\pi_g| \leq 3$, and δ_g is a nuisance parameter, which can be marginalized over. We sample sets of regulators (or covariates) π_g, signal-to-noise parameters δ_g, and noise variances σ_g^2 from the joint posterior distribution with Markov chain Monte Carlo (MCMC), following a Metropolis-Hastings within partially collapsed Gibbs scheme [12].

3. DATA

We generated data from the central circadian gene regulatory network in *Arabidopsis thaliana*, as proposed in [2] and depicted in the top right panel of Figure 1 . Following [14], the regulatory processes of transcriptional regulation and post-translational protein modification were described with a Markov jump process based on Michaelis-Menten kinetics, which defines how mRNA and protein concentrations change in dependence on the concentrations of other interacting components in the system (see appendix of [2] for detailed equations). We simulated mRNA and protein concentration time courses with the Gillespie algorithm [15], using the Bio-PEPA modelling framework [16]. To investigate the influence of recurrent interactions on the network reconstruction, we eliminated feedback loops successively via targeted downregulation of protein translation (knock-outs) and replacement of corresponding concentrations by white Gaussian noise. This gave us five modified network structures, as shown in Figure 1. For each network type we created 11 interventions in consistency with standard biological protocols (e.g. [17]). These include knock-outs of proteins 'GI', 'LHY', 'PRR7,PRR9', 'TOC1', and varying photoperiods of 4, 6, 8, 12, or 18 hours of light in a 24-hour light-dark (LD) cycle. For each intervention we simulated protein and mRNA concentration time courses over 6 days. The first 5 days served as entrainment to the indicated LD cycles. This was followed by a day of persistent darkness (DD) or light (LL), during which concentrations of mRNAs and proteins were measured in 2 hour intervals. Combining 13 observations for each intervention yielded 143 observations in total for each network type. All concentrations were standardized to unit standard deviation. The temporal mRNA concentration gradient was approximated by a difference quotient of mRNA concentrations based on two alternative temporal resolutions: at -2 and

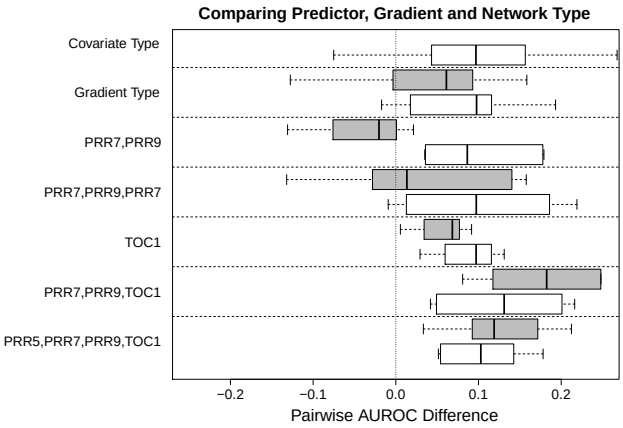

Figure 2. **Pairwise AUROC comparisons including all regression methods.** *Covariate type*: AUROC differences of protein against mRNA as covariates. *Gradient type*: AUROC difference of fine against coarse gradient for mRNAs (grey box) and proteins (white box). The remaining panels show the AUROC differences of various pruned networks as displayed in Figure 1 versus wild-type network for mRNA (grey boxes) and proteins (white boxes) as covariates.

[3] We set: $\nu = 0.01$, $\alpha_\delta = 2$, and $\beta_\delta = 0.2$, as in [12].

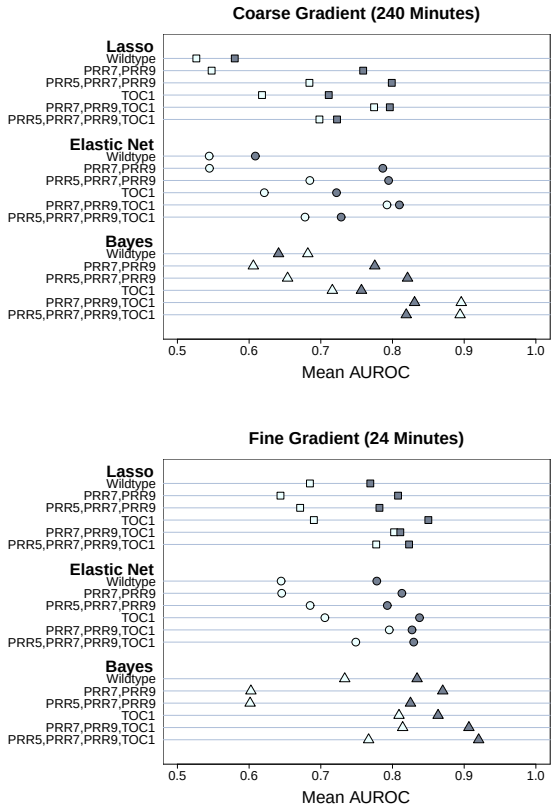

Figure 3. **Method comparison.** Mean AUROC values as a measure of network reconstruction performance for Lasso, Elastic Nets and Bayesian Regression applied to 6 distinct data types generated from networks shown in Figure 1. Empty symbols correspond to mRNA covariates, filled symbols to protein covariates.

+2 hours (coarse gradient), and at -12 and +12 minutes (fine gradient), followed by z-score standardization.

When trying to reconstruct the regulatory network from the simulated data, we ruled out self-loops, such as from LHY (modified) protein to LHY mRNA, and adjusted for mRNA degradation by enforcing mRNA self-loops, such as from the LHY mRNA back to itself. Protein 'ZTL' was included in the stochastic simulations, but excluded from structure learning because it has no direct effect on transcription. We carried out two different network reconstruction tasks. The first was based on complete observation, including both protein and mRNA concentration time series. The second was based on incomplete observation, where only mRNA concentrations were available, but protein concentrations were systematicaly missing. All network reconstructions were repeated on five independent data instantiations.

4. RESULTS

For Bayesian regression, we compute the marginal posterior probabilities of all potential interactions. For Lasso and Elastic Nets, we record the absolute values of non-zero regression parameters. Both measures provide a means by which interactions between genes and proteins

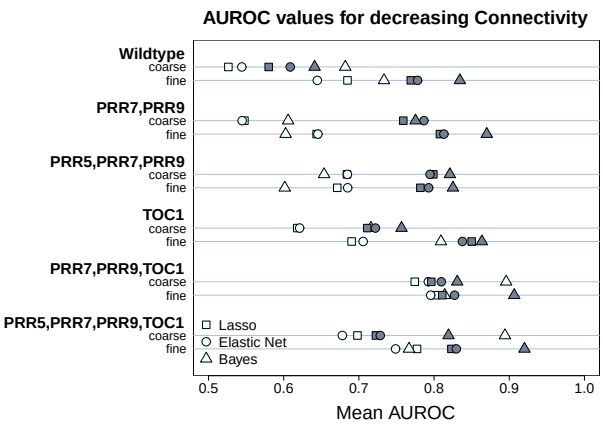

Figure 4. **Performance comparison in dependence on network connectivity.** Mean AUROC values for sparser networks in descending order for coarse (4 hour) and fine (24 minutes) response gradients. Empty symbols correspond to mRNA covariates, filled symbols to protein covariates.

can be ranked in terms of their significance or influence. Given that the true network is known, this ranking defines the Receiver Operating Characteristic (ROC) curve, where 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. The results of our study are shown in Figures 2, 3 and 4 and can be summarized as follows.

Comparison between the methods. The performance of Lasso and Elastic Net is very similar, while Bayesian regression achieves slightly better results, especially when protein concentrations are included and temporal gradients are computed at fine resolution. This indicates a justification of the higher computational costs of inference based on MCMC.

Influence of gradient estimation. A finer temporal resolution for the gradient estimation tends to improve the network reconstruction. This affects in particular data for which the reconstruction based on coarse gradients is close to random expectation, and Bayesian regression models applied to protein concentrations. However, the coarse gradient leads to a noticeable improvement in the reconstruction with Bayesian regression based on mRNA profiles alone for the sparser networks in Figure 1. Preliminary investigations indicate that this unexpected trend is related to confounding correlations between the profiles of the two LHY isoforms, which are more important (propulsive) regulators in the sparser networks. However, a closer analysis is still required.

Influence of missing protein concentrations. With the noticeable exception of Bayesian regression applied to

data with coarse gradient estimation, discussed above, the inclusion of protein concentrations significantly improves the network reconstruction accuracy. Our study allows a quantification of the degree of improvement in terms of AUROC score differences, with a mean improvement of 0.09 and a p-value of 8e-06 (from a two a sided t-test) indicating significant higher values using protein covariates (Figure 2).

Influence of feedback loops. An important aspect of our study is the investigation of how the network reconstruction accuracy depends on the connectivity of the true network and the proportion of recurrent connections. To this end we have successively pruned feedback interactions, as shown in Figure 1. Figures 2 and 4 suggest that there is a noticeable trend, with less recurrent networks appearing to be easier to learn.

5. CONCLUSION

We have carried out a comparative evaluation of three established machine learning methods for regulatory network reconstruction (Lasso, Elastic Nets, Bayesian regression) based on the central gene regulatory network of the circadian clock in *Arabidopsis thaliana*, and a series of synthetic gene knock-outs that affect the proportion of recurrent interactions. Our study allows a quantification of the improvement in network reconstruction accuracy as a consequence of including protein concentrations, the dependence of the performance on the recurrent network connectivity, and the influence of the numerical approximation of the gradient (i.e. transcription rates) by finite-size difference quotients.

6. ACKNOWLEDGMENTS

This work was funded under EU FP7 project "Timet". M.G. is supported by the German Research Foundation (DFG), research grant GR3853/1-1. A.A. is supported by the BBSRC. SynthSys is a Centre for Integrative and Systems Biology partly funded by BBSRC and EPSRC award (BB/D019621).

7. REFERENCES

[1] A. Pokhilko, A. Fernández, K. Edwards, M. Southern, K. Halliday, and A. Millar, "The clock gene circuit in arabidopsis includes a repressilator with additional feedback loops," *Molecular systems biology* 8, 574, 2012.

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

[3] F. Feugier and A. Satake, "Dynamical feedback between circadian clock and sucrose availability explains adaptive response of starch metabolism to various photoperiods," *Frontiers in Plant Science*, 3, 2012.

[4] A. Pokhilko et al., "Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model," *Molecular systems biology*, 6 (1), 2010.

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

[6] M. T. Weirauch et al., "Evaluation of methods for modeling transcription factor sequence specificity," *Nature Biotechnology*, 2013.

[7] D. Husmeier, "Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks," *Bioinformatics*, 19, 2271–2282, 2003.

[8] A. V. Werhli, M. Grzegorczyk, and D. Husmeier, "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks," *Bioinformatics*, 22, 2523–2531, 2006.

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

[10] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 33 (1), 1–22, 2010.

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

[12] M. Grzegorczyk and D. Husmeier, "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*, 11 (4), 2012a, Article 7.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Singapore, 2006.

[14] D. Wilkinson, *Stochastic modelling for systems biology*, 44, CRC Press, 2011.

[15] D. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, 81 (25), 2340–2361, 1977.

[16] F. Ciocchetta and J. Hillston, "Bio-pepa: A framework for the modelling and analysis of biological systems," *Theoretical Computer Science*, 410 (33), 3065–3084, 2009.

[17] K. Edwards et al., "Quantitative analysis of regulatory flexibility under changing environmental conditions," *Molecular Systems Biology*, 6 (1), 2010.