Gene expression

Gene network inference from incomplete expression data: transcriptional control of hematopoietic commitment

Kristin Missal\textsuperscript{1,2}, Michael A. Cross\textsuperscript{3} and Dirk Drasdo\textsuperscript{2,4,*}

\textsuperscript{1}Bioinformatics Group, Department of Computer Science and \textsuperscript{2}Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany, \textsuperscript{3}Interdisciplinary Center for Clinical Research and Division of Hematology/Oncology, University of Leipzig, Inselstrasse 22, D-04103 Leipzig, Germany and \textsuperscript{4}Max-Planck-Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany

Received on May 9, 2005; revised on October 28, 2005; accepted on December 4, 2005

Advance Access publication December 6, 2005

ABSTRACT

Motivation: The topology and function of gene regulation networks are commonly inferred from time series of gene expression levels in cell populations. This strategy is usually invalid if the gene expression in different cells of the population is not synchronous. A promising, though technically more demanding alternative is therefore to measure the gene expression levels in single cells individually. The inference of a gene regulation network requires knowledge of the gene expression levels at successive time points, at least before and after a network transition. However, owing to experimental limitations a complete determination of the precursor state is not possible.

Results: We investigate a strategy for the inference of gene regulatory networks from incomplete expression data based on dynamic Bayesian networks. This permits prediction of the number of experiments necessary for network inference depending on parameters including noise in the data, prior knowledge and limited attainability of initial states. Our strategy combines a gradual ‘Partial Learning’ approach based solely on true experimental observations for the network topology with expectation maximization for the network parameters. We illustrate our strategy by extensive computer simulations in a high-dimensional parameter space in a simulated single-cell-based example of hematopoietic stem cell commitment and in random networks of different sizes. We find that the feasibility of network inferences increases significantly with the experimental ability to force the system into different network states, with prior knowledge and with noise reduction.

Availability: Source code is available under: www.izbi.uni-leipzig.de/services/NetwPartLearn.html

Contact: drasdo@izbi.uni-leipzig.de

Supplementary information: Supplementary Data are available at Bioinformatics online.

1 INTRODUCTION

The lineage choice of hematopoietic stem cells is believed to be determined by network interactions between a relative small number of lineage-associated transcription factors and their corresponding genes (Cross and Enver, 1997; Hoang, 2004), making this clinically relevant system a prime candidate for mathematical modeling. Although regulatory networks can theoretically be inferred from time-series of gene expression levels in cell populations assayed at successive time points by microarray experiments (Thomas et al., 2004; Bar-Joseph, 2004; Zou and Conzen, 2005), this approach is unsuitable for asynchronous populations such as hematopoietic stem cells in which averaging values can misrepresent network features at the single cell level. For cases such as these, experimental procedures have been developed that permit the determination of gene expression levels in single cells individually using a reverse transcription–polymerase chain reaction (RT–PCR) method which maintains representation levels between different mRNAs (Brady et al., 1995). The major drawback of the procedure is that the cell is destroyed during mRNA extraction, so that the gene expression pattern within one cell can be measured only once, whereas network inference requires at least partial knowledge of two successive states. One way to gain knowledge of two successive network states is therefore to up- or down-regulate the expression level of individual genes experimentally by the introduction of recombinant genes or specific inhibitory molecules into the cell (Hannon, 2002; McIvor et al., 2003) followed by a complete monitoring of the network state after a transition has occurred. Since the number of gene expression levels that can be adjusted simultaneously in this way is currently limited practically (typically to just one or two genes), the knowledge of the manipulated network state (the state prior to the transition) is largely incomplete, so that the experimental strategy is feasible only for relatively small gene regulation networks (such as those of hematopoietic lineage commitment) or for subnetworks of larger networks (Pe’er et al., 2001).

Here we use in silico simulations of hematopoietic lineage commitment to study the efficiency and reliability of reverse engineering of gene regulation networks from transition data which are largely incomplete, in order to identify the approximate number of experiments which would be necessary for the reliable inference of the lineage commitment network. For this purpose we generated artificial expression data from Boolean networks [BoolNs; (Kauffman, 1993)] by computer simulations, used reverse engineering strategies to infer networks from the data and compared the inferred networks to those that we originally used to generate the data (similarly to Liang et al., 1998; Akutsu et al., 1998; Yu et al., 2004; Rice et al., 2005). The use of BoolNs is undoubtedly an oversimplification in many biological situations (D’haeseleer...
et al., 2000) but they have been successfully used to model the gene regulatory network in a number of biological systems, e.g. Drosophila melanogaster (Albert and Othmer, 2003). Expression profiles of genes which could serve as a guide to substitute missing gene states (Troyanskaya et al., 2001) are not available, as is often the case in small networks. In the situations studied in our simulations, the number of known gene expression levels from the precursor network state is much smaller than the largest number of input states for one gene. This means that the information transfer on an element divided by the entropy of that element is seldom exactly one and, since this is the criterion used in the REVEAL algorithm (Liang et al., 1998) to determine the input elements for a given element, that REVEAL is not appropriate. Our inference strategy uses a \( \chi^2 \)-test to decide which are the input elements of a given element and may be viewed as an extension of the classical REVEAL algorithm (see Supplementary Material; compare also Murphy and Mian, 1999, http://citeseer.ist.psu.edu/murphy99modelling.html). We use the framework of reverse engineering methods of dynamic Bayesian networks (DBNs) [e.g. Heckerman et al., 1995; Murphy and Mian, 1999; Pe’er et al., 2001; Yu et al., 2004; Zou and Conzen, 2005] to identify the most probable model given the artificially generated data, so that our inferred networks are represented by DBNs. DBNs use a probabilistic approach and are particularly suitable for situations in which (1) the degree of uncertainty is high as a consequence either of the inherent stochasticity of the biological processes during gene regulation, of measurement errors, or of inconsistencies in the data (Ong et al., 2002), (2) variables may be missing, as would be the case if not all influences are known (Beal et al., 2005) and (3) inference is to be improved by the inclusion of prior knowledge (Heckerman, 1995). Peer and co-workers (Sachs et al., 2005) have recently studied experiments on single-cell expression data based on a Bayesian network inference method introduced in the work by Pe’er et al. (2001) in which all network elements are measured. Their approach is based on the maximization of a score function which under some assumptions may be approximated by the Bayesian information criterion (BIC; Heckerman, 1995, see below).

Our initial studies suggested that structural expectation maximization [involving the substitution of missing data by expected values (Friedman, 1997; Friedman et al., 1998)] is inappropriate in cases in which the state of many elements remains unknown. We, therefore, focus on a gradual ‘Partial Learning’ (PartLearn)-strategy which is based solely on observed experimental data (Conant, 1988; Murphy and Mian, 1999). Since BoolNs are a subclass of DBNs (Murphy and Mian, 1999), the PartLearn-strategy includes REVEAL as a special case. Since we use a probabilistic approach we are also able to identify suboptimal network topologies (the wiring) and rules (the quality of the inputs). Using these approaches, we find that the required number of experiments using the technology currently available is large but feasible. (Notice the list of symbols in the Supplement.)

2 MODELING AND REVERSE ENGINEERING METHODS

A variety of published expression and regulation data (Hoang, 2004) were used to formulate a hypothetical network of 11 transcription factor genes (Fig. 1), six of which form a core network which is unaffected by the other five. Each hypothetical interaction was assigned to one of three probability groups: highly probable (label 1), probable (label 2) and possible but not experimentally examined (label 3). This classification concerns only the occurrence and not the quality (the type) of the influences.

We use BoolNs to represent the gene regulation networks. The probability \( p_i, i \in \{1, 2, 3\} \), for the insertion of an edge (an arrow in Fig. 1) is chosen according to the labels; \( p_1 = 0.9, p_2 = 0.66, p_3 = 0.5 \). Edges not shown in Figure 1 correspond to relationships for which we could find no inference from the published literature. Simulations in which we introduced these edges with probability 0.1, resulted in only negligible differences. Repressors (labeled with ‘-‘ in Fig. 1) were assumed to act dominantly (Burke and Baniahmad, 2000) and were modeled by an AND-relation, while activators (labeled with ‘+‘ in Fig. 1) which may act either competitively or synergistically (Merika and Thanos, 2001) were modeled by or functions. The resulting Boolean rules are canalizing functions,\(^1\) in which at least one state of at least one input element guarantees at least one state of the output element independently of the states of all other input elements (Kauffman, 1993). For example, the number of canalizing functions \( k_{can}(k) \) for Boolean functions with \( k \) inputs is \( k_{can}(1) = 2, k_{can}(2) = 8, k_{can}(3) = 88, k_{can}(4) = 3104, \) etc. while the total number of Boolean functions with \( k \) inputs is \( 2^{2^k} \).

The initial value of each element is chosen randomly with equal probability from the set \( \{0,1\} \). We studied the situation in which each initial network state is chosen only once (in order to avoid redundancy) compared with the situation where the initial network states were chosen randomly. The non-redundant case allows us to uniquely identify the number of different initial network states necessary to infer the network topology and parameters. In accordance with the technical feasibility, we assume that either \( n = 1 \) or \( n = 2 \) of the \( N \) elements are known before transition, while the network state that evolves after one transition is assumed to be completely determinable.

In order to eliminate possible limitations due to the particular network choice of Figure 1 we also performed extensive simulations with random networks of \( N = 6, 10, 15, 20 \) elements. The inference results for the random networks were found to differ

\(^1\)This was achieved by investigating the Boolean functions of 1000 randomly created BoolNs, according to Figure 1.
only for \( n = 1 \) to those for the hematopoietic network in Figure 1 and are presented in the Supplementary Material.

**Reverse engineering method:** The topology and parameter of the DBN were inferred by the following steps:

1. Initial determination of model parameters and topology of the DBN. [See lines 1 and 2 Alg. PartLearn (Supplementary Material).]

2. Optimization of structure.
   - (a) Generation of a set of successors of the DBN by inserting edges. [See line 7 Alg. PartLearn (Supplementary Material).]
   - (b) Scoring all successors locally. A local score value evaluates the event to which element \( X_i[t] \) is dependent on its putative parents \( Pa(X_i[t]) \). [See lines 8–10 Alg. PartLearn (Supplementary Material) and Equation (1).]
   - (c) Continue with the best scoring successors, i.e. Hill climbing. [See line 11 Alg. PartLearn (Supplementary Material) and Equations (2), (3) and (5)].

3. Optimization of parameters.
   - Given an optimal topology of the DBN, parameters \( P(X_i[t] = \bar{k}|Pa(X_i[t]) = \bar{j}) \) are estimated by expectation maximization (EM) [Dempster et al., 1977; Friedman, 1997].
   - A dynamic Bayesian network DBN = \((G, \Theta)\) consists of a set of graphs \( G = (G^{(0)}, G^-) \) and a set of parameters \( \Theta = (\Theta^{(0)}, \Theta^-) \). Index \('0'\) denotes the initial situation and \('-'\) the transitions. Since the initial state of each element of a network is chosen randomly, the start structure of the DBN cannot contain correlations. We, therefore, focus below on the transition structures only (\( t \geq 1; t \) time) and drop the index \('-'\). \( G \) is a directed acyclic graph whose vertices correspond to random variables \( X_i[t] = \{X_i[1], \ldots, X_i[t]\} \) such that \( 1 \leq t \leq T \). \( N \) is the fixed number of variables in a time slice \( t \). \( \Theta \) is a set of conditional probability distributions \( \theta_i[t] = P(X_i[t] | X_i[t-1], \ldots, X_i[1]) \) for each random variable \( X_i[t] \) in \( G \). \( G \) expresses conditional independence assumptions for each variable \( X_i[t] \), \( G \) obeys the Markov assumption, namely, that each random variable is independent of its non-descendants given its parents \( Pa(X_i[t]) \). Hence, \( \theta_i[t] = \prod_{k \in \bar{I}} \prod_{j \in \bar{J}_k} P(X_i[t] = k | Pa(X_i[t]) = \bar{j}) \), where \( P(\cdot) \) denotes the probability that \( X_i[t] = k \) given that its parents \( Pa(X_i[t]) \) in \( G \) have state \( \bar{j} \) (note that the vector \( \bar{j} \) contains the state of each parent of element \( X_i[t] \)).
   - Knowing the topological structure of the graph \( G \) permits to denote the state of an element by the states of its parents. The joint probability distribution of the DBN represents the trajectories of the process, which, applying the chain rule of probability can be written in the form \( P(X_0[1], \ldots, X_T[t]) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(X_i[t] | Pa(X_i[t])) \). We assume that the process of gene regulation is Markovian and homogeneous in time meaning that the transition probabilities are time-independent. Under these assumptions only two successive time slices need to be considered [Friedman et al., 1998].

The main problem of reverse engineering from expression data of single cells is the lack of information concerning the expression levels in the previous transition network state. We use PartLearn to deal with this problem.

**PartLearn:** In this approach (Murphy and Mian, 1999) subsets of parents with significant impact on the target gene \( X_i \) at time \( t \) (note that the expression level of gene \( X_i \) at time point \( t \) corresponds to the values of random variable \( X_i[t] \)) are identified based on the mutual information

\[
\text{MI}(X_i[t], Pa(X_i[t])) = \sum_{k_{i-1}} \sum_{j_i} P(X_i[t] = k_{i-1}, Pa(X_i[t]) = j_i) \times \log_2 \frac{P(X_i[t] = k_{i-1}, Pa(X_i[t]) = j_i)}{P(X_i[t] = k_{i-1}) P(Pa(X_i[t]) = j_i)}
\]

(1)

This quantifies the impact of the potential regulatory input genes \( Pa(X_i[t]) \) on the target gene \( X_i[t] \) in time slice \( t \), \( r_i \) is the number of states of the element \( i \) and does not change in different time slices. \( j_i[t] \) denotes the set of possible state vectors of its parents at time \( t \), where \( j_i[t] = \{ j_{i,t}, \ldots, j_{i,T} \} \). Here, \( q_i[t] \) is the number of possible state vectors of its parents at time \( t \) (\( q_i^* = 2^r \) for \( p \) parents), and \( j^*_{l} \) is the \( l \)-th state vector of these parents. The probabilities \( P(\cdot) \) are estimated from the relative frequencies of the occurrences of the event \( '\cdot' \).

The mutual information score of a DBN is decomposable:

\[
\log \text{MI}(G : D) = \sum_{t=1}^{T} \sum_{i=1}^{N} \log \text{MI}(X_i[t], Pa(X_i[t]))
\]

and hence the maximal score is achieved by adding all edges for which \( \text{MI}(X_i[t], Pa(X_i[t])) > 0 \). \( D \) denotes the data. We use a \( \chi^2 \)-independence test to distinguish whether a regulatory influence is highly probable or by chance (Miller, 1955). Our Null hypothesis is that \( H_0 : \text{MI}(X_i[t], Pa(X_i[t])) = 0 \) and our test quantity is

\[
\text{MI}(X_i[t], Pa(X_i[t])) M \ln 4 \sim \chi^2_{(r_i-1)(q_i^* - 1) - 1} \sim a^2,
\]

where \( M \) is the number of observed transition samples. In the resulting model structure there should be edges between \( Pa(X_i[t]) \) and the target gene \( X_i[t] \) if and only if the regulatory relation exists with high probability. In our simulation we studied significance levels of \( \alpha \in [10^{-5}, 0.1] \). In order to assess whether the addition of an element to a set of elements which has already been identified to have a significant regulatory impact leads to a significant increase in the regulatory impact, a further \( \chi^2 \)-independence test is performed. For this test the Null hypothesis is \( H_0 : \text{MI}(X_i[t], Pa(X_i[t])) = \text{MI}(X_i[t], Pa_0(X_i[t])) = 0 \) with a test quantity \( \Delta M \equiv (\text{MI}(X_i[t], Pa(X_i[t])) - \text{MI}(X_i[t], Pa_0(X_i[t]))) M \ln 4 \) which has been shown to follow also approximately a \( \chi^2 \) distribution (Conant, 1988)

\[
\Delta M \sim \chi^2_{(r_i-1)(q_i^* - 1) - 1} \sim a^2 \sim a.
\]

(3)

The set \( Pa(X_i[t]) \) contains one additional element than its subset \( Pa_0(X_i[t]) \). Both sets, \( Pa_0(X_i[t]) \) and \( Pa(X_i[t]) \) denote sets of potential parents of element \( X_i[t] \). Only if \( H_0 \) for \( \Delta M \) is rejected, the additional element of \( Pa(X_i[t]) \) is added to the final list of parents of element \( X_i[t] \).

After having learned a high probable structure of the DBN a full expectation maximization (EM) iteration is applied to obtain parameter estimates \( \hat{P}(X_i[t] = k_{i-1} | Pa(X_i[t]) = j_{i-1})^{2} \)

Two alternatives to identify the network topology may be used.

1. The Bonferroni criterion could be used to determine the significance level for the test. Given, that for \( n = 1 \) maximal \( n_T = N \times N \),

\[\text{We used the EM implementation of the LibBi toolkit (Friedman and Elidan, version 2.1, http://www.cs.huji.ac.il/labs/cmpbio/LibBi/).}\]
and for $n = 2$ maximal $n_T = N^2 + 3N(n^2/2)$ tests are performed, the Bonferroni criterion suggests to use the significance level $a/n_T$. The factors $N^2$ and $3N(n^2/2)$ result from the test in Equation (2), the factor ‘3’ from the test in Equation (3). Our simulations with selected parameter combinations have confirmed the known conservativity of the Bonferroni criterion meaning that both the fidelity and the sensitivity saturate later (Supplementary Material). (2) Alternatively one may maximize the Bayesian information score, $\text{BIC} = \sum_{i=1}^{\infty} \left[ \text{MI}(X_i, \mathbf{P}_{\mathbf{X}}(X_i)) - H(X_i) - (\log(M)/2)Nq_i(r_i - 1) \right]$. Here, the first two terms on the RHS denote the log-likelihood score while the last term is a penalty term to avoid overfitting. $H(X_i[t])$ is the entropy of element $X_i[t]$. We expect the result to be similar.

PartLearn and prior knowledge: We modify the PartLearn reverse engineering strategy to identify the topology such that edges are now identified by a posterior odds ratio (POR) test. This enables us to incorporate prior knowledge directly by treating prior probabilities of the occurrence of a regulatory influence between each gene and its potential regulatory input genes as additional observations. The POR evaluates the posterior distributions as evidence in favor or against the hypotheses $H_0$ (variables are independent) with respect to $H_1$ (variables are dependent)

$$\text{POR} = \frac{P(H_0 | D)}{P(H_1 | D)} = \frac{P(D | H_0)P(H_0)}{P(D | H_1)P(H_1)} = \text{LR} \frac{P(H_0)}{P(H_1)}$$

$$\text{LR} = \frac{\prod_{i, i \neq j} P(X_i[t] = k_i)N_{i,t}^{k_i}}{\prod_{k, i \neq j} P(X_i[t] = k, \mathbf{P}_{\mathbf{X}}(X_i[t]) = j)^{N_{i,t}^{k,j}}}. \quad (4)$$

$N_{i,t}[t]$ is the number of observed transition samples where $X_i[t] = k_i$ and $\mathbf{P}_{\mathbf{X}}(X_i[t]) = j$. $N_{i,t}[t]$ is the number of observations, where $X_i[t] = k_i$ and $N_{i,t}[t]$ is the number of observations, where $\mathbf{P}_{\mathbf{X}}(X_i[t]) = j$. $\text{POR} > 0$ (POR < 0) is evidence in favor of $H_0 (H_1)$. The prior distributions are taken directly from the probabilities that an edge occurs in our hypothetical transcription factor network. For example, if a regulatory relation of an input gene and its target gene is labeled with 1 in Figure 1 then $P(H_0) = 0.1$ and $P(H_1) = 0.9$. Quantifying the impact of a set of potential regulatory input genes $\mathbf{P}_{\mathbf{X}}(X_i[t])$ on a target gene $X_i[t]$ by a POR is equivalent to characterizing the impact by an MI: $-2\text{POR}(X_i[t], \mathbf{P}_{\mathbf{X}}(X_i[t])) = M' \ln(4)\text{MI}'(X_i[t], \mathbf{P}_{\mathbf{X}}(X_i[t]))$. The equality that was originally formulated for the likelihood ratio (LR) (Miller, 1955) can be used for the POR by noting that the influence of the prior $P(H_0)P(H_1)$ can be interpreted as the result of additional (virtual) observations which leads to a modified number of experiments $M'$ and a modified MI' leaving, however, $r_i$ and $q_i$ invariant. Hence the strength of evidence in favor of hypothesis $H_0$ can be again quantified by a chi-square-test reducing the method to PartLearn with prior knowledge:

$$-2\text{POR}(X_i[t], \mathbf{P}_{\mathbf{X}}(X_i[t])) \sim \chi^2_{e-\{1\} \cup \{0\} \cup \{1\} - \alpha}. \quad (5)$$

Assessment of parameters: To assess the learned model parameters we evaluate the relative entropy (Heckerman et al., 1995)

$$H_{\text{rel}}(P, Q) = \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{r_i} \sum_{j=1}^{q_i} P(X_i[t] = k_i, \mathbf{P}_{\mathbf{X}}(X_i[t]) = j) \times \log_2 \frac{P(X_i[t] = k_i, \mathbf{P}_{\mathbf{X}}(X_i[t]) = j)}{Q(X_i[t] = k_i, \mathbf{P}_{\mathbf{X}}(X_i[t]) = j)}. \quad (6)$$

This measure provides an evaluation of how well the model is able to explain an independent test dataset. If a model explains the data well, $H_{\text{rel}}(P, Q)$ is small. $P(\cdot)$ are estimated from the relative frequencies observed in the test dataset generated from the original network and $Q(\cdot)$ are the learned parameters of the DBN.

Classification of simulation results: We introduce the fidelity $f_k \in [0, 1]$ defined as the number of elements for which all parents have been correctly identified divided by the total number of elements with a certain $k$ (i.e. no false positives and no false negatives) to quantify the similarity of the network structure of the inferred DBN and the BoolN. $f_k = 1$ means that those and only those elements for that particular $k$ have been correctly identified. We study how the fidelity $f_k$ behaves with the number of state transition measurements $M$. To evaluate the inferred structure we also calculate the sensitivity, $S_k$, which is defined as the number of correctly identified inputs of the elements $X_i[t]$ (true positives) divided by the number of inputs in the BoolN (sum of true positives and false negatives). This differs from the fidelity in that $S_k$ refers to the inputs of an element and not to the elements itself. For example if two of three inputs of an element $X_i[t]$ are correctly identified they appear in the numerator of $S_k$ but not in the numerator of $f_k$. For the latter it is necessary that all inputs of element $X_i[t]$ are correctly identified. We further investigate the positive predictive value $PPV_k$, which denotes the number of correctly identified inputs of $X_i[t]$ in the inferred network (true positives) divided by the total number of inputs of $X_i[t]$ in the inferred network (sum of true positives and false positives). $PPV_k < 1$ means that false elements have been identified. Finally we calculate the Hamming distance, $H$, as a distance measure between the links between elements in the inferred and the true networks. The Hamming distance is zero only if the inferred and the true network have the same topology. If some elements have not been found, or if false positive elements have been learned, then the Hamming distance is $> 0$. Owing to space limitations the results for this quantity are shown in the Supplement.

Simulated initial conditions: We studied situations in which (1) a limited or (2) the whole state space of network states was accessible by manipulations.

3 RESULTS OF IN SILICO SIMULATIONS

3.1 Random initial network states

Network topology: We have performed extensive computer simulations to test the parameter dependency of the network inference. We studied the fidelity $f_k$, sensitivity $S_k$ and positive predictive value $PPV_k$ as functions of the number of inputs $k$, the number of experiments $M$, the significance levels $\alpha$ and $\alpha_{\text{prior}}$, prior versus non-prior knowledge, redundancy versus non-redundancy of the network states chosen, the network size $N$, the number of manipulable states $n$, and noise versus no noise. In case of noisy data the known states before transition are assumed to have an error rate of 1% and after transition of 5%. The errors subsume errors from manipulation (usually <1% since genetically manipulated cells may be sorted and confirmed by the co-expression of linked fluorescent-protein genes), and from measuring gene expression. In the latter the noise can be largely restricted by the inclusion of suitable PCR controls (Iscove et al., 2002).

A family of typical plots of $f_k(M)$, $S_k(M)$ and $PPV_k(M)$ are shown in Figure 2 in case of prior and no prior knowledge. The fidelity $f_k$ reveals a threshold behavior as a function of $\log_{10}(M)$ and saturates at $f_k^{\text{sat}} \cdot f_k^{\text{sat}}$ reflects the degree to which a network topology can be
The parameters of plots (a) and (b) are $\alpha = 0.01$ (significance level), $n = 1$ (number of elements that can be manipulated), noise and no redundancy. (a) Fidelity $f_k$ versus $M$ ($M$ on log10-scale). For example $M = 100$ experiments we have performed 100 in silico experiments where the first gene has been fixed randomly to either 0 or 1, 100 experiments, where the second gene has been fixed, etc. The dotted lines denote inference including prior knowledge, the full lines for the second test, the PPV $\text{sat}$, also increase to 1. The effect of prior knowledge is that the fidelity saturates at 1 for each $k$ at still a high sensitivity at smaller $M$. Note that in case of noise in simulation data ('np' denotes no prior, 'p' prior and 'red' redundancy). The inset of (c) shows the sensitivity $S_k$ versus $M$ ($M$ on a log10-scale). For large $M$, $S_k$ saturates at $S_k = 1$ for all $k$. The saturation occurs earlier with decreasing values of $k$. Prior knowledge results in an earlier saturation. (c) Condensed information on PPV $k$. The parameters are $\alpha = 0.01$, $n = 2$, noise and no redundancy. The bars denote the 25, 50 and 75% quantiles and include the PPV $\text{sat}$, which we defined as the number of experiments at which $f_k = 0.9/\alpha$, increases approximately exponentially fast with $k$. This permits to conclude $M_{\text{th}}$ for larger $k$. Note that in case of $n = 1$, $\alpha = 0.01$ the fidelity does not converge to 1. In Section 3.2 we chose $\alpha = 0.001$, because then the fidelity saturates again at 1 for each $k$ at still a high sensitivity at smaller $M$. As shown in Figure 2, the sensitivity $S_k$ and the positive predictive value PPV $k$ also increase to 1. The effect of prior knowledge is that the network is inferred at lower $M$ (Fig. 2a). The simple threshold behavior of $f_k(M)$ which may well be fitted by $f_k(M) = f_k^\text{sat} (1 + \tanh(\log_{10}(M - A_1)/A_2))/2$ suggests quantification of the inference by the threshold $M_{\text{th}}$, which we defined as the number of experiments necessary to obtain a fidelity of $0.9 f_k^\text{sat}$, i.e. $f_k^\text{sat}(M_{\text{th}}) \approx 0.9 f_k^\text{sat} \Rightarrow M_{\text{th}} = f_k^\text{sat}^{-1}(0.9 f_k^\text{sat})$. $f_k^\text{sat}$, $A_1$, $A_2$ are fit parameters. The resulting plot (Fig. 2d) suggests $M_{\text{th}}$ may be approximated by

$$M_{\text{th}} \approx A \exp(Bk),$$

735
where \( \beta \) depends on the inference parameters. This permits estimation of \( M_{n_0} \) for larger \( k \) given \( M_{n_0} \) at smaller \( k = 2, 3 \) is known. The shape of \( f_{k-1}(M) \) often deviates slightly from the shapes for \( f_{k-1} \).

Below we summarize our main findings:

1. \( M_{n_0} \) decreases if \( n \) is increased or if prior knowledge is used (Fig. 2a and d). These tendencies are observed for all \( \alpha \in [10^{-5}, 0.01] \) (Fig. 2d). Decreasing \( \alpha \) increases \( M_{n_0} \) via \( \Lambda \) (see Eqn. (7)) both in the presence and in the absence of noise.

2. \( M_{n_0} \) increases slightly for redundancy, which is the expected experimental situation (Fig. 2d). Redundancy also shifts the saturation of \( S_k \) (which is still at 1) towards larger values of \( M \) for both the presence and the absence of prior knowledge.

3. For redundancy, a perfect inference is not achieved if \( \alpha \) is chosen too large. In our computer simulations we found for \( \alpha = 0.01 \) and \( n = 1 \) that the fidelity saturates at \( f_{k}^{\text{sat}} \approx 0.9 \). In this case the PPV_{1} does not saturate at \( PPV_{k}^{\text{sat}} = 1 \) at small \( k \) (we found \( PPV_{5}^{\text{sat}} = PPV_{4}^{\text{sat}} \approx 1 \), \( PPV_{3} = 0.99 \), \( PPV_{2} = 0.98 \), \( PPV_{1} = 0.95 \)). If \( \alpha \) is decreased to \( \alpha \leq 0.001 \), both, the fidelity and the PPV again saturate at 1. That is, for too large \( \alpha \) the DBN finds false positive edges for small \( k \) hence \( \alpha \) has to be chosen sufficiently small.

4. Limiting noise increases the lower threshold of \( \alpha \) (note that small \( \alpha \) increase \( M_{n_0} \)). Generally, noise reduction improves \( M_{n_0} \) and often the quality of inference.

5. For \( n = 2 \) and fixed \( \alpha \), \( M_{n_0} \) increases with decreasing \( \alpha \).

6. The addition of prior knowledge in a noisy situation can lead to an even smaller \( PPV_{k} \) value than under (5) (Fig. 2c).

For example, the saturation value of \( PPV_{k} \) in case of noise, \( n = 2 \) and prior knowledge at \( \alpha = 0.01 \) and \( \alpha_D = 0.01 \) is \( PPV_{1} \approx 0.43 \), while without prior knowledge \( PPV_{1} = 0.95 \) (Fig. 2c). However, given that the effect of the prior should vanish as \( M \rightarrow \infty \) the difference in PPV_{1} might result from the algorithm being trapped in a local optimum (by accumulating too many false positive elements). In the absence of noise in the data (which is never the case in experiments) the PPV_{1} rapidly saturates at \( PPV_{1}^{\text{sat}} = 1 \). For \( n = 1 \) and constant parameter values (noise, prior knowledge, \( \alpha = 0.01 \)), the \( PPV_{k}^{\text{sat}} = 1/k \), so that the learned network topology does not contain false positive edges.

The reason is that for \( n = 1 \) the prior knowledge helps to select the ‘correct’ edges while for \( n = 2 \), even if the correct edges have already been learned, further edges are added in order to fit the noise in the data (overfitting). Accordingly this overfitting can be compensated for by choosing a sufficiently small \( \alpha_D \) to ensure that the learned network topology does not contain false positive edges. We also performed simulations for random networks for \( N = 6, 10, 15, 20 \) for \( n = 1 \) and for selected examples for \( n = 2 \). \( M_{n_0}(N, k) \) for a given \( k \) increases only slightly with \( N \) (Supplementary Figure 3). For \( N = 6 \), \( \sim 10\% \) additional experiments were necessarily compared with the hematopoietic network in order to achieve a saturation of the fidelity. For \( n = 1 \) the saturation level is at \( f_k \approx 1 \) (Supplementary Figure 2). For the hematopoietic network example the saturation occurred at \( f_k = 1 \) only, because each element transferred information individually. At the same time the positive predictive value saturates slightly later with a saturation value that may be below 1.

However, in order to achieve maximal saturation values \( f_{k}^{\text{sat}} \) and \( PPV_{k}^{\text{sat}} \), \( \alpha \) should be decreased by an order of magnitude if the network size is increased from \( N = 6 \rightarrow N = 20 \) [which corresponds to a reduction factor of \( \approx (20/6)^2 \)]. A decrease of \( \alpha \) by one order of magnitude increases \( M_{n_0} \) by \( 20\% - 40\% \). Apart from these differences, the results for random networks were the same, including the observed relation (7).

Parameter learning: We used expectation maximization to learn the correct parameters and quantify the results by the relative entropy. Figure 3 shows the result for \( \alpha = 0.01 \), \( n = 1, 2 \), noise, no redundancy, prior and no prior knowledge. The relative entropy [see Equation (6)] is a cumulative measure and increases with the number of parents (since the number of summands increases with the number of parents). For example, in case the topologies of the inferred and true networks are identical but the expression values in the inferred network are randomly chosen to either 0 or 1, then \( H_{\text{rel}} \approx 40 \). In the case of prior knowledge, the parents are learned more quickly, so that the number of parents (and thereby the number of summands that contribute to \( H_{\text{rel}} \)) at a given \( M \) is larger than in the absence of prior knowledge. This explains why \( H_{\text{rel}} \) is larger with than without prior knowledge (we have confirmed that the individual summands are not larger with than without prior knowledge). Moreover, in case of \( n = 1 \) the relative entropy increases rapidly, regardless of whether prior knowledge is given or not: since due to the large number of missing values the landscape in which EM optimizes the parameters differs perceivably from the true landscape, the optima do not coincide with those in the true landscape. This is supported by the observation that in case more elements are jointly known (\( n = 2 \)) the entropy no longer shows a strong increase (Fig. 3). Convergence to the true landscape can only be obtained if the number of elements which can be jointly manipulated (and hence are known) is sufficiently large which for our example was the case for \( n = 2 \).

If one is primarily interested in a model which does not necessarily reflect the true topology but which is able to predict

Fig. 3. 95% Confidence intervals of mean of relative entropy in case of noise and no redundancy. The means were estimated from 1000 DBNs. \( H_{\text{rel}} \) is calculated for each inferred DBN from a sample of 1000 transition vectors that are generated from the original BoolN. The vertical line is at \( M = 250 \).
observations correctly, then PartLearn already leads to good results if \( M \approx 250 \) and \( n = 2 \).

### 3.2 Initial network states on periodic attractors

In a true experiment the hematopoietic stem cells are assumed to be in a periodic attractor with either two or more states. For this purpose we generated networks with either 2 or \( \geq 3 \) attractor states. The number of network states can be attained by a perturbation of either one (\( n = 1 \)) or at most two (\( n = 2 \)) elements of the attractor states is very limited and usually smaller than \( 2^N \).

According to experimental feasibility we have chosen \( n = 2 \) and the parameters denoted in Figure 4. Again we assume that we are able to completely monitor the network state that follows the perturbation.

The simulated fidelity curves look very different from the case in which an arbitrary initial state could be chosen (Fig. 4). The fidelity increases almost immediately for all \( k \), saturates at the same \( M \approx 300 \) for all \( k > 1 \) but does not converge to 1. \([S_k] \) shows the same qualitative behavior as \( f_k \) (data not shown). The reason is that the limited number of initial states is frequently and repeatedly offered to the DBN so that it learns quickly the topology inherent in the transitions which start from these initial states. On the other hand if the attractors have too few states, then not enough states can be attained to ensure a complete inference. Both \( f_k \) and \( S_k \) depend significantly on the number of attractor states. For prior knowledge the saturation values were the same but \( M_{th} \) was reduced by \( \sim 20\% \) (data not shown).

For random networks of size \( N = 6 \) the saturation occurs at about the same \( M_{th} \) (Supplement). An increase of \( n \) which may be used to increase the saturation level towards its maximum possible value, however, also yields an increase of \( M_{th} \).

Increasing the number of attractor states prior to perturbation changes the fidelity curves from the shape shown in Figure 4 to those shown in Figure 2a so that \( M_{th} \) becomes different for different \( k \) (Supplementary material).

### 4 DISCUSSION

We have simulated gene network inference for small networks in the absence of extensive knowledge concerning the transition states. In order to infer the network topology we used a partial learning strategy which identifies the input of gene elements by assessing the amount of information transmitted to a gene from each gene or from groups of genes. The network topology was represented by a DBN and the joint probabilities for a given topology were used to calculate the mutual information score. Finally, an EM was used to optimize the inference parameters. This inference strategy permits the inclusion of topological prior knowledge by considering the POR [see Equation (4)] in the same way as the MI score. Our studies were guided by a hypothesized core network for hematopoietic stem cell commitment and also apply to random networks. For the hematopoietic stem cell commitment network, we found that each element transmits information individually, so that in principle one influenceable gene is sufficient to determine the whole network topology. This is not the case for random networks. We quantified the degree of knowledge on the network topology by the fraction of correctly learned inputs (which we call fidelity). We found that prior knowledge, a larger number of influenceable elements and a larger number of accessible initial network states, all decrease the number of experiments necessary to obtain the same fidelity (note that the fidelity is monotonic increasing with \( M \)). Redundancy increases the number of experiments as well as the requirement for a larger statistical significance of the resulting topology. Similar tendencies are found for the sensitivities. However, if the chosen significance value is not sufficiently small, then the positive predictive value \( PPV \) may saturate at small values \( \leq 1 \) in the case of prior knowledge for a small number of input genes, and in the presence of noise which probably reflects that the simulation is trapped in a local optimum. Prior knowledge increases the tendency to keep false positive elements in case the prior knowledge does not meet the situation found in the data.

In a real experimental situation networks may often be in attractors prior to experimental perturbation, which may largely limit the number of states accessible by a perturbation of only a small number of elements. For this case, we found a completely different shape of fidelity curves and a saturation value often far below to that found when all network states are in principle accessible. However, the accessibility of states significantly improves with the number \( n \) of gene states that can be experimentally modified and with the number of states of the attractor. As more states become accessible the fidelity curves converge to those for perfect accessibility of network states. Generally, the topology inference and the parameter learning both improve significantly with increasing \( n \). For practical use of our findings we suggest the following procedure:

1. Estimate the size \( N \) of the network of interest.
2. Determine the network state in different cells in order to estimate the size of the network attractor prior to any perturbation.
3. Determine the number of states \( n \) that can be manipulated.
4. Estimate \( M_{th} \) from Supplementary Figure 3. These data are for \( \alpha = 0.001, \alpha_D = 10^{-5} \). For larger \( N \) they result in a slightly worse \( PPV \). In order to stay with the same \( PPV \), the Bonferroni criterion may be used or alternatively, as a rule of thumb, \( \alpha \rightarrow \alpha/N(6)^2 \) (see above).
(5) For the inference the procedure explained in the paper can be used.

For random networks of $N = 6, 10, 15, 20$ elements the fidelities saturate at values $f_k$ that were smaller than one for $n = 1$ and sufficiently large $k$. However, $M_{ab}$ does not increase largely with $N$ and can still be fitted to the form denoted in Equation (7). For a given number $n$ of genes, that can be manipulated jointly, the saturation value is given by the ratio of those Boolean functions in which $n$ genes transmit information on the output to the total number of Boolean functions. Here, increasing $n$ significantly improves network inference.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the discussions with D. Hasenclever and M. Löfler. This work was partly supported by the Interdisciplinary Center for Clinical Research, University of Leipzig (Project N02) and the grant BIZ-6 1/1 from the DFG.

Conflict of Interest: none declared.

REFERENCES


Friedman,N. and Elidan,G. Version 2.1 Llb for Windows/Linux 2.1.


