BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 3 2014, pages 414–419
doi:10.1093/bioinformatics/btt696

Systems biology

Advance Access publication November 30, 2013

Exact likelihood computation in Boolean networks with probabilistic time delays, and its application in signal network reconstruction

Sebastian Dümcke[1,2,*,†], Johannes Bräuer[3,†], Benedict Anchang[4,5], Rainer Spang[4], Niko Beerenwinkel[6] and Achim Tresch[1,2,*]

[1]Institute for Genetics, University of Cologne, 50674 Cologne, [2]Max Planck Institute for Plant Breeding Research, 50829 Cologne, [3]Gene Center, Department of Biochemistry, Ludwig-Maximilians University, 81379 Munich, Germany, [4]Insitute for Functional Genomics, 93053 Regensburg, Germany, [5]Department of Radiology, Center for Cancer Systems Biology, Stanford University, Stanford, CA 94305-5488, USA and [6]ETH Zürich, Department of Biosystems Science and Engineering, Mattenstrasse 26 4058 Basel, Switzerland

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: For biological pathways, it is common to measure a gene expression time series after various knockdowns of genes that are putatively involved in the process of interest. These interventional time-resolved data are most suitable for the elucidation of dynamic causal relationships in signaling networks. Even with this kind of data it is still a major and largely unsolved challenge to infer the topology and interaction logic of the underlying regulatory network.

Results: In this work, we present a novel model-based approach involving Boolean networks to reconstruct small to medium-sized regulatory networks. In particular, we solve the problem of exact likelihood computation in Boolean networks with probabilistic exponential time delays. Simulations demonstrate the high accuracy of our approach. We apply our method to data of Ivanova *et al.* (2006), where RNA interference knockdown experiments were used to build a network of the key regulatory genes governing mouse stem cell maintenance and differentiation. In contrast to previous analyses of that data set, our method can identify feedback loops and provides new insights into the interplay of some master regulators in embryonic stem cell development.

Availability and implementation: The algorithm is implemented in the statistical language R. Code and documentation are available at *Bioinformatics* online.

Contact: duemcke@mpipz.mpg.de or tresch@mpipz.mpg.de

Supplementary information: Supplementary Materials are available at *Bioinfomatics* online.

Received on June 24, 2013; revised on November 7, 2013; accepted on November 25, 2013

1 INTRODUCTION

The inference of signaling networks from biological data is of fundamental importance for a systemic understanding of regulatory processes. The statistical methods that have been developed for that purpose can be grouped according to the type of data that they expect as input. Many approaches use gene expression data. Some methods are based solely on static observations of the unperturbed system; they exploit the fact that fluctuations of interacting components are dependent (Basso *et al.*, 2005; Van Driessche *et al.*, 2005). The use of perturbation data greatly improves network reconstruction (Fröhlich *et al.*, 2008; Niederberger *et al.*, 2012; Tresch *et al.*, 2008). To resolve the order of events in a signaling cascade, time-resolved measurements after perturbation yield further improvements (Friedman *et al.*, 2000; Grzegorczyk *et al.*, 2008). Boolean networks are an appropriate tool for dealing with this type of data (Ideker *et al.*, 2000; Kauffman, 1969; Shmulevich *et al.*, 2002; Silvescu and Honavar, 2001). The most difficult problem lies in accounting for the mostly unknown time delays with which the signal is propagated through the network (Papin *et al.*, 2005).

In this work, we propose Boolean networks with probabilistic time delays as a novel statistical network inference method. There have been attempts to calculate the likelihood of a Boolean network in special cases by using Markov Chain Monte Carlo (MCMC) sampling (Anchang *et al.*, 2009) and for dynamic nested effects models (Failmezger *et al.*, 2013; Fröhlich *et al.*, 2011). Exact results were so far obtained only under strong restrictions on the logic functions involved, as in the context of conjunctive Bayesian networks (Beerenwinkel and Sullivant, 2009; Beerenwinkel *et al.*, 2007). By analytically marginalizing over the unknown delay times, we derive our main result, an exact and efficient recursive likelihood formula for a broad class of Boolean networks with exponentially distributed time delays that may include feedback loops. We evaluate our method in various simulation scenarios for its ability to recover the unknown topology. The method is then applied to a murine stem cell knockdown data set by Ivanova *et al.* (2006), which consists of a set of whole genome gene expression time series after the knockout of six genes (Essrb, Sox2, Nanog, Tcl1, Oct4 and Tbx3) that are considered key regulators in the maintenance and differentiation of mouse embryonic stem cells. Our analysis reveals more feedback loops than previously detected.

2 METHODS

We aim to model central aspects of dynamic signaling networks, namely, combinatorial regulation, and time delayed responses in gene activity. All signaling components are considered either active or inactive, i.e. they are represented as binary variables. The activity of each component is modeled as a Boolean function of its parent variables in the network.

*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.

Signaling in biological networks occurs with time delays, which are suitably modeled by the Boolean networks introduced later in the text.

2.1 Boolean networks with unknown time delays and interventions

Let $\mathcal{G} = \{1, \ldots, N\}$ be a set of N signaling components that dynamically interact with each other via transcriptional regulation, and let $\mathbb{F} = \{0, 1\}$ be a Boolean field. Our model represents intracellular gene regulation by a directed graph given by an adjacency matrix $\Gamma \in \mathbb{F}^{\mathcal{G}\times\mathcal{G}}$. It is understood that $\Gamma_{ij} = 1$ whenever i is a parent, i.e. a regulator of j. At each time point t, a gene $j \in \mathcal{G}$ is characterized by two Boolean variables $A_j(t)$ and $B_j(t)$. The induction state variable $A_j(t)$ tells us whether gene j is either transcribed at its basic rate or whether it exhibits altered transcription ($A_j(t) = 0$ or 1, respectively). The activity state variable $B_j(t)$ reports whether the signaling molecule j is in its basic functional state or whether its function is altered at time point t ($B_j(t) = 0$ or 1, respectively, see Fig. 1). It helps to think of the induction states as genes and their expression, and the activity states as the corresponding gene products (proteins) and their activity as transcription factors. Although protein activity can be measured in some instances, it is generally hard to obtain time-resolved data. Therefore, we will infer the activity variables from the expression of their known target genes. The induction state $A_j(t)$ of j at time t is determined instantaneously by the activity states $B_i(t)$ of its parents $i \in \mathrm{pa}(j) \subseteq \mathcal{G}$ via a Boolean function $f_j : \mathbb{F}^{pa(j)} \to \mathbb{F}$,

$$A_j(t) = f_j(B_i(t);\ i \in \mathrm{pa}(j))\ ,\quad j \in \mathcal{G},\ t \in [0, \infty) \tag{1}$$

If $\mathrm{pa}(j) = \emptyset$, f_j is a constant. The family $\{f_j | j \in \mathcal{G}\}$ of Boolean functions is denoted by $\mathcal{F}$. The changes in the activity state of gene j are transmitted to changes in the corresponding activity state with a constant time delay $d_j \in [0, \infty)$,

$$B_j(t) = \begin{cases} A_j(t - d_j) & \text{for } t \geq d_j \\ A_j(0) & \text{else} \end{cases}\ ,\quad j \in \mathcal{G},\ t \in [0, \infty) \tag{2}$$

Let $\Delta = \{d_j | j \in \mathcal{G}\}$. The graph Γ, together with $\mathcal{F}$ and Δ define the dynamics of all binary variables in the model.

To completely specify the Boolean network, we need to initialize the values of $A_j(t)$ at $t = 0$. Through an intervention experiment, some induction states are actively set to 1, $A_j(0) = 1$ (e.g. by a gene knockdown), whereas the rest of the variables are initialized by 0. At the same time, all feedback to an actively perturbed induction state variable A_j is blocked, which is reflected by the removal of all incoming edges to A_j.

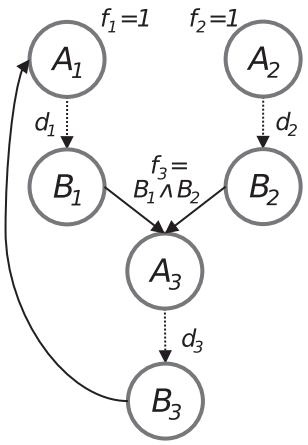

Fig. 1. Schematic of the model for a fixed time point t: A_i and B_i are the induction and activity states, respectively, of each regulator $\{1, 2, 3\}$. The delays in signaling between an alteration of the gene state and an ensuing alteration of the activity state are given by $\Delta = (d_i)$. Given all parent–child relationships of the network, $\mathcal{F} = \{f_1, f_2, f_3\}$ is the family of Boolean functions. Functions for nodes with <2 parents (A_1 and A_2) are constant

In practical situations the delay times Δ are rarely known. We account for this fact by considering the delay times as unknowns for which we specify their prior distribution. The prior is a product of independent exponential distributions, one for each individual delay time,

$$\pi(\Delta;\Lambda) = \prod_{j=1}^{N} \pi_j(d_j;\ \lambda_j)\ ,\quad \pi_j(d_j;\ \lambda_j) = \begin{cases} \lambda_j \exp(-\lambda_j d_j) & \text{if } d_j \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Here, $\Lambda = (\lambda_j)$ is a tuple of appropriately chosen positive hyperparameters, and a complete parametrization of the model is given by the tuple $\mathcal{M} = (\Gamma, \mathcal{F}, \mathcal{L}, \Lambda)$.

2.2 The likelihood function

Let $\mathbf{B} = \{B_j(\tau_k); j \in \mathcal{G},\ k = 0, \ldots, K\}$ the observations of the binary state variables B_j at $K + 1$ time points $0 = \tau_0 < \tau_1 < \tau_2 < \ldots < \tau_K$. Given a parametrization $\mathcal{M}$ of the model and some initial activation pattern, one seeks to calculate the probability of observing $\mathbf{B}$, by integration over the unknown delay times,

$$P(\mathbf{B}|\Gamma, \mathcal{F}, \Lambda) = \int_{\Delta} P(\mathbf{B}|\Gamma, \mathcal{F}, \Delta) \cdot \pi(\Delta;\Lambda) \tag{4}$$

The major technical achievement of this article is the closed-form solution of the integral in Equation (4) for arbitrary Boolean networks (possibly including cycles) that satisfy a rather general admissibility condition (Supplementary Material S1). The class of Boolean networks that can be inferred includes all acyclic networks, and all networks that allow each node to switch only once, yet it is substantially larger. As the derivation of this result requires tedious calculations and elaborate notation, we just give the algorithm for the likelihood calculation in **Algorithms 1 and 2** and refer to Supplementary Methods S2 for details. We also prepared a table of all symbols as Supplementary Materials S6. Some quantities arising during the calculation become extremely small, which bears the risk of underflow errors. Therefore, all necessary computations were performed in log space instead of using standard floating point arithmetic (Supplementary Material S3). Having scored a Boolean network, we search the space of all admissible signaling graphs by Markov Chain Monte Carlo as outlined in Husmeier (2003) (Supplementary Material S5).

Our framework easily allows the modeling of a series of intervention experiments. Each intervention will produce its own sequence of state observations $\mathbf{B}$, and each sequence will be evaluated separately by actively initializing the expression states of perturbed variables with 1 and blocking all feedback to this state by the removal of all incoming edges.

Algorithm 1: Calculation of the likelihood and search through the space of admissible Boolean networks. The scoring function (line 6) is detailed in Algorithm 2

input: maximal scoring state sequence: $\mathbf{B}_{max}$
hyperparameter for the distribution of the delay times: Λ
local probability functions: $P(D|\mathbf{B}, \mathcal{L})$

1 Find $\mathcal{N}(\mathbf{B}_{max})$, all state sequences at Hamming distance 1 of $\mathbf{B}_{max}$
2 Run an MCMC chain over all admissible Boolean networks $(\Gamma, \mathcal{F})$. Acceptance or rejection of proposed models $(\Gamma, \mathcal{F})$ is based on their likelihood L (see Supplementary Materials 3)
3 **foreach** *proposed Boolean network* **do**
4 **foreach** $\mathbf{B} \in \mathcal{N}(\mathbf{B}_{max})$ **do**
  // calculate $S = P(\mathbf{B} \mid \Gamma, \mathcal{F}, \Lambda)$
5 find the set $\mathcal{K}$ of all compatible κ
6 $S_{\mathbf{B}} = \sum_{\kappa\in\mathcal{K}} \mathtt{score}(\mathbf{B}, \kappa, \Lambda)$
7 **end**
  // calculate likelihood L of $(\Gamma, \mathcal{F})$
8 $L = \sum_{\mathbf{B}\in\mathcal{N}(\mathbf{B}_{max})} S_{\mathbf{B}} \cdot P(D \mid \mathbf{B}, \mathcal{L})$
9 **end**

Algorithm 2: Scoring a Single State Sequence B, Given κ and α. The Recursion Will Split into Two Separate Cases Whenever B_k and Its Predecessor $B_{K(k)}$ Switched Values within the Same Observation Interval. If This Happens Too Often, Network Reconstruction Will Be Impossible Anyway. In Practice, A Sufficient Temporal Resolution Will Imply That Scaling of the Algorithm is Roughly Linear in the Number of State Switches

Function `score`$(\mathbf{B}, \kappa, \alpha = \Lambda)$:
input: state sequence **B**,
switch time κ,
parameter of the integral α
For each k, find the interval $[\tau_{i_k}, \tau_{i_{k+1}}]$ where the switch in the state sequence **B** happens calculate:

$$F(j, \beta, \alpha; \kappa) = \int_{t_1=\tau_{i_1}}^{\tau_{i_1+1}} \dots \int_{t_k=\tau_{i_k}}^{\tau_{i_k+1}} \left[exp(\beta t_j) \prod_{i=1}^{k} \pi_i(t_i - t_{\kappa(i)}; \alpha_i) \right] dt_k dt_{k-1} \dots dt_1$$

This is done using the following recursion formula:

$$F(k, \beta, \alpha; \kappa) = \begin{cases} 1 & \text{if } \alpha = \emptyset \\ \hat{c}(k, \beta; \alpha) \cdot F(k, 0, \hat{\alpha}(k, \beta; \alpha); \kappa) & \text{if } \alpha \neq \emptyset, \beta > 0 \\ F(0, 0, (\alpha_1, \dots, \alpha_{k-1}); \kappa) - \exp(-\alpha_k \tau_{i_k+1}) \cdot F(\kappa(k), \alpha_k, (\alpha_1, \dots, \alpha_{k-1}); \kappa) & \text{if } \alpha \neq \emptyset, \beta = 0, t_{\kappa(k)} \geq \tau_{i_k} \\ [\exp(-\alpha_k \tau_{i_k}) - \exp(-\alpha_k \tau_{i_k+1})] \cdot F(\kappa(k), \alpha_k, (\alpha_1, \dots, \alpha_{k-1}); \kappa) & \text{if } \alpha \neq \emptyset, \beta = 0, t_{\kappa(k)} < \tau_{i_k} \end{cases}$$

Here, $\hat{\alpha}(j, \beta; \alpha)$ and $\hat{c}(j, \beta; \alpha)$ are constants defined as

$$\hat{\alpha}(j, \beta; \alpha) = (\hat{\alpha}_i(j, \beta; \alpha))_{i=1,\dots,k}, \ \hat{\alpha}_i(j, \beta; \alpha) = \begin{cases} \alpha_i & \text{if } i \notin \{j, \kappa(j), \kappa^2(j), \dots\} \\ \alpha_i - \beta & \text{if } i \in \{j, \kappa(j), \kappa^2(j), \dots\} \end{cases}$$

$$\hat{c}(j, \beta; \alpha) = \prod_{s \in \{j, \kappa(j), \kappa^2(j), \dots\}} \frac{\alpha_s}{\alpha_s - \beta}$$

3 RESULTS

3.1 Performance on synthetic data

Having in mind the application to stem cell differentiation data with six genes (see Section 3.2), we manually chose five representative topologies with six nodes for our simulation studies with an OR as sole Boolean function. The delay times d_g for each gene g were sampled uniformly from the interval [1, 30] min. The measurements were generated after $t = \{0, 15, 30, 45, 60\}$ min. For each topology, we then calculated the binary activity patterns $B_g(t)$ for each single gene knockout g. The local probability distributions $\mathcal{L} = \{P(D_j|B_j);\ j \in \mathcal{G}\}$ are taken as

$$P(D_j|B_j) \sim \mathcal{N}(D_j; \mu = B_j, \sigma^2)$$

for $\sigma \in \{0.006, 0.12, 0.24, 0.36, 0.46, 0.58\}$. We assume that vague prior knowledge about the delay times is available by choosing the hyperparameter λ_g of the exponential delay time prior such that their expected value equals the respective true delay time d_g.

For each topology, we started 100 MCMC runs with 2000 steps (see Supplementary Material S3 for details on the MCMC method). The likelihood requires the summation over all possible state sequences $\mathbf{B} \in \mathbb{F}^{N \times K}$. This makes calculations infeasible even for medium-sized networks. We address this problem by restricting the model space search to state sequences that are in the immediate vicinity of the best scoring state sequence B_{max}. To find B_{max}, we exploit the fact that in our model, the hidden state variables B_j change their value from 0 to 1 at most once in their time course (due to the monotonicity of the chosen Boolean function, OR). This means that for each B_j we can summarize its time course by denoting the time at which the state change occurs by the random variable T_j called change point. The contribution of the hidden state B_j to $P(D|\mathbf{B}, \mathcal{L})$ is

$$\prod_{k=1}^{K} P\big(D_j(\tau_k)|B_j(\tau_k)\big) = \prod_{k=1}^{K} P(D_j(\tau_k)|B_j = \delta(\tau_k > T_j)) \qquad (5)$$

where $\delta(\tau_k > T_j) = \begin{cases} 1 & \text{if } \tau_k > T_j \\ 0 & \text{else} \end{cases}$ is the indicator function. Thus, there are at most $K + 1$ different time courses for B_j [(0,0,...,0), (0,0,...,0,1),... (0,1,1,...,1), (1,1,...,1)]. Enumerating these, we find the time course for B_j that maximizes the term in (5). Doing so for all $j \in \mathcal{G}$, we find the best scoring state $\mathbf{B}_{max}$. Figure 2 shows the results for all five topologies. The model shows a good overall performance for low and moderate noise levels. It performs best on tree topologies (Fig. 2E), which are often encountered in biological pathways. Another frequent pathway motif is the feed-forward loop, as modeled in Figure 2B. The addition of feedback to the linear topology in Figure 2A decreases performance, but it still remains at a reasonable level. Figure 2F shows the results on a biological network from literature [of the stem cell differentiation pathway from Anchang *et al.* (2009)]. Specificity and sensitivity are comparable with the simpler topologies A–E.

3.2 Application to stem cell differentiation data

Our model calls for time series measurements of protein activities after single gene knockouts. Data of that kind are still sparse. We circumvented this problem and increase the applicability of the method by treating the binary activity state variables as hidden variables. Our data consist of time series of measurements $D = (D_j(\tau_k))$ of the activity states $B_j(\tau_k)$, $j \in \mathcal{G}$, at a finite number of $K + 1$ time points $0 = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_K$.

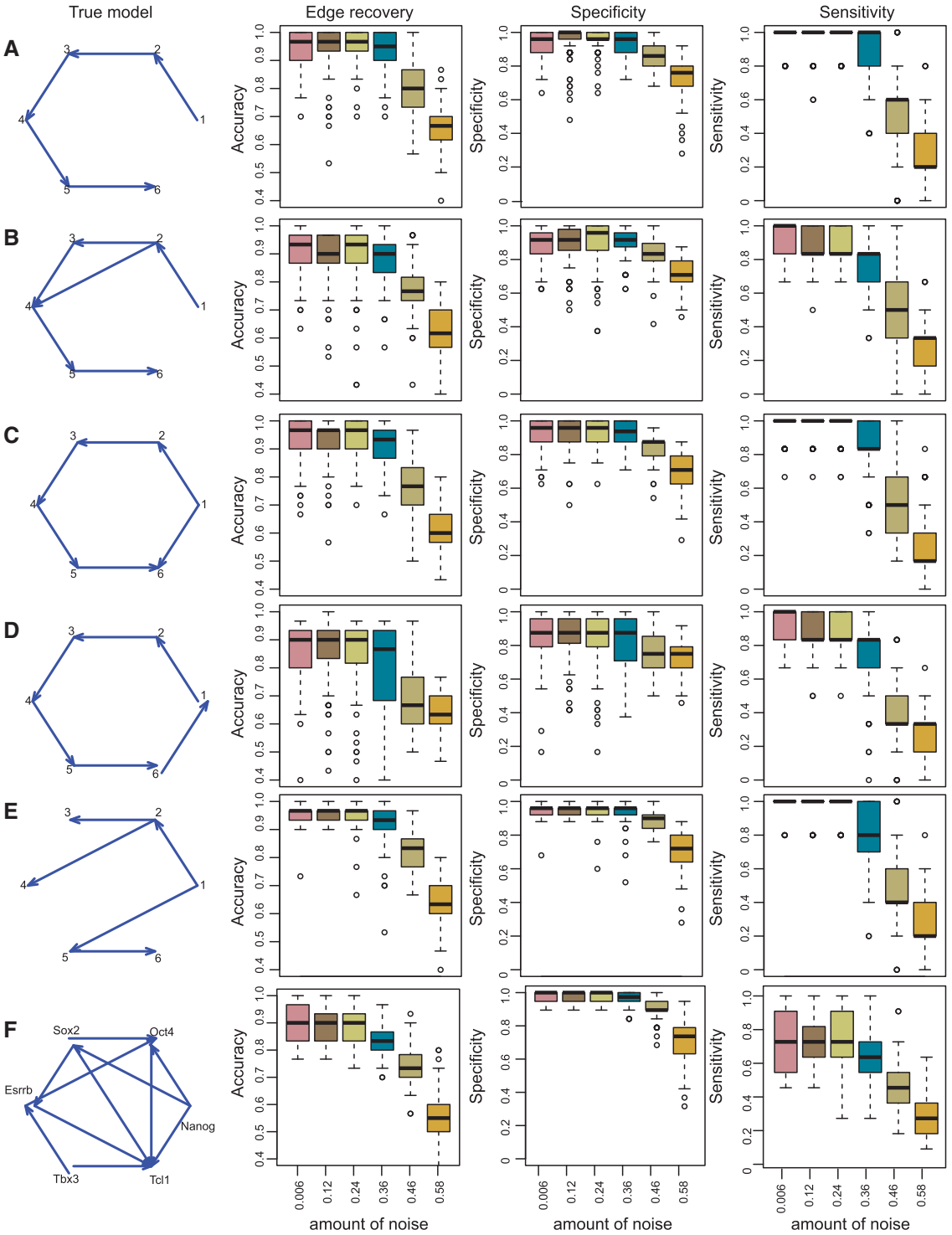

Fig. 2. Results of the simulations study on five topologies (first column). The second column shows the performance as percentage of correctly predicted edges (presence and absence) for different noise levels σ added to the binary activity patterns as a box plot over all 100 runs of the MCMC. The third and fourth columns show the distribution of sensitivity and specificity of network reconstruction across all runs. (**A**) Linear graph. This topology can be predicted with high accuracy up to noise level 0.36. (**B**) Linear graph with feed-forward loop. This topology is also correctly predicted up to noise level 0.36, although we lose 0.1 performance points compared with the linear graph without shortcut. (**C**) Linear graph with forward-jump to the last node. The model can better predict this case than the intra-node forward-jump in B. (**D**) Full cycle. This difficult topology can be predicted by the model with accuracy >80% up to noise level 0.36. Performance then rapidly degrades. This topology has a high variance in sensitivity/specificity values between the different runs even for low levels of added noise. (**E**) Tree structure. The model is well adapted to this topology and shows a high performance until noise level 0.36. Its performance is comparable with the linear topology (A). (**F**) Network of stem cell differentiation as reconstructed by Anchang *et al.*

The data $D_j(\tau_k)$ can be thought of as a noisy possibly replicate quantification of the hidden activity states $\mathbf{B} = \{B_j(\tau_k); j \in \mathcal{G}, k = 0, \ldots, K\}$. We relate measurements to their underlying activity state through time-independent local probability distributions $\mathcal{L} = \{P(D_j|B_j); j \in \mathcal{G}\}$. Given the hidden induction states $\mathbf{B}$ and the local probabilities $\mathcal{L}$, the probability of observing D is

$$P(D|\mathbf{B}, \mathcal{L}) = \prod_{j=1}^{N} \prod_{k=1}^{K} P(D_j(\tau_k)|B_j(\tau_k)) \tag{6}$$

Equation (6) assumes independence of observations. The likelihood then becomes

$$P(D|\Gamma, \mathcal{L}, \mathcal{F}, \Delta) = \sum_{\mathbf{B} \in \mathbb{F}^{N \times K}} P(D|\mathbf{B}, \mathcal{L})(\mathbf{B}|\Gamma, \mathcal{F}, \Delta) \tag{7}$$

Thus, we can apply the method to the dataset of Ivanova *et al.* (2006) who used short hairpin RNA loss of function techniques to downregulate genes whose expression patterns suggest self-renewal regulatory functions in mouse embryonic stem cells. Genome-wide gene expression time series measurements after $t = 0, 1, \ldots, 7$ days were obtained after knockdown of each of the following genes: Nanog, Oct4, Sox2, Tbx3, Esrrb and Tcl1. These genes are known to play a major role in stem cell differentiation and are therefore called 'major genes'. Anchang *et al.* (2009) built a model with this knockdown data using dynamic nested effect models.

The major genes represent the nodes in our network. The variables $A_j(t)$ and $B_j(t)$ correspond to their gene and protein activities, respectively. Because the activity states $B_j(t)$ are not directly measured by Ivanova *et al.* (2006), we use the expression activity of gene groups under the regulatory control of the major genes [the *E*-Genes in the nested effect model of Anchang *et al.* (2009)] as a proxy for their protein activity. To get the local probabilities $P(D_j|B_j)$ needed in the case of assuming hidden $B_j(t)$, we use data from 122 genes given as discretized time series representing admissible patterns [see the Supplementary Materials of Anchang *et al.* (2009) for details]. In accordance with our definition, genes in their basic state were assigned the value 0, and assumed the value 1 on activation. We kept the grouping of the 122 genes into six groups of genes depending on Nanog, Oct4, Sox2, Tbx3, Esrrb or Tcl1 discovered from the *E*-Genes graph from Anchang *et al.* Because the data contain time series representing the undifferentiated cell culture (the negative control), and the cell culture undergoing normal differentiation (the positive control), we filtered for genes whose expression differed more than two-fold at the last time point between the two control experiments. Then, we assigned to each gene at each time point a probability to belong to the basal or the active state, according to whether its expression resembled more the negative or positive control (a likelihood ratio was calculated under the assumption of Gaussian expression distributions). Using the gene groups defined earlier, we calculated a likelihood ratio for each major gene to be active versus inactive as the product of the corresponding likelihood ratios of the assigned genes (this was done separately for each time point and each knockout). The likelihood ratios are then converted into a probability of being active (at a certain time point, in a certain knockdown experiment), which corresponds to the input required for our model.

In this application, we only use the Boolean function AND, leading to monotonic activity states $\mathbf{B}$. As described in Section 3.1, we chose the state sequence $\mathbf{B}_{max}$ that maximizes $P(D|\mathbf{B}, \mathcal{L})$.

Using the same MCMC procedure as in the simulation setting (Supplementary Material S5), the stationary chain comprised 155 unique models. We used model averaging and calculated the weighted frequencies of each edge. Each model was weighted by its number of occurrences in the Markov chain, resulting in a probabilistic adjacency matrix (Fig. 3A). Tcl1 has the lowest connectivity, whereas Nanog has the highest. To compare the results of our model with the model from Anchang *et al.* (Fig. 3C), we converted the probabilistic adjacency matrix into a graph by drawing all edges with a probability >0.5 (Fig. 3B). The most striking difference of Figure 3B compared with Figure 3C is the presence of cycles. In particular, the major genes Oct4, Sox2, Nanog and Esrrb form a maximal clique of the graph.

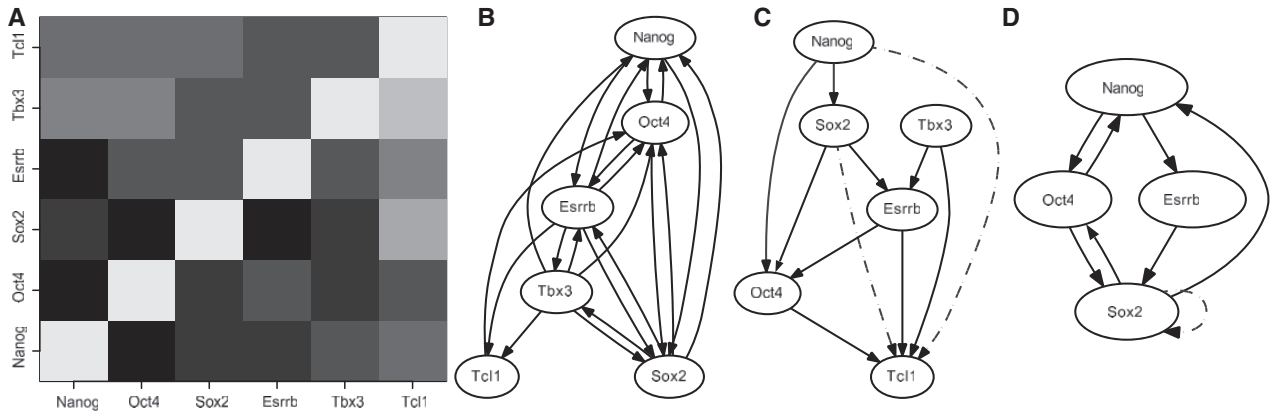

Fig. 3. (**A**) Adjacency matrix of the result of the network inference on the biological dataset. Each entry corresponds to the observed frequency and is colored accordingly with lighter colors representing lower frequencies (**B**) Network obtained from A by setting a threshold of 0.5 on the edge probability (**C**) Network inferred by Anchang *et al.* (2009) (**D**) Extract from the network published in Zhou *et al.* (2007). The authors did not include Tbx3 and Tcl1 in their findings. Dashed edges in B and C represent edges that are not present in our model (A). All other edges from B and C are also found in model A

The two graphs essentially agree on the position of Tcl1, which in both cases is targeted by Tbx3 and Esrrb. Also, Tbx3 is located mostly upstream of the Oct4, Sox2, Nanog, Esrrb clique in both graphs. Still, it is puzzling why our method finds a highly interconnected feedback-loop rich structure, whereas Anchang *et al*. find a sparser solution. The method in Anchang *et al*. assumes an acyclic graph structure, and hence by definition cannot find cycles. As our simulation studies have shown that the model can accurately predict circular structures in regulatory graphs, the feedback in this network might be higher, and the signaling hierarchy less pronounced than previously thought. This is confirmed by a different approach to mouse embryonic stem cell network reconstruction (Zhou *et al*., 2007) that also discovers a large amount of interplay between the key regulators of stem cell differentiation. Zhou *et al*. have also reconstructed a mouse embryonic stem cell network based on transcription factor binding sites, protein interactions and literature annotation. They show bidirectional interactions of Oct4 with Nanog and Sox2 coinciding with our finding (Fig. 3D).

4 CONCLUSION

In this work, we developed an algorithm that permits us to analyze gene knockdown time series experiments, which have high dimensional readouts (such as gene expression). To elucidate the interplay of the major regulators, all of them need to be perturbed and measured individually. On the side of methods development, we have solved the problem of calculating the likelihood function for data generated from a Boolean network with probabilistic exponentially distributed time delays (**Algorithm 2**). The likelihood function can be used for network reconstruction, as has been demonstrated in our simulation studies. Having a closed form solution for the likelihood has several further applications that we did not mention so far. It is possible to sample the joint distribution $P(\mathbf{B}|\Gamma, \mathcal{F}, \Lambda)$ rather efficiently, because many observations $\mathbf{B}$ can be excluded a priori knowing Γ and $\mathcal{F}$. This allows for accounting for some hidden variables B_k among the observed $\mathbf{B}$ by integrating them out. Furthermore, it is possible to calculate the expectation of a certain B_j to be on or off in a given time interval. As an application, we have devised a method to apply it to data in which the values of the Boolean network can only be observed indirectly (**Algorithm 1**). We analyzed murine stem cell differentiation data of Ivanova *et al*. (2006) for the purpose of signaling network reconstruction. Comparison with a previous reconstruction attempt in Anchang *et al*. (2009) revealed a much richer feedback structure than expected. Our method suggests regulatory feedback loops that lead to a better understanding of the dynamic interplay of some master regulators in murine embryonic stem cell development. We expect our method to find numerous applications, as protein abundance data become increasingly available (Fröhlich *et al*., 2009).

ACKNOWLEDGEMENT

The authors thank Holger Fröhlich for useful suggestions during the preparation of the manuscript.

Funding: Deutsche Forschungsgemeinschaft (SFB646 to A.T.).

Conflict of interest: none declared.

REFERENCES

Anchang,B. *et al.* (2009) Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. *Proc. Natl. Acad. Sci. USA*, **106**, 6447–6452.

Basso,K. *et al.* (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet*., **37**, 382–390.

Beerenwinkel,N. and Sullivant,S. (2009) Markov models for accumulating mutations. *Biometrika*, **96**, 645–661.

Beerenwinkel,N. *et al.* (2007) Conjunctive bayesian networks. *Bernoulli*, **13**, 893–909.

Failmezger,H. *et al.* (2013) Learning gene network structure from time laps cell imaging in RNAi knock downs. *Bioinformatics*, **29**, 1534–1540.

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

Fröhlich,H. *et al.* (2008) Analyzing gene perturbation screens with nested effects models in R and bioconductor. *Bioinformatics*, **24**, 2549–2550.

Fröhlich,H. *et al.* (2009) Deterministic effects propagation networks for reconstructing protein signaling networks from multiple interventions. *BMC Bioinformatics*, **10**, 322.

Fröhlich,H. *et al.* (2011) Fast and efficient dynamic nested effects models. *Bioinformatics*, **27**, 238–244.

Grzegorczyk,M. *et al.* (2008) Modelling non-stationary gene regulatory processes with a non-homogeneous bayesian network and the allocation sampler. *Bioinformatics*, **24**, 2071–2078.

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

Ideker,T.E. *et al.* (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac. Symp. Biocomput*., **2000**, 305–316.

Ivanova,N. *et al.* (2006) Dissecting self-renewal in stem cells with RNA interference. *Nature*, **442**, 533–538.

Kauffman,S.A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol*., **22**, 437–467.

Niederberger,T. *et al.* (2012) Mc eminem maps the interaction landscape of the mediator. *PLoS Comput. Biol*., **8**, e1002568.

Papin,J.A. *et al.* (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol*., **6**, 99–111.

Shmulevich,I. *et al.* (2002) Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **18**, 261–274.

Silvescu,A. and Honavar,V. (2001) Temporal Boolean network models of genetic networks and their inference from gene expression time series. *Complex Syst*., **13**, 54–70.

Tresch,A. *et al.* (2008) Structure learning in nested effects models. *Stat. Appl. Genet. Mol. Biol*., **7**, 9.

Van Driessche,N. *et al.* (2005) Epistasis analysis with global transcriptional phenotypes. *Nat. Gen*., **37**, 471–477.

Zhou,Q. *et al.* (2007) A gene regulatory network in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. USA*, **104**, 16438–16443.