International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:1, No:10, 2007

Influence of Noise on the Inference of Dynamic Bayesian Networks from Short Time Series

Frank Emmert Streib, Matthias Dehmer, Gökhan H. Bakır and Max Mühlhäuser

***Abstract*— In this paper we investigate the influence of external noise on the inference of network structures. The purpose of our simulations is to gain insights in the experimental design of microarray experiments to infer, e.g., transcription regulatory networks from microarray experiments. Here external noise means, that the dynamics of the system under investigation, e.g., temporal changes of mRNA concentration, is affected by measurement errors. Additionally to external noise another problem occurs in the context of microarray experiments. Practically, it is not possible to monitor the mRNA concentration over an arbitrary long time period as demanded by the statistical methods used to learn the underlying network structure. For this reason, we use only short time series to make our simulations more biologically plausible.**

***Keywords*— Dynamic Bayesian networks, structure learning, gene networks, Markov chain Monte Carlo, microarray data.**

I. Introduction

DYNAMIC Bayesian networks are a special example of graphical models that combine properties from graph and probability theory [10]. Causally speaking, graphical models allow the visualization of multivariate probability distributions where nodes in a graph represent random variables and connections between nodes indicate dependencies between the random variables [1]. In recent years, Bayesian networks and dynamic Bayesian networks, which are an extention of Bayesian networks in the respect that the underlying directed graph can contain cycles, are used to analyze gene expression data from microarray experiments [4], [13], [6], [8]. Especially, dynamic Bayesian networks seems to be a good choice for this task, because gene networks, e.g., transcription regulatory networks, contain positive and negative feedback loops as Lee et al. demonstrated for yeast [11].

The major objective of this paper is to investigate the influence of external noise on the inference of dynamic Bayesian networks from short time series. This is important, because external noise can be seen as measurement error in microarray experiments which is inevitably present. Hence, we do not deal with principle questions about the learnability of a probabilistic model under ideal conditions, but tackle the practical question if a dynamic Bayesian network is appropriate to be applied in the context of biological data from microarray experiments. To obtain an objective performance measure we generate a short time series by a model that incorporates important features of the biological system. This allows us to evaluate the performance objectively, because the solution, the network structure, is known. To our knowledge, our results are the first in this direction to study this influence systematically. Existing studies in this context investigated, e.g., the appropriate level of description to simulate gene expression data, the influence of the number of time points, the number of categories and the interval length between samples [2], [17], [18], [8], [16].

This paper is organized in the following way: In the next section we present the model we use to generate biological plausible data mimicing the process of, e.g., transcription regulation. In II-C we describe the mathematical framework of dynamical Bayesian networks we use to infer the network structure. In section III we present our results and in IV we finish the article with a discussion and conclusions.

Frank Emmert-Streib is with the Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO 64110, USA, e-mail: fes@stowers-institute.org. Matthias Dehmer is with the Technische Universität Darmstadt, 64289 Darmstadt, Germany, e-mail: dehmer@informatik.tu-darmstadt.de. Gökhan Hasan Bakır is with the Max Planck Institute for Biological Cybernetics, Spemannstrasse 38, 72076 Tübingen, Germany, e-mail: goekhan.bakir@tuebingen.mpg.de. Max Mühlhäuser is with the Technische Universität Darmstadt, 64289 Darmstadt, Germany, e-mail: max@informatik.tu-darmstadt.de.

II. Model

A. Boolean network with external noise

We generate a binary time series $X_i^t \in \{0,1\}$ for all nodes $i \in N$ from a Boolean network consisting of $N = 12$ nodes. The structure of this network is shown in Fig. 1 [7]. The Boolean functions f_j defining the dynamics in this network are deterministic, however, the nodes 2 and 7 receive a random input. That means, the values of these nodes X_2^t and X_7^t are at each time step t randomly chosen from $\{0,1\}$. This prevents the systems dynamics eventually to reach a fixed point after a certain number of time steps. The three different logical functions f_j used are a or-gate, a not-gate and a one-gate which leaves the signal unchanged and just copies the value of the input gene to the output gene. The or-gate is used for all

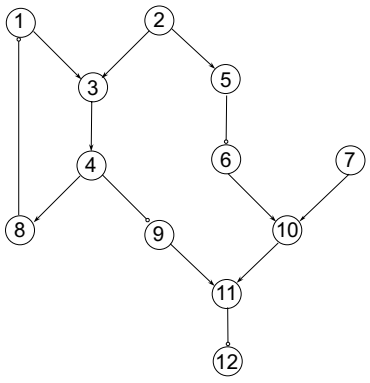

Fig. 1. Network topology of our synthetic network. Arrows represent an excitation between genes and circles an inhibition.

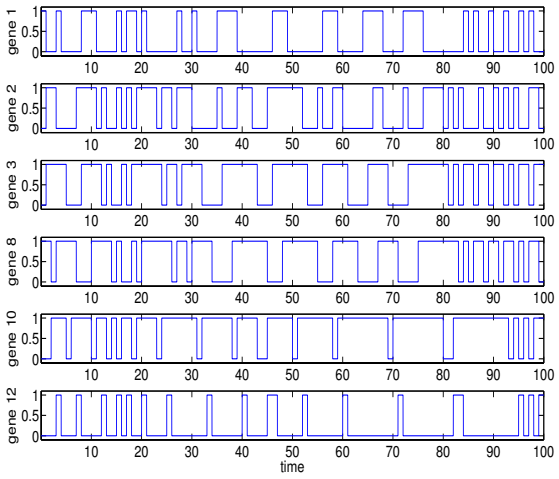

Fig. 2. Time series of expression values for six genes generated with the network structure shown in Fig. 1.

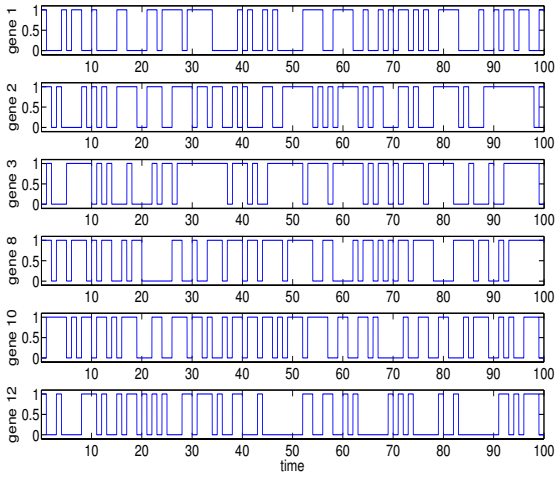

Fig. 3. Time series of expression values for six genes generated with the network structure shown in Fig. 1. The output of a gene was flipped with probability $p_\eta = 0.3$.

nodes which receive two inputs from other genes. The genes which are regulated by a not-gate are depicted in Fig. 1 by a circle and the genes which are regulated by a one-gate by an arrow. Formally, the dynamics of the system is given by

$$X_i^{t+1} = f_j^i(Par(X_i), t) \tag{1}$$

for all $i \in \{1, \ldots, N\}$. That means the value of gene i at time step $t+1$ is determined by the j-th boolean function which has as inputs the parents $Par(X_i)$ of gene i with their values at time step t. Additionally to the dynamics given by the Boolean functions we introduce external noise in the system. We model the influence of external noise on the systems dynamics by flipping the value of a gene X_i at each time step with a probability p_ϵ. This effect is called external noise, because we include by this measurement errors inevitably present in any experiment.

In Fig. 2 we show a resulting time series of length $T = 100$ from the dynamics defined above for 6 genes. In this case the external noise was $p_\epsilon = 0$. One can see that the activity of gene 3 and gene 8 are the same up to a delay of 2 time steps and that gene 1 follows inversely the activity of gene 8. The low activity of gene 12 is a direct consequence of the inhibition of gene 11. To see the influence of the external noise on the activity of the genes we show in Fig. 3 an example for $p_\epsilon = 0.3$. In this case the activity frequency of gene 12 is increased considerably.

B. Stochastic Boolean network with external noise

The second dynamics we study is a network consisting of stochastic Boolean functions. For example, a stochastic or-gate is given by the assignment in table I. The means, every Boolean function is substituted by a conditional probability. For reasons of simplicity we assume always that the value of the deterministic gate is assumed with probability $p(y|Par(y)) = p_1$. By this convention we have only one free parameter p_1 which controls the stochasticity of all conditional probability distributions in the network. Also in this case we introduce external noise in the system by the same mechanism described in section II-A. For the following simulations we use $p_1 = 0.9$ and the network topology shown in Fig. 1.

TABLE I

STOCHASTIC OR-GATE. THE OUTPUT VALUE y IS GIVEN WITH A CERTAIN PROBABILITY $p(y)$. THIS PROBABILITY CORRESPONDS TO THE CONDITIONAL PROBABILITY $p(y|x_1, x_2)$ GIVEN IN THE TABLE. FOR SIMPLICITY WE ASSUME THAT THE VALUE OF THE DETERMINISTIC OR-GATE IS ALWAYS CHOSEN WITH PROBABILITY p_1.

y	x_1	x_2
$p(y=0) = p_1$	0	0
$p(y=1) = p_1$	1	0
$p(y=1) = p_1$	0	1
$p(y=1) = p_1$	1	1

To make the following simulations more realistic we add to the network 30 genes whose dynamics is randomly and uniformly drawn from $\{0, 1\}$. These genes serve as distractors for our learning algorithm. Biologically, this corresponds to, e.g., microarray experiments which include genes that are not involved in the process under investigation. The problem is, that it is not known in advance which genes are relevant for a certain biological process and, hence, they can not be excluded for the analysis per se. This is another difficulty normally neglected in the study of dynamic Bayesian networks.

C. Dynamic Bayesian Networks

A Bayesian network $\mathcal{M}$ is a graphical model in form of a directed acyclic graph (DAG) $\mathcal{G}$ together with conditional probability distributions, depending on parameters Θ, for each node i in the graph that depends only on its parents, $P(n_i|Pa^{\mathcal{G}}(n_i))$ [15]. This provides a graphical representation

of the joint probability distribution of N random variables n_i by

$$P(n_1, n_2, \ldots, n_N) = \prod_i^N P(n_i | Pa^{\mathcal{G}}(n_i)) \quad (2)$$

In the context of genetic networks we identify the random variables n_i with (discrete) expression values X_i of genes and connections between random variables as interactions. The problem we are facing is to infer the structure of the network $\mathcal{G}$ from given data $\mathcal{D}$, that means we want to maximize the conditional probability $P(\mathcal{G}|\mathcal{D})$.

$$\mathcal{G}^* = \underset{\mathcal{G}}{\operatorname{argmax}}\{P(\mathcal{G}|\mathcal{D})\} \quad (3)$$

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G}) \quad (4)$$

The optimal network structure is denoted by $\mathcal{G}^*$ and the posterior distribution $P(\mathcal{G}|\mathcal{D})$ is given via the Bayes rule in Eq. 4 up to a normalizing factor. The likelihood $P(\mathcal{D}|\mathcal{G})$ is obtained by integrating over the parameters of the conditional probabilities Θ by

$$P(\mathcal{D}|\mathcal{G}) = \int P(\mathcal{D}|\Theta, \mathcal{G})P(\Theta|\mathcal{G})d\Theta \quad (5)$$

It was suggested [8] that the maximum a-posteriori (MAP) approach Eq. 3 is not the most efficient if the available data are incomplete. Instead, sampling from the posterior probability Eq. 4 results in a collection of networks with comparable quality rather than just in a single network [8]. The problem with this approach is that sampling from the posterior is not directly possible because the denominator can only be calculated if the size of the graph is very small. However, this can be overcome by applying a *Markov chain Monte Carlo* simulation (MCMC) [12]. Here we use the algorithm of *Metropolis-Hastings* (MH). This algorithm is based on local modifications of the old structure $\mathcal{G}_{old}$ leading to a new structure $\mathcal{G}_{new}$. Possible local modifications are to delete, reverse or add an edge to the graph. If the new structure is accepted or rejected is decided based on the following criterion,

$$p_{accept} = min\left\{1, \frac{P(\mathcal{G}_{new}|\mathcal{D})}{P(\mathcal{G}_{old}|\mathcal{D})}\frac{T(\mathcal{G}_{old}|\mathcal{G}_{new})}{T(\mathcal{G}_{new}|\mathcal{G}_{old})}\right\} \quad (6)$$

The transition probabilities $T(\mathcal{G}'|\mathcal{G})$ are given by $1/\#\mathcal{G}$. Here $\#\mathcal{G}$ denotes the number of possible structures which can be obtained by the allowed local modifications (delete, reverse or add an edge). For more technical details about the algorithms the reader is referred to HUSMEIER [8].

So far we discussed only Bayesian networks. This class of graphical models is restricted to acyclic graphs as mentioned above. However, one characteristic property of genetic networks is that they can contain feedback loops. For example in Fig. 1 the genes $1, 3, 4$ and 8 are forming a feedback loop. This limitation of Bayesian networks can be overcome by using *dynamic Bayesian networks* [3]. Dynamic Bayesian networks are directed graphs together with conditional probability distributions which can contain cycles. Practically, we solve the problem to determine the structure of the network which fits best to the data by unfolding the dynamic Bayesian network in time. This results in a normal Bayesian network that can be treated in the way described before.

III. RESULTS

Our major objective in this paper is to study the influence of external noise on the inference of dynamic Bayesian networks from short time series to see if this mathematical framework is suitable to be applied to experimental data from microarray experiments. This is an important question, because external noise which models measurement errors are inevitable in any kind of experiments and especially in biological experiments. In the following simulations we used the network structure consisting of 12 genes as shown in Fig. 1 together with 30 random genes serving as distractors (not shown in the figure) and two different dynamics for the gene activities as described in section II-A and II-B. We corrupted the dynamics of both models by external noise by inverting the activity of each gene X_i^t at each time step t with probability p_ϵ. From these models we generated time series of length $T = 100$ to infer the network structure with a dynamic Bayesian network.

A. Influence of external noise on the overall performance

The results for the Boolean network are shown in Fig. 4 and for the stochastic Boolean network in Fig. 5. The rows in each figure correspond the the amount of external noise $p_\epsilon = \{0.1, 0.3, 0.4\}$ used to disturb the dynamics. The results are visualized by the *receiver operator characteristics* (ROC) curves. A ROC curve plots the sensitivity $= TP/(TP + FN)$ against the complementary specificity $= 1-TN/(TN+FP) = FP/(TN+FP)$. Due to the fact, that we approximated the posterior distribution $P(\mathcal{G}|\mathcal{D})$ in Eq. 4 by MCMC simulation rather than determined its corresponding MAP we have only probabilities for the presence of an edge in a network [8]. That means, we have to choose a threshold $\gamma \in [0, 1]$ if we decide to accept an edge,

$$W_{ij} = \begin{cases} 1 & : \quad P(W_{ij}|\mathcal{D}) \geq \gamma \\ 0 & : \quad P(W_{ij}|\mathcal{D}) < \gamma \end{cases} \quad (7)$$

Hence, the sensitivity as well as the complementary specificity depend on γ implicitly. With other words, the ROC curves shown in Fig. 4 and 5 are parameterized by γ. The diagonals shown as dashed line in Fig. 4 and 5 correspond to a completely random prediction.

From the top and middle Fig. 4 one can see that the Boolean networks can almost completely be reconstructed for external noise up to $p_\epsilon = 0.2$. This is remarkable if one bears in mind that the time series used to extract this information from consisted of only 100 time steps. For $p_\epsilon = 0.3$ the performance dropped, but is still much better than a random predictor as indicated by the dashed line. These results can be viewed as ideal cases which give theoretical bonds to what can be expected from more realistic models or even experimental data themselves, because of the underlying deterministic dynamics of the genes. The results for the more biologically realistic model, the stochastic Boolean network, are shown in Fig. 5. The performance decreases evidently. For $p_\epsilon = 0.1$ we obtain still good results, because one gets, e.g.,

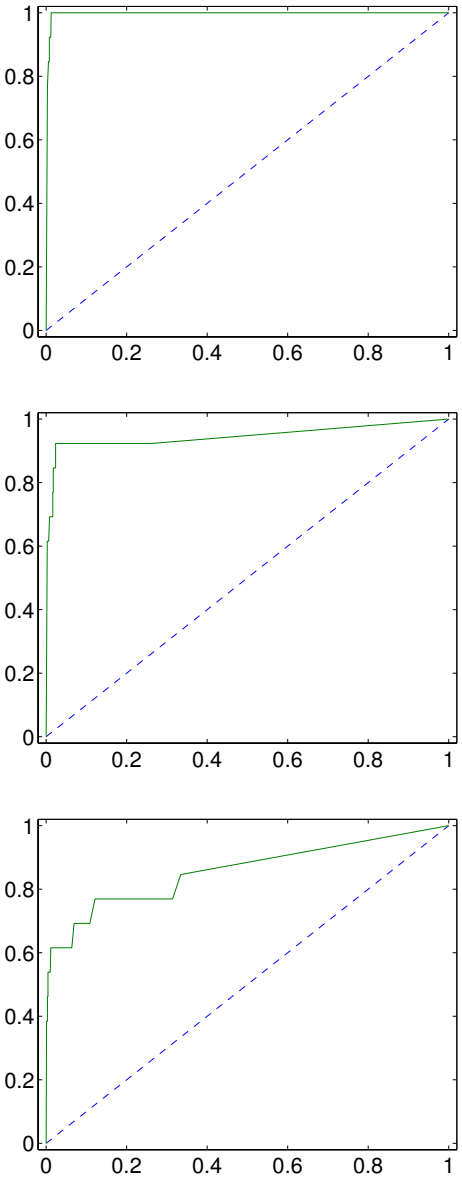

Fig. 4. Sensitivity in depends on the complementary specificity for a Boolean network with 12 connected genes and additionally 30 genes serving as destructors. The structure of the connected network is shown in Fig. 1. Different rows correspond to different values of external noise p_ϵ. Top: $p_\epsilon = 0.1$. Middle: $p_\epsilon = 0.2$. Bottom: $p_\epsilon = 0.3$. The time series used to infer the network structure was $T = 100$ time steps long.

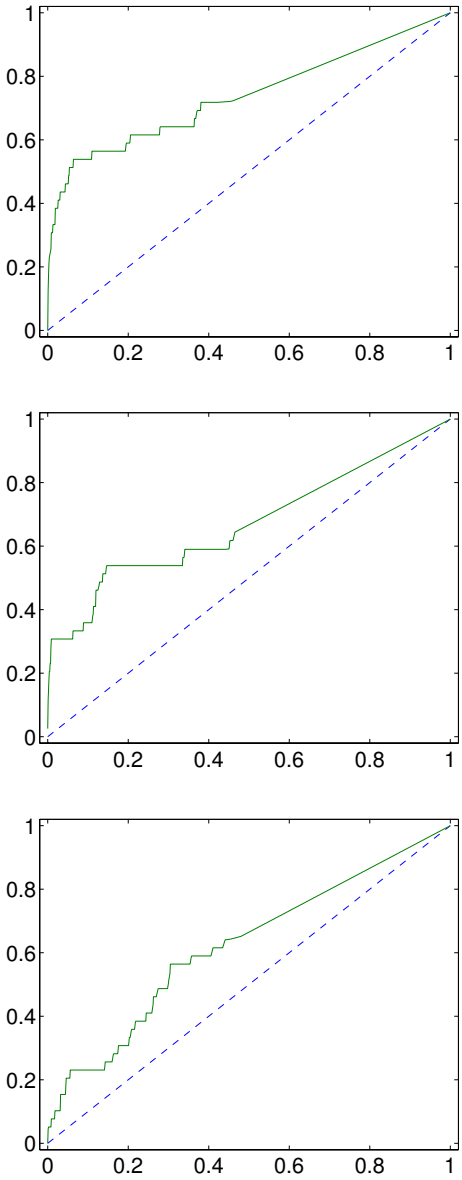

Fig. 5. Sensitivity in depends on the complementary specificity stochastic Boolean network with 12 connected genes and additionally 30 genes serving as destructors. The structure of the connected network is shown in Fig. 1. Different rows correspond to different values of external noise p_ϵ. Top: $p_\epsilon = 0.1$. Middle: $p_\epsilon = 0.2$. Bottom: $p_\epsilon = 0.3$. The time series used to infer the network structure was $T = 100$ time steps long.

about 55% sensitivity for only 10% complementary specificity. This means the number of true positives - correctly predicted edges in the network - is much higher than the number of false positives - incorrectly predicted edges in the network. From an experimental point of view this enables the possibility to test some predicted edges, by other experimental methods, with a high probability to confirm these results. For increasing values of the external noise p_ϵ the results getting worse. However, even for $p_\epsilon = 0.3$ the recovered network structure is clearly better than random.

B. Influence of external noise on edge-probabilities

Now, we want to take a closer look to the obtained results by studying the edge-probabilities $p(E_{i,j}|\mathcal{D})$ of the posterior distribution in Eq. 4. In table II and III we give these values for all 13 edges in the graph shown in Fig. 1. For the Boolean network and $p_\epsilon = 0.1$ all edges probabilities are very high as expected. For $p_\epsilon = 0.2$ the values are still high except for $E_{10,7}$. There are only four edges which are predicted for all noise cases with 100%, $E_{1,8}, E_{4,3}, E_{6,5}$ and $E_{9,4}$. None of these edges contributes to an or-gate, but three of them ($E_{1,8}, E_{6,5}$ and $E_{9,4}$) to a not-gate. Interestingly,

TABLE II

EDGE-PROBABILITIES $E_{i,j}$ OF THE POSTERIOR DISTRIBUTION IN EQ. 4 FOR THE NETWORK WITH BOOLEAN FUNCTIONS TO CONNECT GENE i WITH GENE j. THE THREE COLUMNS CORRESPOND TO $p_\epsilon = \{0.1, 0.2, 0.3\}$ (SECOND, THIRD, FORTH COLUMN).

$E_{1,8}$	1.0000	1.0000	1.0000
$E_{3,1}$	0.9485	0.6818	0.1371
$E_{3,2}$	1.0000	0.9087	0.1948
$E_{4,3}$	1.0000	1.0000	1.0000
$E_{5,2}$	0.7429	1.0000	1.0000
$E_{6,5}$	1.0000	1.0000	1.0000
$E_{8,4}$	1.0000	1.0000	0.6700
$E_{9,4}$	1.0000	1.0000	1.0000
$E_{10,6}$	1.0000	1.0000	0.4664
$E_{10,7}$	1.0000	0	0.0118
$E_{11,9}$	1.0000	0.4740	0
$E_{11,10}$	1.0000	0.6134	0
$E_{12,11}$	0.8637	1.0000	0.7242

TABLE III

EDGE-PROBABILITIES $E_{i,j}$ OF THE POSTERIOR DISTRIBUTION IN EQ. 4 FOR THE NETWORK WITH STOCHASTIC BOOLEAN FUNCTIONS TO CONNECT GENE i WITH GENE j. THE THREE COLUMNS CORRESPOND TO $p_\epsilon = \{0.1, 0.2, 0.3\}$ (SECOND, THIRD, FORTH COLUMN).

$E_{1,8}$	0.0302	0.0471	0.0255
$E_{3,1}$	0.3256	0.4826	0.1585
$E_{3,2}$	1.0000	0.9896	0.0774
$E_{4,3}$	0.2062	0.0569	0.0190
$E_{5,2}$	0.2034	0.1264	0.0218
$E_{6,5}$	0.1861	0.0521	0.0824
$E_{8,4}$	0.0440	0.0013	0.0126
$E_{9,4}$	0.0056	0	0.0047
$E_{10,6}$	0.4332	0.7617	0.4345
$E_{10,7}$	1.0000	0.7118	0.7067
$E_{11,9}$	0.7455	0.3333	0.0917
$E_{11,10}$	0.9832	0.3726	0.1135
$E_{12,11}$	0.0041	0.0551	0.0413

exactly these three edges are among the lowest probabilities for the stochastic Boolean network in table III and the edge-probabilities for edges contribution to an or-gate are among the highest. This raises the question if this is a systematic effect or happened just by change in the small network structure used.

IV. CONCLUSIONS

In this paper we examined the influence of external noise on the inference of dynamic Bayesian networks from short time series. Here external noise represents perturbations of the observed systems dynamic, e.g., due to measurement errors inevitably present in every experiment. Hence, we were dealing with the principle question if dynamic Bayesian networks are suitable to be applied under such conditions. Practically, these conditions correspond, e.g., to data from microarray experiments which allow to monitor the mRNA concentration of thousands of genes simultaneously within a cell. In this context the reconstructed network structure corresponds to a gene network, e.g., the transcription regulatory network.

We found that increasing values of external noise reduces significantly the overall performance. However, the results are still much better than random and provide by this reasonable predictions which can be tested in further experiments. This holds even for the stochastic Boolean networks which provide certainly a better abstraction of the real biological processes than the deterministic Boolean networks. Interestingly, we found that the logical functions (deterministic or stochastic) can be learned with different precision in both settings. For example, the not-gate can be learned best for Boolean networks, but worst for stochastic Boolean networks. This could have important consequences for inferring transcription regulatory networks, because this implies that there is a bias for certain types of transcription regulation logic which causes difficulties to be inferred from the data for theoretical reasons.

ACKNOWLEDGMENT

We would like to thank Kevin Patrick Murphy [14] and Dirk Husmeier [9] for providing freely MatLab software.

REFERENCES

[1] Buntine, W.L.: Operations for Learning with Graphical Models. Journal of Artificial Intelligence Research **2** (1994) 159–225.
[2] Chen, T., He., H.L., Church, G.M.: Modeling gene expression with differential equations. Pac. Symp. Biocomput. **4** (1999) 29–40.
[3] Friedman, N., Murphy, K., Russel, S.: Learning the Structure of Dynamic Probabilistic Networks. In Cooper, G.F. and Moral, S. (eds), Proceedings of the Fourteenth Conference on Uncertainty in Artifical Intelligence (UAI). Morgan Kaufmann Publishers, San Francisco, CA (1998).
[4] Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian Networks to Analyze Expression Data. Journal of Computational Biology **7:3/4** (2000) 601–620.
[5] Gardner, T.S., di Bernardo, D., Lorenz, D., Collins, J.J.: Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. Science **301** (2003) 102–105.
[6] Hartemink, A.J., Gifford, D., Jaakkola, T., Young, R.: Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac. Symp. Biocomp. **6** (2001) 422–433.
[7] Hartemink, A.J.: Reverse engineering gene regulatory networks. Nature **23:5** (2005) 554–555.
[8] Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics **19:17** (2003) 2271–2282.
[9] Musmeier, D.: Inferring Dynamic Bayesian Networks with MCMC (DBmcmc). www.bioss.sari.ac.uk/∼dirk/software/DBmcmc/ (2003).
[10] Jordan, M.I.: Learning in Graphical Models. MIT Press (1998).
[11] Lee, T.I. et al.: Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science **298** (2002) 799-804.
[12] Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer-Verlag, New York (2001).
[13] Murphy, K.P., Mian, S.: Modelling Gene Expression Data using Dynamic Bayesian Networks, Technical Report (1999).
[14] Murphy, K.P.: Bayes Net Toolbox. Technical Report, MIT Artificial Intelligence Laboratory (2002).
[15] Perl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco, CA, USA (1988).
[16] Schlitt, T., Brazma, A.: Modelling gene networks at different organizational levels. FEBS Letters **579** (2005) 1859–1866.
[17] Smith, V.A., Jarvis, E.D., Hartemink, A.J.: Evaluating functional network inference using simulations of complex biological systems. Bioinformatics **18** (2002) 164–175.
[18] Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D.: Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics **20:18** (2004) 3594–3603.