BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 21 2006, pages 2674–2680
doi:10.1093/bioinformatics/btl440

Systems biology

Systematic component selection for gene-network refinement

Nicole Radde[1,*,†], Jutta Gebert[1,*,†] and Christian V. Forst[2,*]

[1]Center for Applied Computer Science, University of Cologne, Weyertal 80, 50931 Cologne, Germany and
[2]Los Alamos National Laboratory, PO Box 1663, Mailstop M888, Los Alamos, NM 87545, USA

Received on June 2, 2006; revised on August 9, 2006; accepted on August 10, 2006
Advance Access publication August 22, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: A quantitative description of interactions between cell components is a major challenge in Computational Biology. As a method of choice, differential equations are used for this purpose, because they provide a detailed insight into the dynamic behavior of the system. In most cases, the number of time points of experimental time series is usually too small to estimate the parameters of a model of a whole gene regulatory network based on differential equations, such that one needs to focus on subnetworks consisting of only a few components. For most approaches, the set of components of the subsystem is given in advance and only the structure has to be estimated. However, the set of components that influence the system significantly are not always known in advance, making a method desirable that determines both, the components that are included into the model and the parameters.

Results: We have developed a method that uses gene expression data as well as interaction data between cell components to define a set of genes that we use for our modeling. In a subsequent step, we estimate the parameters of our model of piecewise linear differential equations and evaluate the results simulating the behavior of the system with our model.

We have applied our method to the DNA repair system of *Mycobacterium tuberculosis*. Our analysis predicts that the gene *Rv2719c* plays an important role in this system.

Contact: {radde.gebert}@zpr.uni-koeln.de, chris@lanl.gov

1 INTRODUCTION

Research in Systems Biology is aiming at the understanding and modelling of cellular processes on molecular level. With advanced techniques for concentration measurements of macromolecules being developed, time series of mRNA concentrations for whole organisms are now available. One of the studied organisms is *Mycobacterium tuberculosis* (Mtb) which is the causative agent of the disease tuberculosis. Yearly, tuberculosis is responsible of over 1.7 million deaths worldwide according to the WHO (data from 2004). We focus on the DNA repair system of this bacterium which is essential for its survival in hostile environments, e.g. induced by anti-microbial drugs. One particular drug is mitomycin which specifically destructs DNA. Boshoff *et al*. (2004) conducted time series expression experiments on mytomycin response of Mtb which are accessible online at the NCBI/GEO expression repository (Gene Expression Omnibus). Additional information about background intensities has been made available by Boshoff *et al*. (2004) An overview as well as specific aspects of the DNA repair system in Mtb and *Escherichia coli* can be found in Dullaghan *et al*. (2002), Mizrahi and Anderson (1998), Rand *et al*. (2003) and Walker (1996).

Building a model of the Mtb DNA repair system first raises the question which components determine this special subsystem. A subsystem of an organism is usually not closed, therefore components of a subsystem also interact with other components of the organisms. With respect to the DNA repair system major contribution originates from identified DNA repair genes and encoded transcription factors. But the DNA repair system is neither closed nor yet completely understood. Thus, novel key-players are still to be discovered and their mode of action determined. Therefore we present a novel method to expand a known gene regulatory core network by including new genes with strong influence on the genes in the core network. This method utilizes expression and interaction data to provide an extended network model.

The characterization and modeling of gene-regulatory networks have been pursuited since the early 1970s. An overview of different types of models is given by de Jong (2002). Seminal contributions by Glass and Kauffman (1973) have used Boolean networks for the description of the behavior of such systems. Owing to its simplicity, a Boolean approach can even be used for large networks and is able to make qualitative statements about the behavior of the network, but a quantitative analysis is not possible. Other approaches to analyze gene regulation use Bayesian networks (Friedman *et al*., 2000) that are based on directed acyclic graphs and include the stochastic nature of the considered biological processes. A Bayesian approach is typically used to identify and infer the topology of gene regulatory networks. On the other hand, this approach needs to be extended to study the dynamic behavior.

Differential equations are predestined for this task, especially because for small subnetworks detailed knowledge is often available which can be included in the parameter estimation. The inclusion of such information is often necessary even for simple differential equations owing to restrictions in the parameter space. Parameter estimations for differential equations require more time series data than for Boolean networks.

Gustafsson *et al*. (2005) have shown that linear differential equations can capture important features of a large-scale gene regulatory network. A recent paper by Bansal *et al*. (2006) uses simple linear differential equations to study the influence of genes in the *E.coli* SOS response pathway. Owing to the non-linear nature of gene regulatory functions, linear differential equations are not

*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.

optimally suited for quantitative gene-network modeling. To address this issue as well as to keep our model as simple as possible but as accurate as required, we will use piecewise linear differential equations.

Our paper is structured as follows: In Section 2 we will explain our model and our approach to identify the components for a chosen subsystem. This method is then applied to the DNA repair system in Section 3, leading to the result that the inclusion of gene *Rv2719c* improves the simulations of the chosen subsystem. In Section 4 we conclude with a short summary and a discussion of the results.

2 METHODS

In order to model a gene regulatory network and to estimate parameters from time series data, we proceed in three steps. We start with a set of seed genes, which belong to the subsystem we want to consider, and for which a regulatory core network is known from literature. First, in Subsection 2.2, an algorithm is used that searches for additional genes, called candidate genes, which may play an important role for the subsystem we want to model. Thus, we extend the number of components that are included in our analysis. Second, the graph-topology of the extended network is determined in Subsection 2.3 by applying a statistical analysis on the correlation coefficients between seed genes and candidate genes. And finally, the estimation of the model parameters is detailed in the last Subsection 2.4.

2.1 Network model

We define a gene regulatory network as a directed graph with a set of nodes $V = \{v_1, \ldots, v_n\}$, representing products of genes, and a set of edges $E = \{e_{ij}\}$ between these nodes. An edge e_{ij} is present between node v_j and node v_i, if the product of gene j regulates the transcription rate of gene i via binding to the promoter region of gene i. The influence can either be positive or negative. A piecewise linear regulation function $l_{ij}^{+/-}$, describing this regulatory effect, is assigned to each edge e_{ij} in the network. The expression value $x_i(t)$ of node i depends on the regulation functions of the incoming edges, and the dynamic behavior is described as

$$\dot{x}_i(t) = s_i - \gamma_i x_i(t) + \sum_{j \in \Omega_i^+} l_{ij}^+(x_j) + \sum_{k \in \Omega_i^-} l_{ik}^-(x_k). \tag{1}$$

Here, s_i, $\gamma_i \in \mathbb{R}^+$ are basic rates for synthesis and degradation, which determine the temporal change of gene product i when all regulators of v_i are inactive. Degradation of a component is assumed to be a first order decay process. Ω_i^+ and Ω_i^- are disjoint subsets of V that contain all regulators with a positive or negative effect on the expression rate of gene i, respectively. Different regulators are assumed to act independently, so the total effect on the regulated component is the sum of the single effects. Such a simplification keeps the system piecewise linear and therefore analytically solvable.

The regulation functions $l_{ij}^+(x_j)$ and $l_{ik}^-(x_k)$ describe the temporal changes in the expression value of a gene i depending on the expression value of the corresponding regulator j. Yagil and Yagil (1971) experimentally verified that these functions have a sigmoidal shape. This can also be derived theoretically, when we consider the binding reaction of the transcription factor j, TF_j, to the promoter region of gene i, i.e. the binding site B_{ij}, as a reverse chemical reaction (see e.g. Jacob and Monod, 1961):

$$m_{ij} \cdot \mathrm{TF}_j + B_{ij} \overset{K}{\rightleftharpoons} C \tag{2}$$

The factor m_{ij} corresponds to the cooperation between single transcription factors and is often denoted as Hill-coefficient in literature. In reaction (2), several transcription factors first have to form a complex for activation, and m_{ij} denotes the number of transcription factors in this complex. However, m_{ij} can be interpreted more generally as a cooperation between transcription factors and does not have to be an integer. We assume that reaction (2) is always in equilibrium, since the time-scale of our system, describing changes in gene expression, is much slower than the binding of a transcription factor to DNA. Thus, the reaction constant K uniquely determines the relation between concentrations of educts and products according to the law of mass action. K is the relation between reaction rate constants k_1 for the complex formation and k_2 for the dissociation. It depends on temperature and binding energy of the complex, as described in Djordjevic *et al*. (2003) and Gerald *et al*. (2002).

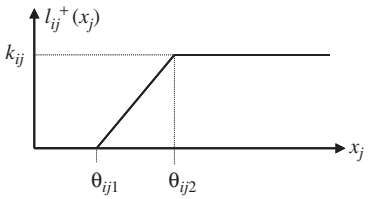

Fig. 1. Regulation function $l_{ij}^+(x_j)$ that describes the temporal change of the expression value of the regulated component i as a function of the regulator's expression value x_j. A negative influence, $l_{ij}^-(x_j)$, is described in a similar way with k_{ij} being negative.

Calculating the steady state of the corresponding systems of differential equations

$$\begin{aligned} [\dot{B}_{ij}] &= -k_1[B_{ij}][\mathrm{TF}_j]^{m_{ij}} + k_2[C] \\ [\dot{\mathrm{TF}}_j] &= m_{ij} \cdot [\dot{B}_{ij}] \\ [\dot{C}] &= -[\dot{B}_{ij}] \end{aligned}$$

and assuming that the number of bound transcription factors is much smaller than the total concentration of transcription factors, x_j, the probability that binding site B_{ij} is occupied can be written as

$$P_{B_{ij}\,\mathrm{bound}}(x_j) = \frac{x_j^{m_{ij}}}{x_j^{m_{ij}} + K^{-1}} \tag{3}$$

If we now assume that this probability is proportional to the effect on the transcription rate of gene i, we can parameterize the effect on the regulated component by

$$\dot{x}_i(x_j) = k_{ij} \cdot \frac{x_j^{m_{ij}}}{x_j^{m_{ij}} + \theta_{ij}^{m_{ij}}} \tag{4}$$

with $k_{ij} > 0$ in case of a positive regulation and $k_{ij} < 0$ in case of a negative regulation. The parameter θ_{ij} is a threshold value depending on the reaction constant of the binding reaction. Equation (4) is used as a basis to build a piecewise linear model with regulation functions of the form

$$l_{ij}(x_j) = \begin{cases} 0 & \text{for } x_j \le \theta_{ij1} \\ \dfrac{k_{ij}}{\theta_{ij2} - \theta_{ij1}}(x_j - \theta_{ij1}) & \text{for } \theta_{ij1} < x_j < \theta_{ij2} \\ k_{ij} & \text{for } \theta_{ij2} \le x_j \end{cases} \tag{5}$$

with θ_{ij1}, $\theta_{ij2} \in \mathbb{R}^+$ and $k_{ij} \in \mathbb{R}$. As shown in Figure 1, the effect on the regulated component vanishes when the expression value of the regulator is below the first threshold value θ_{ij1}. It changes linearly between θ_{ij1} and θ_{ij2}, and saturates when the expression value of the regulator exceeds the second threshold θ_{ij2}. The first threshold is determined mainly by the Hill-coefficient m_{ij}, whereas the second one depends on both m_{ij} and the reaction constant K.

With this parameterization we have a piecewise linear description of system (1). The threshold values θ_{ij} partition the state space into cuboids and within one cuboid Q the system becomes simply linear:

$$\dot{\mathbf{x}}(t) = A_Q \mathbf{x}(t) + \mathbf{c}_Q \tag{6}$$

The vectors $\mathbf{x}(t), \dot{\mathbf{x}}(t) \in \mathbb{R}^n$ contain concentrations of all genes at time t and their time derivatives, respectively. The vector $\mathrm{c_Q} \in \mathbb{R}^n$ has constant

entries coming from the basic synthesis rates as well as from regulation functions, and $A_Q \in \mathbb{R}^{n\times n}$ summarizes the linear parts of the regulation functions and the degradation terms. The model and the unique form of its general solution is derived in detail in Gebert *et al*. (2005). In order to solve such types of systems, the equations have to be decoupled by transforming the system into Jordan canonical form, see Luenberger (1973).

The piecewise linear description has several advantages in comparison with the non-linear one. First, it can be solved analytically, thus simplifying, for example, the analysis of robustness against changing of parameters. Furthermore, the partition of the state space provides a decoupling, such that both the parameter estimation and the solution can be considered separately for every cuboid. This is especially interesting in the case of locally concentrated data in state space, since our piecewise linear description can easily be limited to certain cuboids, thus reducing the number of parameters to be estimated. For the parameter estimation in the general case, one can apply methods for linear systems, as for example the hinging hyperplane algorithm developed by Breiman (1993). This algorithm finds a partition of the state space and simultaneously estimates parameters using linear regression methods. In the special case that the thresholds θ_{ij} are known in advance, the parameter estimation is trivial. One problem with piecewise linear systems consists in the behavior at the thresholds. The form of the differential equations exactly at the thresholds are essential for the system's dynamics. Thresholds can influence the dynamic behavior of the system, since they can contain additional steady states or limit cycles as described in de Jong and Page (2000). However, this is not a problem in our model owing to the special form of the regulation function.

Our model was originally developed to uncover regulations on a transcriptional level, but it can easily be extended to posttranscriptional interactions, when experimental data are available. In this case, the variables x_i no longer just denote concentrations, but more generally describe concentrations or activities. This is important if a distinction between an active and an inactive form of a protein has to be made.

2.2 Searching for potentially important genes

We start our modeling with a core network that consists of a set of genes, the seed genes, and connections between these genes, which are known from literature. In the first step we search for additional genes that may also play an important role in our subsystem. Thus, the set of network nodes is extended by so-called candidate genes. For this purpose, we use an algorithm developed by Cabusora *et al*. (2005). The input of this algorithm is a set of seed genes, a table containing interaction information and gene expression data. The set of seed genes consists of genes that are known to belong to one regulatory subsystem, i.e. the cell cycle or the DNA repair system. Interaction information can be any kind of known or predicted interactions between genes or proteins of the organism, e.g. transcriptional regulation or the knowledge that several genes are regulated by the same transcription factors.

A large network is constructed using only the table of interaction information. Then, the algorithm creates a subgraph by calculating the k-shortest paths between the seed nodes with an upper limit l for the path length. The method used for this purpose is explained in Jimenez and Marzal (1999) and Hershberger *et al*. (2003). The parameter k and the maximal path length l have to be chosen manually by the user. The distance between two nodes v_i and v_j, $z_{i,j}$, is determined by $z_{i,j} = \Phi^{-1}(1 - \tau_{i,j})$, where Φ^{-1} is the inverse of the cumulated Normal distribution and $\tau_{i,j}$ the correlation coefficient between genes i and j, calculated using expression values of genes i and j. For details see Cabusora *et al*. (2004).

In our analysis, we apply the Kendall correlation coefficient

$$\tau_{i,j} = \frac{P - I}{\sqrt{(n(n-1)/2 - T_i)(n(n-1)/2 - T_j)}} \tag{7}$$

with P being the number of proversions, I the number of inversions and T_i and T_j the numbers of bindings. Here, every pair of concentrations of component i at two different time points t and $\tilde{t}(x_i(t), x_i(\tilde{t}))$, is compared with the corresponding pair of component j, $(x_j(t), x_j(\tilde{t}))$. A proversion is a homogeneous change in both variables, i.e. $(x_i(t) > x_i(\tilde{t})$ and $(x_j(t) > x_j(\tilde{t}))$ or $(x_i(t) < x_i(\tilde{t}))$ and $x_j(t) < x_j(\tilde{t}))$, respectively. A change in the opposite direction, i.e. $(x_i(t) > x_i(\tilde{t})$ and $(x_j(t) < x_j(\tilde{t}))$ or $(x_i(t) < x_i(\tilde{t}))$ and $x_j(t) > x_j(\tilde{t}))$, is defined as an inversion. In case of $x_k(t) = x_k(\tilde{t}))$ we have a binding.

In comparison to the Pearson correlation coefficient, which is frequently used, the Kendall correlation coefficient needs more computing time, since correlations between $n(n-1)/2$ pairs of genes have to be compared. Instead, it can discover not only linear, but also other monotonous relations, and is less sensitive to outliers. As microarray data are very noisy and as the relations we want to find are expected to be highly non-linear, the Kendall correlation coefficient may be more appropriate than the Pearson correlation coefficient for our analysis.

The output of the algorithm is an undirected subgraph, that contains seed genes as well as the components of the k-shortest paths. These genes provide a set of candidate genes with potential influence on the subsystem. This set is then further investigated in the following analysis.

2.3 Identifying statistically significant edges

In the first step, we started with a set of seed genes with known gene regulation mechanism. As described in the previous step we have also obtained additional genes that may have an important influence on the subsystem. Since we do not know how to include these new genes into the network, we introduce a statistical procedure in order to find significant edges between candidate genes and seed genes. Therefore, we create a correlation distribution $\mathcal{D}$ by calculating the Kendall correlations between seed genes and candidate genes to all other measured genes in the organism. $\mathcal{D}$ is supposed to represent the real distribution of correlations within the whole gene regulatory network, and is used to assign probabilities to every correlation coefficient. Therefore, we consider the deviations from the mean m of $\mathcal{D}$ and define two subsets of the whole set S of all correlation coefficients τ in $\mathcal{D}$ according to a significance level α: The subset $S^{\min}$ contains $\alpha/2\%$ of all τ with the smallest values and $S^{\max}$ contains $\alpha/2\%$ of all τ with the largest values. The maximum of $S^{\min}$, denoted $\tau_{\min}$, and the minimum of $S_{\max}$, $\tau_{\max}$, are used to decide whether an edge is significant or not:

$$e_{ij} \text{ is significant} \Leftrightarrow (\tau_{ij} \leq \tau_{\min} \text{ or } \tau_{ij} \geq \tau_{\max}) \tag{8}$$

With the significance level α, one determines the sparseness of the network. The significant edges define a network structure that is further used in the following Subsection 2.4 to parameterize the model and estimate the parameters.

2.4 Parameter estimation

In the last step of our method we build a parameterized model, given a fixed network structure. The structure of the network contains the edges from the core network and the interactions found in Subsection 2.3. Depending on what is known about the underlying regulatory mechanisms, determining directions for undirected edges could be problematic. Using correlation coefficients that incorporate time delays may help to solve this potential issue. However, in our model directions of all edges are known.

Finally, we have to estimate the parameters of the equations with time series expression data by minimizing the squared error between time derivatives obtained from the data and the prediction derived from our system. Therefore, in a first iteration, we try to find a convenient partition of the state space, i.e. determine the thresholds θ_{ij}. In the case of insufficient data, the thresholds have to be determined beforehand. In our application we are able to set the thresholds manually by examining the expression time courses.

The partition of the state space also leads to a partition of the measurements according to the different parts of the state space. All measurements belonging to one cuboid Q, i.e. the measurements at time points t_{Qz}, $z = 1, \ldots, z_Q$, are then used to estimate the parameters for this cuboid.

This can be formulated as an optimization problem with a quadratic objective function:

$$\min_{\pi} \sum_{z=1}^{z_Q} \|\dot{\mathbf{x}}(t_{Q_z}, \pi) - \hat{\dot{\mathbf{x}}}(t_{Q_z})\|^2 \quad (9)$$

The components of the vector π are the parameters of the system which have to be estimated, that are synthesis and degradation rates s_i and γ_i, as well as the strengths of regulations k_{ij}. The components $\dot{x}_i(t)$ of the vector $\dot{\mathbf{x}}(t)$ are the time derivatives predicted from the differential equations and include the parameters to be estimated according to Equation (6). The components of $\hat{\dot{\mathbf{x}}}(t)$ are estimates for the derivatives directly derived from the experimental data. We use polynomial regression to obtain values for $\hat{\dot{\boldsymbol{x}}}(t)$ using expression measurements at the following and the previous time points:

$$\begin{aligned}\hat{\dot{\boldsymbol{x}}}_i(t_k) = &\frac{x_i(t_{k-1})}{(t_{k-1}-t_k)(t_{k-1}-t_{k+1})}\cdot(t_k - t_{k+1}) \\ &+\frac{x_i(t_k)}{(t_k - t_{k-1})(t_k - t_{k+1})}\cdot(2t_k - t_{k-1} - t_{k+1}) \\ &+\frac{x_i(t_{k+1})}{(t_{k+1}-t_{k-1})(t_{k+1}-t_k)}\cdot(t_k - t_{k-1})\end{aligned}$$

In addition, we add constraints with respect to expected signs of parameters, i.e. s_i, $\gamma_i \geq 0$ for all i.

Optimization problem (9) can be rewritten into the classical form

$$\min_{\mathbf{v}} \|M\mathbf{v} - \mathbf{w}\|^2 \text{ subject to } v_i \geq 0. \quad (10)$$

3 APPLICATION—*MYCOBACTERIUM TUBERCULOSIS* AND ITS DNA REPAIR SYSTEM

3.1 Biological background

Currently, over one-third of the world's population is infected with tuberculosis (TB). *M. tuberculosis* (Mtb) primarily causes infection of the lungs, although it can attack any part of the body such as the kidney, spine, and brain. If not treated properly, TB disease can be fatal. The emergence of multiple drug resistant strains is an increasing concern, specifically for immuno-compromised patients. The availability of the complete genomic Mtb sequence in the year 2000 was an important step for understanding the bacterium and for the development of novel drugs, intervening in regulatory processes on molecular level. The intervention should be able to circumvent existing drug resistance in Mtb.

The DNA repair system is switched on in the case of damaged DNA, resulting in single-stranded DNA, and has been extensively studied in *E.coli* (Walker, 1996). Some of the insights gained from these studies can be transfered to Mtb. The main components are the proteins RecA and LexA, which together regulate ~35–40 genes in Mtb. RecA binds to single-stranded DNA and changes the structure of the protein LexA, such that it is prevented from binding to the SOS boxes. These SOS boxes are specific binding sites in the promoters of the so called SOS genes. If LexA binds to these boxes, the expression of the SOS genes is inhibited, thus an upregulation of SOS genes is induced by DNA damage. RecA and LexA themselves belong to these genes making it possible to rapidly change the expressions of the regulated genes. In contrast to *E.coli*, there exists an alternative mechanism to upregulate some of the SOS genes. Moreover, other genes being involved in the repair system have no SOS box in their promoter region, indicating the existence of such an alternative mechanism. Both results have been found by Dullaghan *et al.* (2002). Our method to identify additional key-genes with important functions in the DNA repair system may also contribute to learn more about the alternative mechanism.

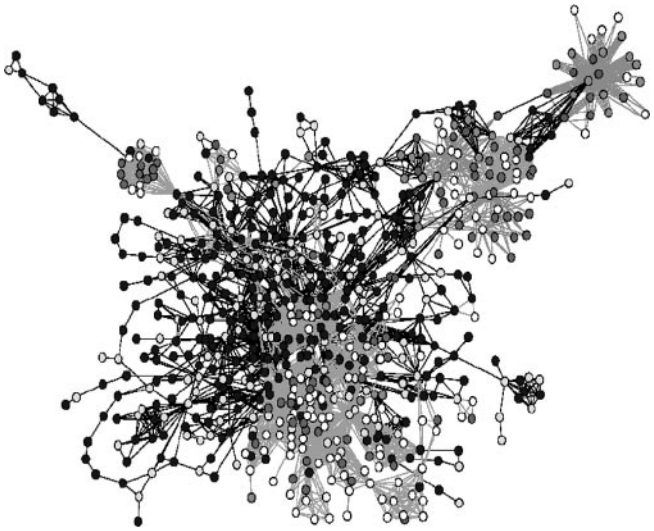

Fig. 2. Mtb source network including approximately 1000 genes and 70 000 interactions. Nodes represent genes, edges indicate interaction between two genes.

The experimental data used for the model building process has been generated by Boshoff *et al.* (2004) and can be accessed at the NCBI's, Gene Expression Omnibus (Gene Expression Omnibus, http://www.ncbi.nih.gov/geo). This data set includes 16 experiments, in which Mtb is treated with 0.2 μg mitomycin. Gene expression was measured at 0.33, 0.75, 1.5, 2, 4, 6, 8 and 12 h after treatment. Analyses were performed using the BRB ArrayTools developed by Dr Richard Simon and Amy Peng (BRB Array Tool, http://linus.ncbi.nih.gov/brb/download.htm). We use the ratio between treated cells and control, no filters and a median normalization.

3.2 Finding possibly important genes

In the first step the algorithm described in Cabusora *et al.* (2004) uses interaction data as well as gene expression data of Mtb. The interaction data consists of a list of interactions found in experiments, in databases and in the literature. This list is used to build a source network for Mtb as described in Section 2 and is displayed in Figure 2.

As we are interested in the DNA repair system of Mtb, we use the genes *recA, lexA, ruvC* and *linB* as the input set of genes called seed genes hereafter. The genes *ruvC* and *linB* are chosen as representatives of the SOS genes. Among the SOS genes there exists a group which is regulated solely by *recA* and *lexA*, such as *linB, dnaE2* and *lexA*, but some genes are upregulated despite a perturbation of the *recA–lexA* mechanism. Genes belonging to this second group are *ruvC, recA* and *Rv2100*. Together with the SOS regulatory mechanism, these genes are also regulated by an alternative mechanism (Rand *et al.*, 2003).

The output of the algorithm (Cabusora *et al.*, 2004), which is shown in Figure 3, yields a response network of genes which possibly have an important influence on the seed genes. These genes include *Rv2719c, infB* and *dnaE2*. Interestingly, the gene *Rv2719c* has been detected by Dullaghan *et al.* (2002) to be a DNA damage inducible gene. In order to select genes which should be added to the model, we evaluate in

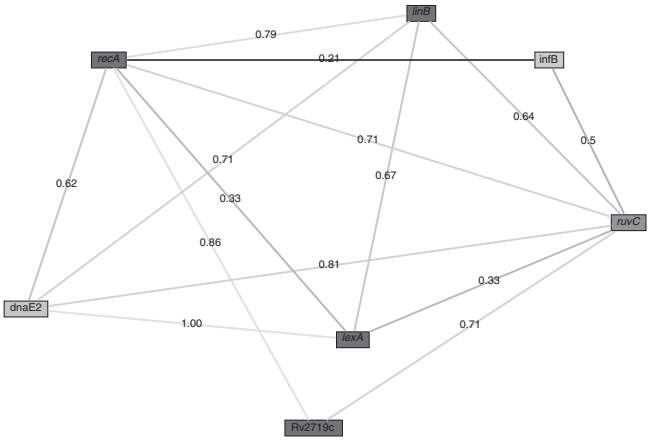

Fig. 3. Output for the seed genes *recA, lexA, ruvC* and *linB* using 9-shortest paths with maximal path length $l = 10$. Edges are labelled with Kendall correlation coefficients.

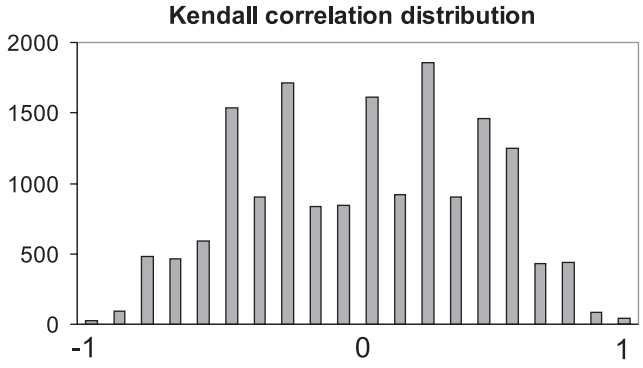

Fig. 4. Distribution of correlation coefficients for the network genes with all other measured genes.

Table 1. Kendall correlation coefficients for each pair of genes in our system

Genes	*lexA*	*recA*	*ruvC*	*linB*	*infB*	*dnaE2*	*Rv2719c*
lexA	1.00	0.33	0.33	0.67	−0.67	1.00	0.00
recA	0.33	1.00	0.71	0.79	0.21	0.62	0.86
ruvC	0.33	0.71	1.00	0.64	0.5	0.81	0.71
linB	0.67	0.79	0.64	1.00	0.29	0.71	0.79
infB	−0.67	0.21	0.50	0.29	1.00	0.62	0.36
dnaE2	1.00	0.62	0.81	0.71	0.62	1.00	0.52
Rv2719c	0.00	0.86	0.71	0.79	0.36	0.52	1.00

Table 2. Probabilities to get the correlation coefficient for every pair of genes or an even higher deviation from the mean m of the distribution $\mathcal{D}$

Genes	*lexA*	*recA*	*ruvC*	*linB*	*infB*	*dnaE2*	*Rv2719c*
lexA	0.00	0.51	0.51	0.07	0.07	**0.00**	1.00
recA	0.51	0.00	**0.04**	**0.03**	0.62	0.13	**0.01**
ruvC	0.51	**0.04**	0.00	0.07	0.25	**0.01**	**0.04**
linB	0.07	**0.03**	0.07	0.00	0.57	**0.04**	**0.03**
infB	0.07	0.62	0.25	0.57	0.00	0.13	0.45
dnaE2	**0.00**	0.13	**0.01**	**0.04**	0.13	0.00	0.20
Rv2719c	1.00	**0.01**	**0.04**	**0.03**	0.45	0.20	0.00

For significant values we use bold face.

the following subsection each pair of genes from the list by considering the correlation strengths between them.

3.3 Statistically significant edges in the network

We now want to determine the statistical significance of the Kendall correlation coefficients for each pair of genes in our system. The coefficients are listed in Table 1. We use 3923 measured genes/ORFs in the experiments to calculate correlation coefficients between these genes/ORFs and genes of our DNA repair model (*recA, lexA, ruvC, linB, Rv2719c, infB* and *dnaE2*). Figure 4 shows a histogram of the distribution $\mathcal{D}$ of correlation coefficients. The mean of $\mathcal{D}$ is $m = -0.004$, the standard deviation is $\sigma = 0.415$. In Table 1, the value 0.86 shows a strong correlation between the genes *Rv2719c* and *recA*. In order to evaluate if the correlations are significant, we apply the statistical procedure described in subsection 2.3 with significance level $\alpha = 5\%$. This leads to the cutoff values $\tau_{\min} = -0.714$ and $\tau_{\max} = 0.714$. The resulting probabilities are shown in Table 2. As expected, the probabilities for the correlation coefficients between the seed genes are very low, many of them fall below the significance level of $\alpha = 5\%$. The probabilities for the correlation coefficients of gene *infB* with other network genes are not significant and are therefore omitted in the following analysis. However, the genes *Rv2719c* and *dnaE2* show significant correlation coefficients to some of the seed genes. For example, there are significant values for the pairs (*dnaE2*, *lexA*) and (*Rv2719c*, *recA*). When inspecting the time series of these two genes, we detect that the expression of gene *dnaE2* has as its highest upregulation a 2.1-fold expression and for gene *Rv2719c* we have up to 12-fold expression. Therefore, gene *dnaE2* is not significantly up- or downregulated, thus we conclude that this gene is not involved in the considered system, whereas gene *Rv2719c* seems to play an important role.

3.4 Modeling and simulation of the DNA repair system

In this subsection we will model the basic system consisting of the genes *recA, lexA, ruvC* and *linB* as well as an extended system including additionally the gene *Rv2719c*. Already known or assumed regulations are summarized in Figure 5. The nodes of our gene regulatory network correspond to the products of genes.

In the core network the protein RecA is activated by single-stranded DNA, thus ensures that LexA can no longer bind to the SOS box in the case of damaged DNA. LexA is still present in the cell, but in a different, non-binding form. Therefore we need two variables for the description of LexA. LexA refers to the total amount and LexASOS denotes the fraction of the protein amount which can bind to DNA. LexASOS inhibits *recA* and all other SOS genes, such as *ruvC* and *linB*. Three additional regulations are inserted in the extended network owing to the presence of *Rv2719c*. This gene itself has a SOS box (Dullaghan *et al*., 2003), therefore it is inhibited by LexASOS. Moreover, we propose that

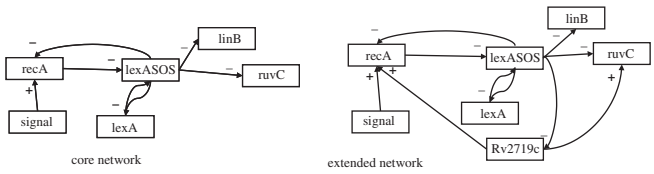

Fig. 5. Gene regulatory networks of the DNA repair system with and without *Rv2719c*.

this gene plays an important role in the, so far, unknown regulation mechanism of Mtb DNA repair. Therefore we include activation functions from gene *Rv2719c* to *recA* and *ruvC* in our model.

The differential equations describing the DNA repair system are built according to our model approach described in subsection 2, using basic synthesis and degradation rates, inhibitions and activations. In the following, *recA* corresponds to index 1, *lexA* to 2, *lexASOS* to 3, *ruvC* to 4, *linB* to 5 and *Rv2719c* to 6.

We derive the differential equations for x_1, x_2, x_4 and x_5 without *Rv2719c* as follows:

$$\begin{aligned}
\dot{x}_1(t) &= c_1 - \gamma_1 \cdot x_1(t) + \alpha_1 \cdot \text{signal} + k_{1,3} \cdot b^-(x_3(t), \theta_{2,3}) \\
\dot{x}_2(t) &= c_2 - \gamma_2 \cdot x_2(t) + k_{2,3} \cdot b^-(x_3(t), \theta_{2,3}) \\
\dot{x}_4(t) &= c_4 - \gamma_4 \cdot x_4(t) + k_{4,3} \cdot b^-(x_3(t), \theta_{4,3}) \\
\dot{x}_5(t) &= c_5 - \gamma_5 \cdot x_5(t) + k_{5,3} \cdot b^-(x_3(t), \theta_{5,3})
\end{aligned}$$

with $\alpha_1 \in \mathbb{R}^+$, c_i, $\gamma_i \in \mathbb{R}^+$ and $k_{i,j} \in \mathbb{R}^-$. The Boolean functions $b^-(x_j(t), \theta_{i,j})$ are simplifications of the piecewise linear functions described in subsection 2, because the number of measurements are not sufficient for a more detailed description. The variable $x_j(t)$ denotes again the expression value of the mRNA of the influencing gene, and $\theta_{i,j}$ is the value of x_j at which the influence reaches its maximal strength. The function is equal to zero for $x_j(t) \geq \theta_{i,j}$ and equal to one for $x_j(t) > \theta_{i,j}$.

Introducing *Rv2719c* into the model results in additional activation functions with respect to *ruvC* and *recA*:

$$\begin{aligned}
\dot{x}_1 &= c_1 - \gamma_1 \cdot x_1 + \alpha_1 \cdot \text{signal} + k_{1,3} \cdot b^-(x_3, \theta_{2,3}) + k_{1,6} \cdot b^+(x_6, \theta_{1,6}) \\
\dot{x}_2 &= c_2 - \gamma_2 \cdot x_2 + k_{2,3} \cdot b^-(x_3, \theta_{2,3}) \\
\dot{x}_4 &= c_4 - \gamma_4 \cdot x_4 + k_{4,3} \cdot b^-(x_3, \theta_{4,3}) + k_{4,6} \cdot b^+(x_6, \theta_{4,6}) \\
\dot{x}_5 &= c_5 - \gamma_5 \cdot x_5 + k_{5,3} \cdot b^-(x_3, \theta_{5,3})
\end{aligned}$$

with $k_{1,6}$, $k_{4,6} \in \mathbb{R}^+$.

There is no basis to build a differential equation for x_6 as no regulatory influences on *Rv2719c* are known. For the parameter estimation of these four equations we therefore have to use the measured data for *Rv2719c*. Moreover, the mRNA of *lexA* has been measured without the distinction if the protein LexA can bind or not, thus we can only make assumptions about the amount of binding LexA and omit to describe the variable quantitatively. Instead, we have used a Boolean description:

$$x_3(t) = \begin{cases} 1 & \text{for time 0 h up to 4 h} \\ 0 & \text{for time 4 h up to 6 h} \\ 1 & \text{for time after 6 h} \end{cases}$$

Parameters are estimated using the method of least squares with constraints as described in Section 2. For this purpose, we need to assign the data to the different linear differential equations which

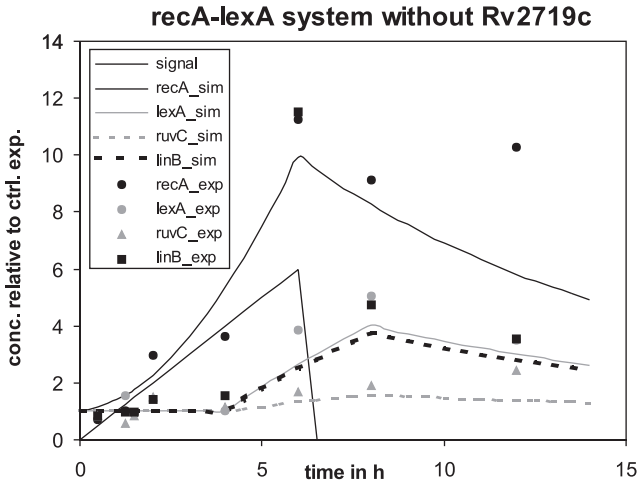

Fig. 6. Simulation for the core network without *Rv2719c* with initial conditions $x_i = 1$ and a signal which increases until time 6h and disappears thereafter.

is equivalent to setting the threshold values for the regulation functions. We decided to group the data as follows: Data for the time points $t = 0.33, 0.75, 1.5, 2, 4$ h are assigned to the differential equations where the signal (single-stranded DNA) affects RecA, but LexA still binds to the SOS boxes, therefore $x_3 = 1$. From time point $t = 4$ h until time point $t = 6$ h, LexA does no longer bind to the DNA, thus we set $x_3 = 0$. For time points t = 6, 8, 12 h, LexA again inhibits the SOS genes.

The derivatives are estimated using polynomial regression of second order as described in subsection 2.4. As the time behavior of all components can already be reproduced using only two of the three parameters for each component, we have set the degradation rates for each component to $\gamma_i = 0.1$. Constraints for the minimization are positive synthesis rates, positive maximal activations and negative maximal inhibitions. In the core network without *Rv2719c* we achieve the following parameters:

$$\begin{aligned}
&\gamma_i = 0.1\,\text{h}^{-1},\ \alpha_1 = 0.548\,\text{h}^{-1},\ k_{1,3} = 0\,\text{h}^{-1}, \\
&k_{2,3} = -0.898\,\text{h}^{-1},\ k_{4,3} = -0.168\,\text{h}^{-1},\ k_{5,3} = -0.82\,\text{h}^{-1}
\end{aligned}$$

The estimations of the parameters in the extended network are

$$\begin{aligned}
&\gamma_i = 0.1\,\text{h}^{-1},\ \alpha_1 = 0.511\,\text{h}^{-1},\ k_{1,3} = -0.898\,\text{h}^{-1}, \\
&k_{2,3} = -0.898\,\text{h}^{-1},\ k_{4,3} = -0.013\,\text{h}^{-1},\ k_{5,3} = -0.82\,\text{h}^{-1}, \\
&k_{4,6} = 0.233\,\text{h}^{-1},\ k_{1,6} = 0.464\,\text{h}^{-1}
\end{aligned}$$

We compare the simulations of the core and the extended network to draw a conclusion about the importance of gene *Rv2719c*. In Figure 6 the simulation using the core network is shown. The simulation using the extended network with the same initial conditions as for the previous simulation is illustrated in Figure 7. In each figure, experimental data or mean values for multiple measurements are also shown. Both simulations show similar behaviors between $t = 0$ h and $t = 6$ h. Then the behaviors diverge due to the positive influence of *Rv2719c* on *ruvC* and *recA*. In the simulation based on the core network, *ruvC* and *recA* decrease to their original levels, i.e. to their fixed points, thus the response abates fast. Contrarily, in the extended network these genes demonstrate a

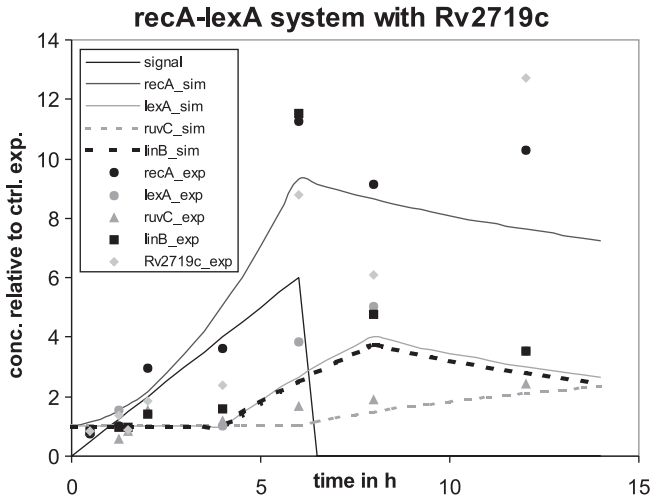

Fig. 7. Simulation for the extended network with initial conditions $x_i = 1$.

slightly higher expression, which indicates that the response persists for a longer time. This behavior is in better accordance with the experimental data.

4. DISCUSSION

We have developed a novel method to identify and verify the importance of specific genes for the function and dynamics of gene regulatory networks. By a multi-step process we first construct a response network from interaction data and gene expression profiles. We further refine this response network by applying statistical analysis methods and calculating correlation coefficients for each connection in the network. We then use such a refined response network as scaffold to construct a dynamic gene regulation network based on a hybrid model of piecewise linear differential equations built on Boolean regulation functions. Finally, we estimate the parameters for this model and evaluate the results by simulating the dynamic behavior of the network.

We have applied our method to the DNA repair system in *M. tuberculosis*. Our simulation results indicate that the hypothetical gene *Rv2719c* plays an important role in the SOS repair mechanism. This result is in agreement with the analysis of the SOS genes by Dullaghan *et al*. (2002), who suggested that this gene also takes part in the DNA repair mechanism.

In the last few years, the amount of available experimental data to infer interactions between cell components has grown tremendously, justifying quantitative modeling approaches that provide detailed insights into the dynamics of the underlying regulatory processes. Thus, a tendency exists to use more complex models to capture regulatory mechanisms, i.e. moving from Boolean networks to time and state continuous models or using models that contain nonlinear equations instead of just linear descriptions. However, in practice it is usually not possible to determine the parameters of quantitative models from gene expression data alone. Thus, one has to restrict the solution space either by including prior knowledge about the system or by incorporating further data sources.

With our method, we are able to contribute to quantitative modeling efforts using multiple data sources and by including biological knowledge. We are not only capable to predict the dynamic behavior of regulatory processes but to pinpoint key-elements in these processes for further investigation and experimental verification.

ACKNOWLEDGEMENTS

This work was supported by grants from the Los Alamos National Laboratory (LDRD-20040184ER), the DAAD and the BMBF (CUBIC). We gratefully acknowledge Dr. Boshoff, NIAID, for providing *M. tuberculosis* gene-expression data and for continuous support.

This paper is dedicated to Prof. Peter Schuster on the occasion of his 65th birthday.

REFERENCES

Bansal,M. *et al.* (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, **22**, 815–822.

Boshoff,H. *et al.* (2004) The transcriptional response of *Mycobacterium tuberculosis* to inhbitors of metabolism: novel insights into drug mechanisms of action. *Biol. Chem.*, **279**, 40174–40184.

Breiman,L. (1993) Hinging hyperplanes for regression, classifictaion and function Approximation. *IEEE Trans. Inform. Theory*, **39**, 999–1013.

Cabusora,L. *et al.* (2005) Differential network expression during drug and stress response. *Bioinformatics*, **21**, 2898–2905.

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

Glass,L. and Kauffman,S.A. (1973) The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.*, **39**, 103–129.

de Jong,H. and Page,M. (2000) Qualitative simulation of large and complex genetic regulatory Systems. In Horn,W. (ed.), In *Proceedings of the ECAI 2000*, IOS Press, pp. 191–195.

de Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comp. Biol.*, **9**, 67–103.

Djordjevic,M. *et al.* (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**, 2381–2390.

Dullaghan,E.M. *et al.* (2002) The role of multiple SOS boxes upstream of the *Mycobacterium tuberculosis* lexA gene—identification of a novel DNA-damage-inducible gene. *Microbiology*, **148**, 3609–3615.

Gebert,J. *et al.* (2006) Modeling gene regulatory networks with piecewise linear differential equations, accepted for publication in: *EJOR Chall. of Cont. Opt. in Theory and Applications*, in press.

Gerland,U. *et al.* (2002) Physical constraints and functional characteristics of transcription factor-dna interaction. *Proc. Natl. Acad. Sci. USA*, **99**, 12015–12020.

Gustafsson,M. *et al.* (2005) Constructing and analyzing a large-scale gene-to-gene regulatory network—Lasso-constrained inference and biological validation. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2**, 254–261.

Hershberger,J. *et al.* (2003) Finding the *k*-shortest simple paths: a new algorithm and its implementation. In *proceedings of the 5th Workshop on Algorithm Engineering and Experiments*, ALENEX 2003, Baltimore, USA, 26–36.

Jacob,F. and Monod,J. (1961) Genetic regulatory machanisms in the synthesis of proteins. *J. Mol. Biol.*, **3**, 318–356.

Jiménez,V.M. and Marzal,A. (1999) Computing the *k*-shortest paths: a new algorithm and an experimental comparison. In *Lecture Notes in Computer Science* 1668, 15–29, 3rd International Workshop on Algorithm Engineering (WAE 99) July 19–21, 1999, London, UK.

Luenberger,D.G. (1973) *Introduction to Linear and Nonlinear Programming. Addison Wessley, Massachusetts.*

Mizrahi,V. and Anderson,S.J. (1998) DNA repair in *Mycobacterium tuberculosis*. What have we learnt from the genome sequence? *Mol. Microbiol.*, **29**, 1331–1339.

Rand,L. *et al.* (2003) The majority of inducible DNA repair genes in *Mycobacterium tuberculosis* are induced independently of Rec A. *Mol. Microbiol.*, **50**, 1031–1042.

Walker,G.C. (1996) The SOS response of *Escherichia coli*. In Neidhardt,F.C. (ed.), *Escherichia coli and Salmonella*. Washington: ASM Press, pp. 1400–1416.

Yagil,G. and Yagil,E. (1971) On the relation between effector concentration and the rate of induced enzyme synthesis. *Biophys. J.*, **11**, 11–27.