Supplementary material

for the paper:

'Bayesian regularization of non-homogeneous dynamic Bayesian networks by globally coupling interaction parameters'

(*AISTATS* 2012)

This paper is a supplement to the main paper 'Bayesian regularization of non-homogeneous dynamic Bayesian networks by globally coupling interaction parameters', which appears in Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012). Volume 22 of JMLR: W&CP 22.
Table 1 summarizes the details of the four *A. thaliana* gene expression time series from Subsection 4.2. In Appendix I we derive equations (17-18) from Subsection 3.1 of the main paper. In Appendix II we provide the mathematical details of the point process prior and more detailed information on our convergence diagnostics and the employed network reconstruction criterions. Finally, in Appendix III we demonstrate that our findings on the network reconstruction accuracies can also be produced when the are under the precision recall curve rather than the area under the ROC curve is employed for evaluation.

	Segment 1	Segment 2	Segment 3	Segment 4
Source	Mockler et al.(2007)	Edwards et al. (2006)	Grzegorcyk et al. (2008)	Grzegorcyk et al. (2008)
Time points	12	13	13	13
Time interval	4h	4h	2h	2h
Pretreatment entrainment	12h:12h light:dark cycle	12h:12h light:dark cycle	10h:10h light:dark cycle	14h:14h light:dark cycle
Measurements	Constant light	Constant light	Constant light	Constant light
Laboratory	Kay Lab	Millar Lab	Millar Lab	Millar Lab

Table 1: **Gene expression time series segments for Arabidopsis.** The table contains an overview of the experimental conditions under which each of the gene expression experiments was carried out. See Subsection 4.2 of the main paper for further details.

Appendix I - Derivations of equations

Equation (17) in Subsection 3.1 of the main paper can be derived as follows:
For updating the signal-to-noise ratio hyperparameters $\{\delta_g\}$, $1 \leq g \leq N$, with a Gibbs sampling scheme note that

$$
\begin{aligned}
&P(\mathbf{y}_{g,.}, \mathbf{w}_{g,.}, \boldsymbol{\sigma}^2_{g,.}, \delta_g | \mathbf{X}_{\pi_g,.}) \\
&= P(\delta_g^{-1}|\alpha_\delta, \beta_\delta) \prod_{h=1}^{K_g} P(\sigma_{g,h}^{-2}|\alpha_\sigma, \beta_\sigma) P(\mathbf{y}_{g,h}|\mathbf{w}_{g,h}, \sigma^2_{g,h}, \mathbf{X}_{\pi_g,h}) P(\mathbf{w}_{g,h}|\sigma^2_{g,h}, \delta_g, \mathbf{m}_g) \\
&= \mathrm{Gam}(\delta_g^{-1}|\alpha_\delta, \beta_\delta) \prod_{h=1}^{K_g} \mathrm{Gam}(\sigma_{g,h}^{-2}|\alpha_\sigma, \beta_\sigma) \mathcal{N}(\mathbf{y}_{g,h}|\mathbf{X}^\mathsf{T}_{\pi_g,h}\mathbf{w}_{g,h}, \sigma^2_{g,h}\boldsymbol{\Sigma}_{g,h}) \mathcal{N}(\mathbf{w}_{g,h}|\mathbf{m}_g, \delta_g\sigma^2_{g,h}\mathbf{C}_{g,h}) \\
&\propto [\delta_g^{-1}]^{\alpha_\delta - 1} \exp(-\beta_\delta \delta_g^{-1}) \prod_{h=1}^{K_g} [\sigma_{g,h}^{-2}]^{\alpha_\sigma - 1} \exp(-\beta_\sigma \sigma_{g,h}^{-2}) \\
&\cdot \frac{1}{\sqrt{(2\pi\delta_g\sigma^2_{g,h})^{k_g}|\mathbf{C}_{g,h}|}} \exp\left(-\frac{1}{2\delta_g\sigma^2_{g,h}}[\mathbf{w}_{g,h} - \mathbf{m}_g]^\mathsf{T}\mathbf{C}^{-1}_{g,h}[\mathbf{w}_{g,h} - \mathbf{m}_g]\right) \\
&\cdot \frac{1}{\sqrt{(2\pi\sigma^2_{g,h})^{T_{g,h}}|\boldsymbol{\Sigma}_{g,h}|}} \exp\left(-\frac{1}{2\sigma^2_{g,h}}[\mathbf{y}_{g,h} - \mathbf{X}^\mathsf{T}_{\pi_g,h}\mathbf{w}_{g,h}]^\mathsf{T}\boldsymbol{\Sigma}^{-1}_{g,h}[\mathbf{y}_{g,h} - \mathbf{X}^\mathsf{T}_{\pi_g,h}\mathbf{w}_{g,h}]\right)
\end{aligned}
$$

where $\mathbf{y}_{g,.} = (\mathbf{y}^\mathsf{T}_{g,1}, \ldots, \mathbf{y}^\mathsf{T}_{g,K_g})^\mathsf{T}$, $\mathbf{w}_{g,.} = (\mathbf{w}^\mathsf{T}_{g,1}, \ldots, \mathbf{w}^\mathsf{T}_{g,K_g})^\mathsf{T}$, $\boldsymbol{\sigma}^2_{g,.} = (\sigma^2_{g,1}, \ldots, \sigma^2_{g,K_g})$, $\mathbf{X}_{\pi_g} = (\mathbf{X}_{\pi_g,1}, \ldots, \mathbf{X}_{\pi_g,K_g})$, k_g is the cardinality of $\boldsymbol{\pi}_g$, and $|.|$ denotes the determinant of a matrix.
On collecting all the terms that depend on δ_g^{-1} and normalization this gives:

$$
\begin{aligned}
&P(\delta_g^{-1}|\mathbf{y}_{g,.}, \mathbf{w}_{g,.}, \sigma^2_{g,.}, \mathbf{X}_{\pi_g,.}) = \\
&\mathrm{Gam}\left(\delta_g^{-1}|\alpha_\delta + \frac{K_g k_g}{2}, \beta_\delta + \frac{1}{2}\sum_{h=1}^{K_g} \frac{1}{\sigma^2_{g,h}}[\mathbf{w}_{g,h} - \mathbf{m}_g]^\mathsf{T}\mathbf{C}^{-1}_{g,h}[\mathbf{w}_{g,h} - \mathbf{m}_g]\right)
\end{aligned}
$$

where K_g is the number of segments for node g.

Equation (18) in Subsection 3.1 of the main paper can be derived as follows:
For the inverse variances $\sigma^{-2}_{g,h}$ we use a collapsed Gibbs sampler in which the interaction parameters $\mathbf{w}_{g,h}$ have been integrated out. From equations (12-13) of the main paper we obtain:

$$
\begin{aligned}
&P(\mathbf{y}_{g,h}, \sigma^{-2}_{g,h}|\mathbf{X}_{\pi_g,h}, \sigma^2_{g,h}, \delta_g) = \mathcal{N}(\mathbf{y}_{g,h}|\tilde{\mathbf{m}}_{g,h}, \sigma^2_{g,h}\tilde{\boldsymbol{\Sigma}}_{g,h}) \mathrm{Gam}(\sigma^{-2}_{g,h}|\alpha_\sigma, \beta_\sigma) \\
&\propto \frac{1}{\sqrt{\sigma^{2T_{g,h}}_{g,h}|\tilde{\boldsymbol{\Sigma}}_{g,h}|}} \exp\left(-\frac{\Delta^2_{g,h}}{2\sigma^2_{g,h}}\right) [\sigma^{-2}_{g,h}]^{\alpha_\sigma - 1} \exp(-\beta_\sigma \sigma^{-2}_{g,h}) \\
&\propto [\sigma^{-2}_{g,h}]^{\frac{T_{g,h}}{2} + \alpha_\sigma - 1} \exp\left(-\left[\frac{\Delta^2_{g,h}}{2} + \beta_\sigma\right]\sigma^{-2}_{g,h}\right)
\end{aligned}
$$

where $\Delta^2_{g,h}$ was defined in equation (16) of the main paper. By normalization we get:

$$
P(\sigma^{-2}_{g,h}|\mathbf{y}_{g,h}, \mathbf{X}_{\pi_g,h}, \delta_g) \;=\; \mathrm{Gam}\left(\sigma^{-2}_{g,h}|\frac{T_{g,h}}{2} + \alpha_\sigma, \frac{\Delta^2_{g,h}}{2} + \beta_\sigma\right)
$$

Appendix II - Mathematical details

The point process prior for the changepoint sets $\boldsymbol{\tau}_g$

In Section 3.2 of the main paper we did not specify the prior $P(\{\boldsymbol{\tau}_g\})$ on the gene-specific changepoint sets $\boldsymbol{\tau}_g = \{\tau_{g,1}, \ldots, \tau_{g,K_g-1}\}$ $(g = 1, \ldots, N)$ explicitly. Here, we provide the details: We assume that the gene-specific changepoint sets $\boldsymbol{\tau}_g$ are independently distributed $P(\{\boldsymbol{\tau}_g\}) = \prod_{g=1}^{N} P(\boldsymbol{\tau}_g)$, and we follow Fearnhead (2006) and employ a point process prior to model the distances between successive changepoints for each gene g $(g = 1, \ldots, N)$. In the point process model $s(t)$ $(t = 1, 2, 3, \ldots)$ denotes the prior probability that there are t time points between two successive changepoints $\tau_{g,j-1}$ and $\tau_{g,j}$ on the discrete interval $\{2, \ldots, T-1\}$. The prior probability $P(\boldsymbol{\tau}_g)$ of the changepoint set of gene g, $\boldsymbol{\tau}_g = \{\tau_{g,1}, \ldots, \tau_{g,K_g-1}\}$, containing $K_g - 1$ changepoints $\tau_{g,j}$ $(j = 1, \ldots, K_g - 1)$ with $1 < \tau_{g,j-1} < \tau_{g,j} < T$ $(j = 2, \ldots, K_g - 1)$, is:

$$P(\boldsymbol{\tau}_g) = P(\tau_{g,1}, \ldots, \tau_{g,K_g-1}) = s_0(\tau_{g,1}) \left(\prod_{j=2}^{K_g-1} s(\tau_{g,j} - \tau_{n,j-1}) \right) (1 - S(\tau_{g,K_g} - \tau_{g,K_g-1})) \tag{1}$$

where $\tau_{g,0} = 1$ and $\tau_{g,K_g} = T$ are two pseudo change-points, $s_0(.)$ is the prior distribution of the first changepoint $\tau_{g,1}$, and

$$S(t) = \sum_{s=1}^{t} s(t); \quad S_0(t) = \sum_{s=1}^{t} s_0(t) \tag{2}$$

are the cumulative distribution functions corresponding to $s(.)$ and $s_0(.)$. For $s(.)$ we follow Fearnhead (2006) and use the probability mass function of the negative binomial distribution[1] NBIN(p,k) with hyperparameters p and k:

$$s(t) = \left(\begin{array}{c} t-1 \\ k-1 \end{array} \right) p^k (1-p)^{t-k} \tag{3}$$

In a point process model on the positive *and* negative integers the probability mass function of the first changepoint $\tau_{g,1} \in \{2, \ldots, T-1\}$ is a mixture of k negative binomial distributions:

$$s_0(\tau_{g,1}) = \frac{1}{k} \sum_{i=1}^{k} \left(\begin{array}{c} (\tau_{g,1} - 1) - 1 \\ i - 1 \end{array} \right) p^i (1-p)^{(\tau_{g,1}-1)-i} \tag{4}$$

For our analysis of the in vivo data from *S. cerevisiae* in Section 6 of the main paper we used a fixed value for k $(k = 1)$ and we varied the hyperparameter p. In the first instance, we started with six different values: $p \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ that cover the whole area from zero changepoints (per gene) to about 8 changepoints per gene. Then, to shed more light onto the more interesting area with fewer changepoints per gene, we additionally employed: $p \in \{0.001, 0.01, 0.02, 0.025, 0.03, 0.04, 0.05, 0.075, 0.15\}$.

MCMC sampling and convergence diagnostics

For all MCMC simulations we set the lengths of the burn-in and the sampling phase to 50,000 MCMC iterations each. For the simulated network data, described in Subsection 4.1 of the main paper, and the expression data from *Arabidopsis thaliana*, described in Subsection 4.2 of the main paper, we assumed that the true segmentations (i.e. the true changepoints) are known ("supervised approach"). In each single MCMC iteration for *all* genes g $(g = 1, \ldots, N)$ the gene-specific variance σ_g^2 and the signal-to-noise hyperparameter δ_g are re-sampled, and the performance of one basic operation on the parent node set $\boldsymbol{\pi}_g$ is proposed. The move type is randomly chosen

[1] Note that the negative binomial distribution can be seen as a discrete version of the Gamma distribution.

from three basic operations on $\boldsymbol{\pi}_g$: (i) adding a new parent node to $\boldsymbol{\pi}_g$, (ii) deleting one of the parent nodes from $\boldsymbol{\pi}_g$, and (iii) substituting a parent node from $\boldsymbol{\pi}_g$ for another one, as proposed in Grzegorczyk and Husmeier (2011). During the sampling phase of length 50,000 we sampled every 500 iterations to obtain a network sample of size 100. From this sample of networks the marginal edge posterior probabilities of the individual edges can be computed.
For the gene expression time series in *S. cerevisiae*, described in Subsection 4.3 of the main paper, we assumed that the gene-specific changepoint sets $\boldsymbol{\tau}_g$ are unknown. Therefore, in each single MCMC step, for *all* genes g we also performed one move on the changepoint set $\boldsymbol{\tau}_g$. The move type is randomly chosen from three basic operations: (i) a changepoint birth adds a new changepoint to $\boldsymbol{\tau}_g$, (ii) a changepoint death move removes a changepoint from $\boldsymbol{\tau}_g$, and (iii) a changepoint re-allocation move substitutes a changepoint from $\boldsymbol{\tau}_g$ for another one.
We applied the standard diagnostic based on trace plots (Giudici and Castelo (2003))and potential scale reduction factors (Gelman and Rubin (1992)) to determine appropriate MCMC simulation lengths. In particular for the real data from *S. cerevisiae* we started five MCMC simulations from different network $\mathcal{M}$ and changepoint set $\boldsymbol{\tau}_g$ $(g = 1, \ldots, N)$ initializations for half a dozen point process hyperparameters $p = 0, 0.1, 0.2 \ldots, 0.5$ in equation (3) to assess convergence. MCMC convergence was monitored in terms of the potential scale reduction factors (PSRFs) based on the marginal edge posterior probabilities. For the above mentioned MCMC run lengths we observed a sufficient degree of convergence ($PSFR < 1.1$ for *all* individual edges). Because of the computational costs this convergence diagnostic could not be determined for every MCMC simulation that was employed in our study. However, we assume that the MCMC simulations with $p = 0, 0.1, 0.2 \ldots, 0.5$ are representative, and since we consistently observed a sufficient degree of convergence, according to the above mentioned criterion, we concluded that the run lengths also ensure convergence for other hyperparameters p. In Section 6 of the main paper we consistently report results of MCMC simulations that were seeded with empty parent sets ($\boldsymbol{\pi}_g = \emptyset$ for all g), empty changepoint sets ($\boldsymbol{\tau}_g = \emptyset$ for all g) and the following hyperparameters: $\sigma_g^2 = 1$ and $\delta_g = 1$ for all g. Details on further hyperparameter settings can be found in Section 5 of the main paper.

Network reconstruction accuracy

The network reconstruction accuracy can be evaluated as follows: Let $\mathcal{M}^{\diamond}(n, j) = 1$ indicate that the true graph possesses the edge $X_n \rightarrow X_j$, while $\mathcal{M}^{\diamond}(n, j) = 0$ indicates that there is no edge from X_n to X_j. For both Bayesian network models inference via RJMCMC sampling yields a marginal edge posterior probability $e_{n,j}$ for every edge $\mathcal{M}^{\diamond}(n, j)$. For $\zeta \in [0, 1]$ we define $E(\zeta) := \{\mathcal{M}(n, j) | e_{n,j} \geq \zeta\}$ as the set of all edges $\mathcal{M}(n, j)$ whose posterior probabilities exceed the threshold ζ. Since the true edges are known, for each $E(\zeta)$ the number of true positive $TP[\zeta]$, false positive $FP[\zeta]$, true negative $TN[\zeta]$, and false negative $FN[\zeta]$ edges can be counted. From this we can compute the true positive rate $TPR[\zeta] = TP[\zeta]/(TP[\zeta] + FN[\zeta])$ (also called *recall* or *sensitivity*), the false positive rate $FPR[\zeta] = FP[\zeta]/(TN[\zeta] + FP[\zeta])$, and the precision $PRE[\zeta] = TP[\zeta]/(TP[\zeta] + FP[\zeta])$. Plotting the $TPR[\zeta]$ values (vertical axis) against the corresponding $FPR[\zeta]$ values (horizontal axis) and connecting neighboring points by linear interpolation gives the receiver operator characteristic (ROC) curve. The area under the ROC curve (AUC or AUC-ROC) is a quantitative measure that can be obtained by numerically integrating the ROC curve on the interval $[0, 1]$; larger AUC values indicate a better network reconstruction accuracy, whereby 1 indicates perfect prediction, whereas 0.5 corresponds to a random expectation. An alternative score of the network reconstruction accuracy can be obtained by numerically integrating the Precision-Recall (PR) curve, so as to obtain the area under the PR curve (AUC or AUC-PR). PR curves can be obtained as follows: (i) The $PRE[\zeta]$ values (vertical axis) are plotted against the corresponding $TPR[\zeta]$ values (horizontal axis). (ii) Different from ROC curves, neighboring points cannot be connected by straight lines and

a nonlinear interpolation is required[2]. In our implementation we use the interpolation scheme described in Davis and Goadrich (2006). (iii) As the precision is not defined for TP=0 and FP=0 (PRE=0/0), we integrate the PR curve on the interval $[(1/E), 1]$ where E is the number of edges of the true graph $\mathcal{M}^{\diamond}$; i.e. we restrict on the area where at least one of the true edges has been learned.

Appendix III - An alternative evaluation criterion: The area under the precision recall curve

In this third appendix, we show that all our findings on the network reconstruction accuracies can be reproduced with an alternative network reconstruction accuracy criterion. Two scores have been established for evaluating the network reconstruction accuracy in systems biology research. The area under the receiver operator characteristic curve (AUC-ROC) and the area under the precision recall curve (AUC-PR). A comparison of these two criteria can be found in Davis and Goadrich (2006). These two criteria have been used for evaluating the results of the regularly held DREAM network reconstruction challenge (e.g. see Prill *et al.* (2010)). Therefore, we evaluated our network inference results with both criteria independently, and we found that both criteria yield very similar results. Figures 1 and 2 of this supplementary paper show the network reconstruction accuracy in terms of areas under the precision recall curves (AUC-PR); Figure 1 corresponds to Figure 1b) of the main paper, and Figure 2 corresponds to Figure 3b) of the main paper.

[2]The linear interpolation has to be done in terms of the true positives (TPs) and false positives (FPs) which corresponds to a nonlinear interpolation in the precision recall representation.

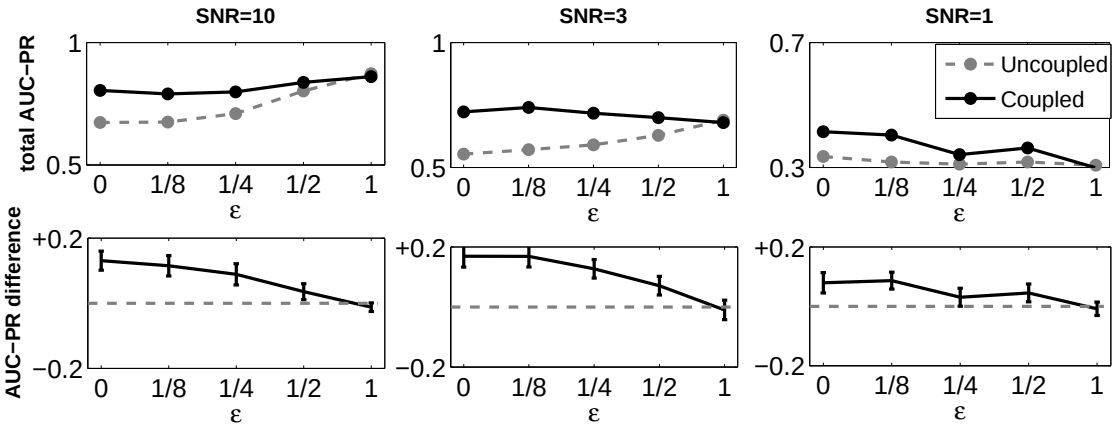

Figure 1: **Network reconstruction accuracy in terms of areas under the precision recall curve for the RAF pathway.** This figure corresponds to Figure 1b) of the main paper. The RAF pathway topology, as reported in Sachs *et al.* (2005), can be found in Figure 1a) of the main paper. Different from Figure 1b) of the main paper, the graphs monitor the network reconstruction accuracy in terms of *areas under the precison recall curve* (AUC-PR scores) for the conventional uncoupled (dotted gray lines) and the proposed coupled (solid black lines) non-homogeneous dynamic Bayesian network (NH-DBN) model. The graphs demonstrate how the proposed Bayesian regularization scheme is affected by increasing violations of the prior assumption inherent in equation (10) of the main paper. Simulated data were generated as described in Subsection 4.1 of the main paper. The abscissa represents the amplitude ε by which the global parameter vector is perturbed. The columns represent three different signal-to-noise (SNR) levels 10, 3, and 1. The top row shows the absolute values of the AUC-PR scores, while the bottom row shows the differences between the proposed regularization scheme (coupled NH-DBN) and the conventional unregularized method (uncoupled NH-DBN). All MCMC simulations were repeated on 25 independent data instantiations, with error bars indicating two-sided 95% confidence intervals .

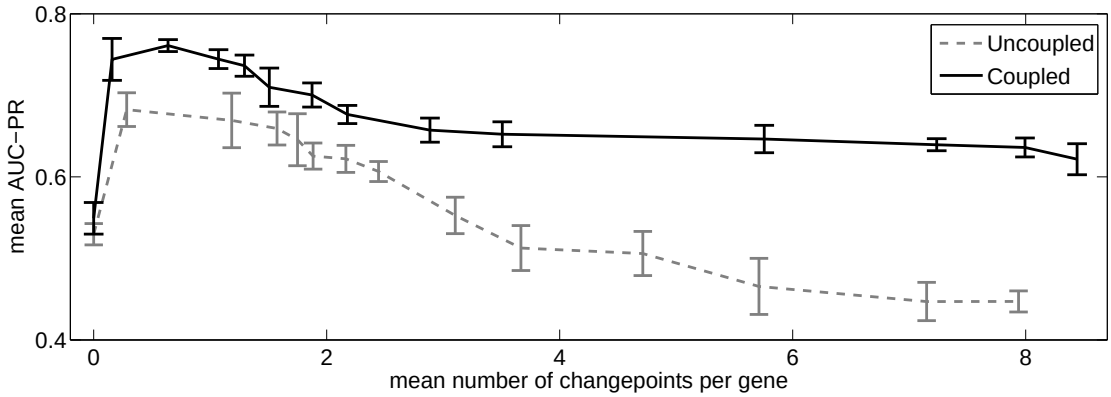

Figure 2: **Network reconstruction accuracy in terms of areas under the precision recall curve for the *S. cerevisiae* network.** This figure corresponds to Figure 3b) of the main paper. Cantone *et al.* (2009) designed the network shown in panel Figure 3a) of the main paper and measured in vivo expression levels for the five genes; see Subsection 4.3 of the main paper for details. The graph shows the network reconstruction accuracy (ordinate) plotted against the mean number of changepoints per gene (abscissa) for the conventional uncoupled method (dashed line) and the proposed Bayesian coupling scheme (solid line). Different from Figure 3b) in the main paper, the network reconstruction accuracy (ordinate) is quantified in terms of *mean areas under the precison recall curves* (AUC-PR scores), averaged over 5 MCMC simulations, with the vertical bars indicating standard errors.

References

Cantone, I., Marucci, L., Iorio, F., Ricci, M. A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D. and Cosma1, M. P. (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*, **137**, 172–181.

Davis, J. and Goadrich, M. (2006) The relationship between precision-recall and ROC curves. In *ICML '06: Proceedings of the 23rd international conference on Machine Learning*, pp. 233–240. ACM, New York, NY, USA.

Fearnhead, P. (2006) Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, **16**, 203–213.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Giudici, P. and Castelo, R. (2003) Improving Markov chain Monte Carlo model search for data mining. *Machine Learning*, **50**, 127–158.

Grzegorczyk, M. and Husmeier, D. (2011) Non-homogeneous dynamic Bayesian networks for continuous data. *Machine Learning*, **83**, 355–419.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G. and Stolovitzky, G. (2010) Towards a rigorous assessment of systems biology models: The dream3 challenges. *PLoS ONE*, **5**, e9202.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. and Nolan, G. P. (2005) Protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.