World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:1, No:10, 2007

First Studies of the Influence of Single Gene Perturbations on the Inference of Genetic Networks

Frank Emmert-Streib and Matthias Dehmer

Abstract**—** Inferring the network structure from time series data is a hard problem, especially if the time series is short and noisy. DNA microarray is a technology allowing to monitor the mRNA concentration of thousands of genes simultaneously that produces data of these characteristics. In this study we try to investigate the influence of the experimental design on the quality of the result. More precisely, we investigate the influence of two different types of random single gene perturbations on the inference of genetic networks from time series data. To obtain an objective quality measure for this influence we simulate gene expression values with a biologically plausible model of a known network structure. Within this framework we study the influence of single gene knock-outs in opposite to linearly controlled expression for single genes on the quality of the infered network structure.

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

I. Introduction

CONSIDERABLE progress has been made in the last decade in the understanding of the molecular biological processes underlying life. This can be attributed first of all to the technological advances rather than to theoretical break-throughs which could provide a general mathematical framework for living-matter comparable to our knowledge in physics. The technological advances manifest in two major streams, experimental and computational. The experimental technologies in modern molecular biology, e.g., microarrays, proteomics, ChIP-chip, allow nowadays to monitor the behavior of, e.g., the gene expression, on a systems level. Systems level means, that it is in principle possible to measure, e.g., the gene expression, of all genes within an organism and not only some dozen. The resulting amount of information gathered by these experimental technologies can not be processed without high-performance computers and efficient algorithms. Two prominent example of such computer-intensive methods are the bootstrap algorithm [2] to calculate, e.g., standard errors or confidence intervals and *Markov chain Monte Carlo* simulations [11] to sample, e.g., from probability distributions that are very high dimensional and are otherwise intractable. The sequential application of these technics, first the experiment then the computational data analysis, experienced limitations in the application to the inference of genetic networks based on high-throughput data [4], [7] due to the inherent difficulty of the problem.

In this paper we tackle the problem how to design microarray experiments so that computational methods can extract more information form the resulting data of the biological process under investigation. More precisely, we focus on microarray experiments and computational methods to infer the network topology of genetic networks, e.g., transcription regulation networks. We investigate the influence of different types of random single gene perturbations on the quality of the infered network structure. To obtain reliable results we generate the time series of the expression values of a genetic network with a biologically plausible model. This ensures that we know the network structure of the genetic network we want to infer and gives us a clear criterion to judge the estimated network structure. To our knowledge, our results are the first in this direction. 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 [1], [15], [16], [8], [14].

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-B 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.

II. Model

A. Generation of gene expression data

We generate the gene expression values X^t with a linear model which is very similar to the model suggested by Yu et

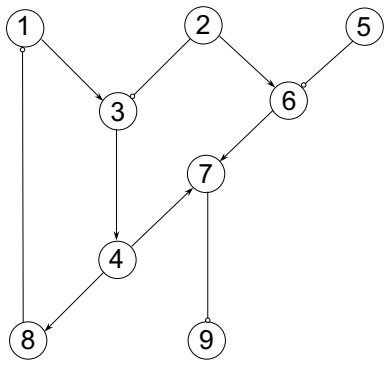

Fig. 1. Network topology of our synthetic network. Arrows represent an excitation between genes and circles an inhibition.

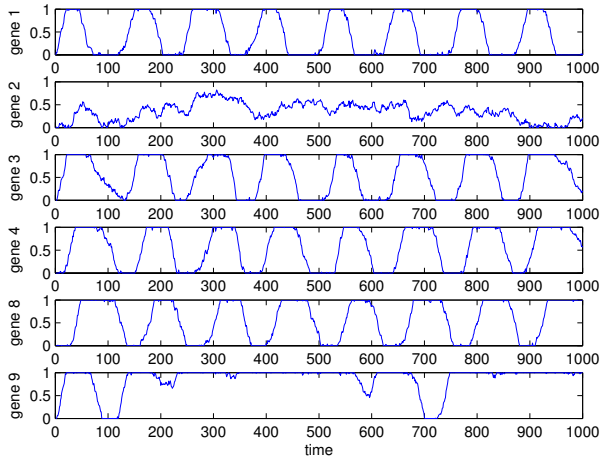

Fig. 2. Time series of expression values for six genes generated with the network structure shown in Fig. 1.

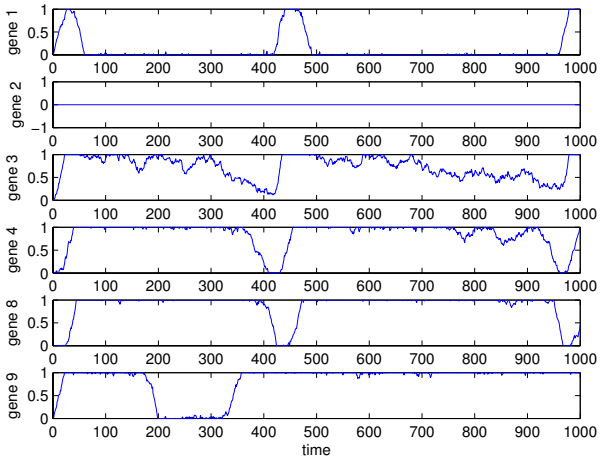

Fig. 3. Time series of expression values for six genes generated with the network structure shown in Fig. 1. Gene 2 was knocked-out.

al. [16].

$$X_i^{t+1} = X_i^t + \Delta X_i^t \quad (1)$$

$$\Delta X_i^t = \sum_j^{N_g} \delta W_{ij}(X_j^t - B_j) + \epsilon_i^t \quad (2)$$

We assume discrete time points t and the expression values X^t of the N_g genes are restricted to the continuous interval $[0, 1]$. This gives the interpretation that gene i is expressed at time step t if $X_i^t = 1$ or it is not expressed if $X_i^t = 0$. The expression values of the genes are updated at every time step by Eq. 2. The influence of genes on each other is defined by the coupling matrix W. We assume only three possible interactions: 1. excitation $W_{ij} = 1$, 2. inhibition $Wij = -1$ or 3. independence $Wij = 0$. For reasons of simplicity the strength of an excitation or an inhibition shall be the same indicated by δ. The vector B contains threshold values, e.g., for $W_{ij} > 0$ gene j can only up-regulate gene i if $X_j^t - B_j > 0$ otherwise gene i gets down-regulated. The last term in Eq. 2 ϵ_i^t represents Gaussian white noise independently drawn for each component at each time step from $G(\mu = 0, \sigma = 3.0)$. This is in contrast to Yu et al. [16], because they used noise drawn from a uniform distribution. The difference between both noise models might not be dramatic, however, we prefer Gaussian noise because of absence of more detailed information about the real situation and its omnipresence in nature. As connectivity between the genes we use two different networks. One is shown in Fig. 1. This network consists of 9 connected genes. The arrows indicate excitation the circles inhibition from one gene to another. The second network has the same structure but additionally we include 11 genes which have no connections to other genes at all. This represents a distraction corresponding to gene in a microarray that are not involved in the process under investigation. In Fig. 2 we show a simulated time series of gene expression values for six genes. The time series of gene 2 is just a random walk, because it does not interact with any other gene. The feedback loop consisting of genes $1, 3, 4$ and 8 leads to a periodic expression pattern of these

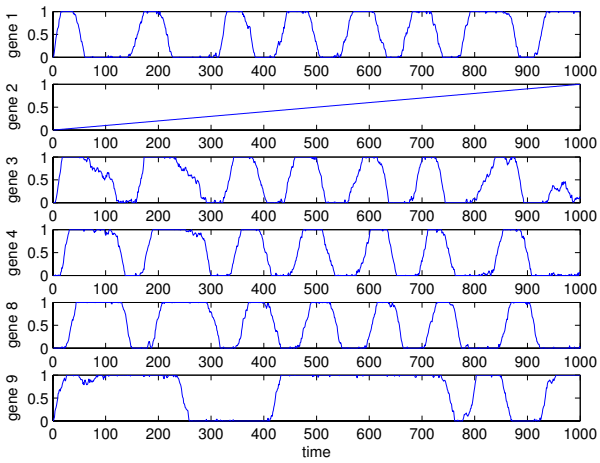

Fig. 4. Time series of expression values for six genes generated with the network structure shown in Fig. 1. Gene 2 was linearly increased from 0 to 1 in 1000 time steps.

genes, because gene 8 represses the expression of gene 1. The outcome of gene 9 is less obvious, because it receives indirect input from two genes $(2, 5)$ with random activity.

In this paper we want to investigate the influence of two different types of perturbations on the gene expression on the inference of the underlying genetic structure. We perturb the gene expression by single gene knock-out and by controling the expression linearly from no expression $X_i = 0$ to expression $X_i = 1$. This is inspired by the genetic toggle switch introduced by GARDNER et al. [5]. Fig. 3 and 4 show the corresponding time series for two examples. In Fig. 3 gene 2 is knocked-out. The periodic patterns of the resulting time series are dramatically changes. This is because gene 2 can no longer inhibit the expression of gene 3. In Fig. 4 the expression of gene 2 was linearly increased over time. Here one can see the opposite effect, high X_2 values repress gene 3.

B. 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 depend only on its parents, $P(n_i|Pa^{\mathcal{G}}(n_i))$ [13]. This provides a graphical representation of the joint probability distribution of N random variables n_i by

$$P(n_1, n_2, \ldots, n_N) = \prod_i^N P(n_i|Pa^{\mathcal{G}}(n_i)) \tag{3}$$

In the context of genetic networks we identify the random variables n_i with (discretized) 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 $P(\mathcal{G}|\mathcal{D})$.

$$\mathcal{G}^* = \underset{\mathcal{G}}{\operatorname{argmax}}\{P(\mathcal{G}|\mathcal{D})\} \tag{4}$$

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G}) \tag{5}$$

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. 5 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

$$P(\mathcal{D}|\mathcal{G}) = \int P(\mathcal{D}|\Theta, \mathcal{G})P(\Theta|\mathcal{G})d\Theta \tag{6}$$

It was suggested [8] that the maximum a-posteriori (MAP) approach Eq. 4 is not the most efficient if the available data are incomplete. Instead, sampling from the posterior probability Eq. 5 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) [11]. 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,

$$p_{accept} = min\{1, \frac{P(\mathcal{G}_{new}|\mathcal{D})}{P(\mathcal{G}_{old}|\mathcal{D})}\frac{T(\mathcal{G}_{old}|\mathcal{G}_{new})}{T(\mathcal{G}_{new}|\mathcal{G}_{old})}\} \tag{7}$$

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 refered 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

The major objective of this paper is to study the influence of two different gene perturbation strategies on the quality of the infered network structure. To make our simulations biologically realistic we allow only the observation of short time series (20 time points) and the perturbation of only 5 genes. More precisely, e.g., we knock-out gene 4 then we simulate the gene expression levels according to Eq. 1 and 2 for 100 time steps but observe the expression values X_i only every 5-th time step. This results in 20 measurement that are used to infer the network structure. Because microarrays do not only contain genes which are relevant for a certain biological process under investigation, but contain also a certain number of genes which are not involved in the pathway we want to infer, we repeat our simulations for the identical network structure in Fig. 1, but add additionally 11 genes as destructors which have no connection to any other gene in the network.

A. Single gene perturbations

The results for the single gene knock-out experiments are shown in the first and third figure in 5 and the corresponding results for the controlled expressions in the second and forth figure in 5. The first two figures correspond to the network in Fig. 1 with 9 genes the following two includes additionally 11 genes as destructors. 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. 5 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,

$$W_{ij} = \begin{cases} 1 & : \quad P(W_{ij}|\mathcal{D}) \geq \gamma \\ 0 & : \quad P(W_{ij}|\mathcal{D}) < \gamma \end{cases} \tag{8}$$

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

AUROC values	knock-out	linear control
9 genes	0.78	0.76
20 genes	0.66	0.63

TABLE I
THE AREA UNDER THE ROC CURVE FOR THE RESULTS IN FIG. 5.

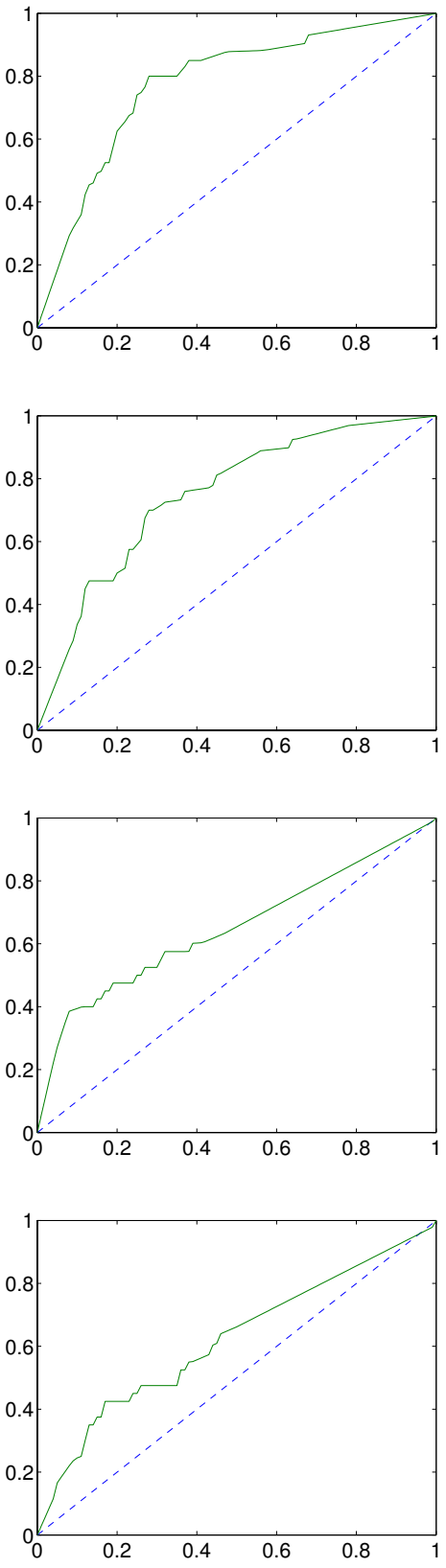

Fig. 5. From top to bottom: 1. Network with 9 genes and knock-out perturbations. 2. Network with 9 genes and linear controlled expression perturbations. 3. Network with 20 genes and knock-out perturbations. 4. Network with 20 genes and linear controlled expression perturbations.

As expected the larger the network the more difficult is the structure to infer. This can be seem by visual comparison of the figures or by calculation the values of the *area under the ROC curve* (AUROC) which are given in table I. In general, the larger the AUROC value is the better is the quality of the infered network. A comparison between the two types of perturbations reveals that the single gene knock-out experiments give slightly better results than the linear control of genes.

IV. Conclusions

In this paper we investigated the influence of two different types of random single gene perturbations on the quality of the inference of the underlying genetic network. We generated biologically realistic time series mimicing the expression values of genes and perturbed the system by knocking genes out or controlling the expression values linearly from not expressed to expressed. To infer the network structure based on the discretized time series of 20 time points and 5 different perturbations we applied MCMC simulation to estimate the posterior distribution of the network structures for the given data. Our results reveal that there is only a minor influence of the perturbation type on the quality of the inference. The single gene knock-out experiments give slightly better results for the network consisting of 9 as well as 20 genes. This could have three reasons: First, we selected the perturbed genes randomly. Intuitively, if the network structure would be know in advance one should always be able to decide what kind of perturbation provides the most information gain for the inference of the network in combination with the utilized algorithm. However, if the structure is completely unknown how could we make such a decision? For example, Gardner et al. [6] demonstrated for the known SOS pathway in *E. coli* that induced over-expression of single genes is sufficient to infer the underlying structure to a high degree. Second, the linear expression control over the complete time interval of the measurement is not too different from the knock-out perturbation. At the beginning the expression values might be comparable in its influence on other genes with a knock-out whereas at a later stage (after about $4/5$ of the observation time, see Fig. 4) the gene appears to other genes as completely expressed. However, this time period could be too short to have an efficient influence. To investigate this effect one could study the influence of a ramp function with varying slope. Third, there is no difference between both types of perturbations. We are of the opinion that for an sufficient amount of data available about the underlying system there should be not much difference between both perturbation strategies. However, data about biological systems, e.g., transcription regulation networks, do not fulfill this condition, because the possible state space of expression values can only be sampled sparsely. In this case we would expect differences. It might be possible that the number of different perturbations used (5) in contrast to the total number of genes in the network (9 and 20) is too high bringing us near to the situation of sufficient data. Further studies about networks containing more connected genes and the influence of the number of experiments on the quality of network inference will shed light on this point.

Our results are the first investigating the influence of the perturbation strategy on the quality of the infered genetic network for time series data from biologically plausible simulations. This completes studies about 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 [1], [15], [16], [8], [14]. In general, we think that simulation studies can help to design more efficient high-throughput experiments leading themselves to a more efficient computational analysis of the resulting data.

ACKNOWLEDGMENT

We would like to thank Wing Hung Wong for fruitful discussions and Kevin Patrick Murphy [12] and Dirk Husmeier [9] for providing freely MatLab software for Bayesian Networks.

REFERENCES

[1] Chen, T., He., H.L., Church, G.M.: Modeling gene expression with differential equations. Pac. Symp. Biocomput. **4** (1999) 29–40.
[2] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall/CRC (1994).
[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., Cantor, C.R., Collins, J.J.: Construction of a genetic toggle switch in Escherichia coli. Nature **403:20** (2000) 339–342.
[6] 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.
[7] 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.
[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] Lee, T.I. et al.: Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science **298** (2002) 799-804.
[11] Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer-Verlag, New York (2001).
[12] Murphy, K.P.: Bayes Net Toolbox. Technical Report, MIT Artificial Intelligence Laboratory (2002).
[13] Perl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco, CA, USA (1988).
[14] Schlitt, T., Brazma, A.: Modelling gene networks at different organizational levels. FEBS Letters **579** (2005) 1859–1866.
[15] Smith, V.A., Jarvis, E.D., Hartemink, A.J.: Evaluating functional network inference using simulations of complex biological systems. Bioinformatics **18** (2002) 164–175.
[16] 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.