Reverse Engineering Non-Linear Gene Regulatory Networks Based on the Bacteriophage λ cI Circuit

Jochen Supper, Christian Spieth and Andreas Zell
Centre for Bioinformatics (ZBIT)
University of Tübingen
Sand 1, 72076 Tübingen, Germany
Email: supper@informatik.uni-tuebingen.de

***Abstract*— The ability to measure the transcriptional response of cells has drawn much attention to the underlying transcriptional networks. To untangle the network, numerous models with corresponding reverse engineering methods have been applied. In this work, we propose a non-linear model with adjustable degrees of complexity. The corresponding reverse engineering method uses a probabilistic scheme to reduce the reconstruction problem to subnetworks. Adequate models for gene regulatory networks must be anchored on sufficient biological knowledge. Here, the cI auto-inhibition circuit (cI circuit) is used to validate our reverse engineering method. Simulations of the cI circuit are used for the reconstruction, whereas a simplified cI circuit model assists the modeling phase. Several levels of complexity are evaluated, subsequently the reconstructed models show different properties. As a result, we reconstruct an abstract model, capturing the dynamic behavior of the cI circuit to a high degree.**

I. Introduction

An increasing amount of transcriptional data is being collected, providing valuable insight to cellular processes under various conditions. The interpretation of this data can be focused on a particular biochemical reaction or broadened to global-scale networks. Focused models often employ detailed kinetic and stochastic reactions [17] obtained by forward engineering. Global modeling employs abstract representations [12], [14], [21], [23] used by reverse engineering approaches. Many models have been proposed on both sides of this spectrum, whereas intermediate models are scarce because it is hard to incorporate the contradicting requirements in one model. Unfortunately, most reverse engineering attempts impose a simple structure on the data, guided by computational requirements, where the true structure may only be revealed with detailed models.

In this work, we use a adjustable network model covering different degrees of complexity, from basic linear to various non-linear models. A major aspect of the modeling will be the degree and type of complexity employed. Therefore, utilization of complexity - including non-linear terms - is carefully considered. To provide the functional design we validate our approach on a kinetic model.

Thieffry et al. [20] found that most regulation circuits in Escherichia coli are one-element auto-inhibition circuits. Whereas Bundschuh et al. [3] provide a detailed kinetic model of the bacteriophage λ cI circuit. The cI circuit is part of the bacteriophage λ lytic and lysogenic pathway, which has been and still is intensively studied [8]. This and several other biochemical models have been thoroughly investigated with a component-centric focus [1]. Various forward modeling approaches have been animated by this wealth of knowledge, including detailed stochastic [1] and kinetic [13] models. Such subnetworks give valuable examples of gene expression, control and dynamics. With these models general questions regarding the proper conceptual and mathematical representation and the sufficient amount of detail required for suitable predictability can be addressed.

Several methods have been developed to reverse engineer global networks. These models include Boolean networks [14], (non)-linear networks [22], [23], S-Systems [19] and differential equations [4]. Boolean networks employ discretized data, describing the state of a gene as either ”on” or ”off”. This renders them as biologically not very realistic. Linear systems of equations are used to describe linear and non-linear models. The description of linear models is straight forward, where non-linear models have to be linearized. Weaver et. al [22] and de Jong et al. [23] employ different linearizations, which will be discussed later in further detail. S-Systems employ complex interaction terms and can not be solved analytically. Overall, these methods aim to reconstruct the global network and implement abstract interaction terms.

Today, modeling is guided by a rich flow of experimental data. The stream is still widened by an increasing pool of measurement techniques including mRNA microarray technology [18], chromatin immunoprecipitation (ChIP) [2], quantitative RT-PCR [11] and microarray-based immunoassays [15]. Despite of all this information, detailed knowledge regarding network models is still almost exclusively collected by biologists. They collect and integrate data, expand and refine their models and finally validate them. On a global scale integration of data and validation of the results are major problems with no obvious solution. Since we are not able to cope with all of these problems here, we restrict ourself to mRNA time-series and simulation of the cI circuit to validate our model.

The restriction to mRNA data-series provides us with a sampling of M time points, where M is usually small compared to the number of genes N. Under this condition the reconstruction of a genetic network is underdetermined. Mathematically this describes an infinite ensemble of possible

0-7803-9387-2/05/$20.00

solutions. Occam's razor[1] and the assumption that genetic networks are sparsely connected [20] suggest to search for the minimally connected network (also referred to as sparse network). Different regression and norm-based approaches as well as linear programming have been employed to find minimal genetic networks [9], [23]. A problem thereby is – among linearity – that their objective function (i.e. L_1 and L_2 norm) optimize something different than sparsity. More complex models like S-Systems can not be solved analytically any more. Here, we introduce a novel analytical reconstruction method which finds the minimal network by reducing the problem to subnetworks.

In contrast to the mathematical reconstruction of the network, biological plausibility often demands for non-linear interaction terms. These interaction terms should be incorporated while maintaining an analytically solvable model. Weaver et. al [22] use an approach known from neural networks that employs sigmoidal activation functions by applying their inverse to the output data. De Jong et al. [23] use piecewise-linear input functions, which discretize the phase space. Each of these models employs specific properties utilizable during the modeling process. However, no specific utilization of modeling terms has emerged as the model of choice and it is unlikely that one will. Therefore, we introduce a general interaction term for which any function can be used.

To converge detailed biochemical models and abstract reverse engineering approaches, the biochemical model should be simplified. Regulation processes incorporate a diverse set of molecules interacting on different time scales. Transcription and translation operate on the timescale of minutes to hours. Other reactions like the dimerization of a protein occur on the timescale of seconds [1]. On the timescale of transcription, fast reactions equilibrate and lose their dynamic behavior. Removing such fast reactions or lumping them into simpler mathematical representations leads to a significant reduction of the model. However, intrinsic non-linearities cannot be removed and the behavior is altered to some extent.

We propose a model (sec. II), which is solved by a probabilistic reconstruction method (sec. II-B) and can contain non-linearity (sec. II-D). To evaluate the validity of this approach, we adjust our model to reconstruct and predict the cI circuit (sec. IV), where a simplified form of the cI circuit will guide the modeling process. The results of the reverse engineering will be evaluated for predictability and topology under different levels of complexity (sec. V).

II. Mathematical And Computational Methods

The proposed reconstruction method is separated into three distinct modules throughout this section and in the implementation. A linear solver is at the core of the reconstruction method, a probabilistic reduction scheme is used to search for the minimal network and a mapping function introduces non-linearities. To allow the application of analytical solvers the mathematical model is described by linear equations. The probabilistic search scheme is introduced to find the minimal network by reducing the problem to subnetworks. Finally, we incorporate non-linearities into the model while maintaining the analytical solvability.

A. Mathematical Model

A gene regulatory model based on mRNA abundance data describes how a gene is regulated through the expression of other genes. The basic linear model of interaction is given in Eq. (1). The regulatory interactions between the genes are represented by a weight matrix W, where each row of W represents all regulatory inputs for a specific gene. The regulatory effect of gene x_j on gene x_i at time-point t is the expression level of x_j multiplied by its regulatory influence on x_i, w_{ij}. The total regulatory input to x_i is derived by summing over all genes in the system.

$$x_i(t + \Delta t) = \sum_{j=1}^{N} w_{ij} x_j(t) \tag{1}$$

A positive value for w_{ij} indicates that gene x_j is stimulating the expression of gene x_i. Similarly, a negative value indicates repression, while a value of zero indicates that gene x_j does not influence the transcription of gene x_i. This modeling of regulatory interactions enables us to use analytical approaches for solving linear systems of equations.

B. Reconstruction Method

As stated above, Eq. (1) describes the regulation of a gene. A system of linear equations can be set up by all equations describing the behavior of a particular gene. Thus, if $M + 1$ measurements – equidistant in time – are available, it is possible to create a linear system with M equations. Because the number of measurements is scarce, M is much smaller than the number of genes N. Therefore, the solution is mathematically underdetermined. This leads to a high-dimensional solution space, wherein one can pick any point to reconstruct the network. Since there is an infinite number of solutions such a reconstruction will be arbitrary and only fit the data, while having no resemblance to the biological system or the minimal solution.

To find a biologically plausible system we impose sparseness and restrict ourselves to the minimal network within the solution space. Therefore, we assume that k is the maximal in-degree of any gene, satisfying $k < M$. With this assumption, a straight-forward approach would be to screen for every possible combinatorial regulation of one gene by k other genes. This creates $\binom{N}{k}$ overdetermined systems of linear equations with k variables and M equations. These systems can be solved by least-square analysis, ranking the results according to their fit. Thus, the overall problem is reduced to $\binom{N}{k}$ smaller problems which can be solved and ranked. This approach is called MWSLE (Minimum Weight Solutions to Linear Equations) and is NP-complete. Chen et al. [5]

[1]Occam's Razor states that one should make no more assumptions than needed. This renders the simplest explanation the best.

proposed to bound k reducing the complexity to $O(M \cdot N^k)$, still rendering it inapplicable for large N.

To further relieve the computational complexity we propose a new reduction method. Again, we try to solve an underdetermined linear system of equations with N genes and M equations. To make the system solvable, we reduce the problem by restricting ourselves to M genes. Thus, we consider a subnetwork containing M genes instead of N. The fully connected subnetwork describing the regulation of a particular gene contains at most $k+1$ genes and is contained in some network of size M. Such a network can be solved because it is described by M equations and M variables. Since we do not know which subnetwork to pick, we consider the probability of randomly picking one that contains all regulatory interactions for one gene. Overall, we have $S_N := \binom{N}{k}$ possible interactions. In a network of the size M the number of possible interactions is $S_M := \binom{M}{k}$. Therefore, it covers a fraction of all possible solutions. The probability to capture one particular solution is $\frac{S_M}{S_N}$. Repeating this procedure by independent choices of M increases the overall probability to find the minimal network. This probability can be calculated with Eq. (2), where r is the number of runs.

$$p = \left(1 - \frac{S_M}{S_N}\right)^r \tag{2}$$

For every repetition, we have to solve a linear system of equations, providing a possible network structure. This creates an ensemble of r network structures. Within this ensemble we search for the minimally connected network. To decide whether one gene has a regulating influence, we check if its weight exceeds a certain threshold. If this is the case a regulatory edge is drawn in our subnetwork. The solution ensemble may contain many network structures with more than k connections, violating our assumption. These networks are neglected. Network structures having less than $k+1$ connections are ranked according to their connectivity and a least-square analysis of the weights smaller than the threshold. Finally, we select the structures with the smallest connectivity and within them the structure with the smallest least-squares value.

Overall, we relieve the computational complexity by implicitly covering a large number (S_M) of potential regulatory interactions. Note that this relief is dependent on the number of measurements taken. By increasing M, S_M grows exponentially. For a desired probability p we can calculate the number of necessary runs by solving Eq. (2) for r. The runtime of the reconstruction can be calculated by multiplying the number of runs r with the runtime for solving a linear system of M equations. The runtime for solving the linear system depends on the solver used.

To give an example, we calculate the runs necessary to reconstruct a network of 100 genes with 21 measurements and $k = 3$. If every possible solution is searched $S_N := \binom{N}{k} = 161,700$ combinatorial choices have to be evaluated. By reducing the problem to a subnetwork of size M, $S_M := \binom{M}{k} = 1,140$ combinatorial choices are implicitly evaluated. By applying statistics we can calculate the number of picks r needed for a desired confidence of reconstruction p. For instance, a confidence of 99.9% is reached, if 650 reductions are evaluated, whereas an exhaustive search would require 161,700 evaluations. This provides a significant relief in computation time.

The probabilistic reduction scheme generates an ensemble of linear systems that need to be solved. Several methods are available for this task, differing in runtime and robustness to ill-conditioned linear systems. We use singular value decomposition (SVD) [16] because of its robustness and its previous use in reverse engineering of genetic networks [23].

C. Evaluation of the Reconstruction Method

The reconstruction method is evaluated with a simple artificial model and the cI circuit. The artificial model is designed to give a proof of the concept and explore the properties of the reconstruction method. The cI circuit is reconstructed validating non-linear modeling and predictability.

The artificial network used for evaluation consists of 40 genes and a maximal in-degree of 3. The network is randomly generated with weights ranging from -1 to 1 and an in-degree ranging from 1 to 3. Measurement series are created with different initial values. Gaussian noise with a standard deviation of 5% is applied to every measurement. For the evaluations one property is modified while the others are kept constant. All runs are repeated ten times to obtain an average performance. The confidence of the reconstruction p is set to 99.9% throughout this work, with Eq. (2).

The evaluation of the reconstruction is tested under varying conditions including the number of measurements, maximum in degree, measurement noise and the network size. We also test the minimal number of equations needed under noise-free conditions, where mathematically $k+1$ is expected. To benchmark the reconstruction, we evaluate the specificity and sensitivity. A connection between two genes is considered as correctly reconstructed if the topology is captured while distinguishing induction from repression. Since some weights in the matrix will have small values a threshold of 0.1 is chosen to ignore minimal deviations from zero.

As second step the reconstruction is evaluated with the cI circuit. Modeling and reconstruction demands for thorough consideration of the underlying system. Therefore the complete model will be introduced in section III, and subsequent modeling will be discussed in section IV.

D. Introduction of Non-Linearity

The linear model fulfils our requirement for analytical solvability. Non-linear models are introduced to provide more flexibility. This is addressed in two steps. First, we extend the linear model Eq. (1) at several points to achieve greater flexibility. Then, we resolve the extensions by mapping them back into a linear model. The complete extension is given in Eq. (3). It is not necessarily intended to explore the complete complexity and flexibility of the model, but to utilize any subset of it.

$$g^{-1}\left(x_i(t+\Delta t)\right) = \sum_{j}^{N} w_{ij} \cdot h\left(x_j(t)\right) + b_i - \lambda_i x_i(t) \quad (3)$$

All extensions on the right hand side of Eq. (3) were previously implemented by different research groups. Along with the explanation of this model, we will refer to previous work in which the particular expansion was considered. The first-order degradation of RNA is described by $\lambda_i x_i(t)$, with λ_i denoting the degradation rate. This term is frequently used and discussed in the review of de Jong et al. [6]. The basal expression is incorporated by the term b_i which defines the expression level in absence of any regulatory input. Weaver et al. [22] incorporate this term into a linear model by using a constantly expressed "on-gene" which mimics the basal expression.

So far, the model is still linear, non-linearity is introduced by g and h. Both functions essentially act in the same way, but in different directions. The function h maps the input values to a non-linear form by applying the desired function to each value of $x(t)$. The function g maps the output values to a non-linear form, by applying the inverse of the desired function to each value of $x(t+\Delta t)$. If this is done for all time-steps, the non-linear model can be reconstructed with a linear solver. This model is sketched in Fig. 1, where the node functions correspond to g and the edge functions to h.

The function g was introduced by Weaver et al. [22]. By applying a method known from neural networks they are able to linearize the non-linear model. The function g is a sigmoidal dose-response function with the parameters α and β, which can be merged into the linear model as well. In principle any other invertible function can be applied for g, although the parameterization may not be linearizable. Apart from being a very elegant reconstruction method, it suffers from a strong sensitivity to noise, because the inverse of the sigmoidal function gets very steep and the maximal expression level of every protein has to be specified $a\ priori$.

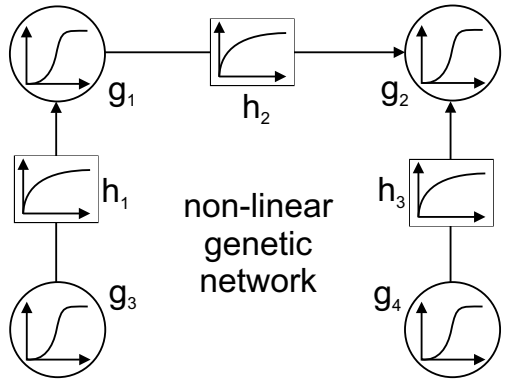

Fig. 1. Sketch of a non-linear genetic network. The edges are associated with a Michaelis-Menten activation function, and the nodes are associated with a sigmoidal dose-response function. The edge function correspond to h and the node functions correspond to g.

We propose a novel modeling term by mapping the non-linear input values to a linear model by the function h. The functions h and g act similarly except that g is applied with its inverse, where h is applied to $x(t)$ with the normal function. This simple trick linearizes a model which behaves non-linear on the activation, illustrated by the edge functions in Fig. 1. The function of activation h has to be specified prior to the reconstruction and for a parameterized function every parameter set has to be evaluated separately. De Jong et al. [7] proposed a step function for g, where we propose a method allowing for general functions.

To illustrate the mapping of a non-linear to a linear model by h, we employ a simple example. For those interested in the detailed method of incorporating g, we refer to [22]. In our example, gene A influences gene B according to the repression function $\alpha \cdot (1 + {}^{A(t)}/_{\beta})^{-1}$, derived from Michaelis-Menten kinetics with the parameter values $\alpha = \frac{1}{2}$ and $\beta = 1$. If measurement values of A are i.e. $(1, 2, 3, 4)$, B is efficiently expressed with $({}^1/_4, {}^1/_6, {}^1/_8, {}^1/_{10})$. Thus, the relationship between A and B is non-linear. To linearize their relationship, the repression function has to be applied to the measurement values of A. The parameter β has to be estimated, where the parameter α corresponds to the value in the weight matrix W. In this example the parameter β is assumed to be known. The repression function is applied to A without the α parameter yielding A' $({}^1/_2, {}^1/_3, {}^1/_4, {}^1/_5)$. Now the relationship of A' and B is linear. The qualitative form of the non-linear function has to be introduced as a prior to the model, where α is inferred by the reconstruction method. Because it is not obvious how to tailor this function, we will discuss the choice of how to tailor this function in the following section.

III. Model of the Regulatory Network

The aim of this work is to validate the reconstruction with the cI circuit. The qualitative structure and the quantitative parameters of the cI circuit are provided in several publications [1], [13]. Theoretical analysis was applied by Bundschuh et al. [3] to simplify the model while preserving its original behavior. This simplification will be important and necessary for the modeling phase.

The cI circuit is part of a larger regulatory process, which functions as a stochastic bistable switch [1]. This property may hamper the reconstruction process on biological data. However, in this study the model is treated as a generic model. For further analysis it may be of importance that high temporal-resolution data is available [13]. This allows us to investigate larger models or reconstructions based on in vitro measurement data.

The complete model is shown in Fig. 2. DNA (D) is coding for the mRNA (M) which is translated by the RNA polymerase (R). The protein monomer (P) transcribed from the mRNA also occurs in a dimerized form (P_2). The dimer reversibly binds to the DNA creating an inactive DNA-dimer complex (Q). Free DNA is bound by the RNA polymerase (D^*), which dissociates when releasing the translated mRNA. Protein and

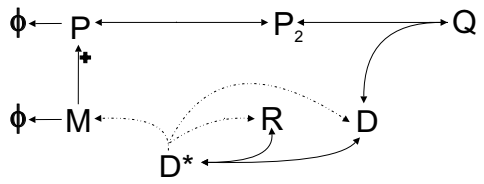

Fig. 2. Complete kinetic model of cI circuit. The metabolites are DNA (D), protein (P), RNA polymerase (R), mRNA (M), protein dimer (P_2), DNA with bound RNA polymerase (D^*) and the DNA-P_2 complex (Q). The two-sided arrows indicate reversible reactions, where the one-sided arrows indicate production or degradation. The dashed arrows from D^* describe the production of mRNA followed by the dissociation of the RNA polymerase from the DNA.

mRNA molecules are constantly degraded, described by a reaction to ϕ.

The complete kinetic model is more complex than the proposed reverse engineering model. To overcome this gap the kinetic model can be simplified or the mathematical model can be extended. The work of Bundschuh et al. [3] deals with the simplification of the kinetic model by removing fast reactions. These reactions are assumed to equilibrate on longer timescales. Slow reactions are either kept or lumped into a single non-linear term. The non-linear terms were modeled by Michaelis-Menten and Hill kinetics, yielding indistinguishable protein distributions in case of the Michaelis-Menten equations. Therefore we choose the Michaelis-Menten equations as abstract representation of regulator-gene interactions.

The simplified model contains only three components shown in Fig. 3. There, the mRNA produces proteins which in turn form dimers. Again, protein and mRNA molecules are constantly degraded. Dimer abundance regulates mRNA production through a Michaelis-Menten transcription term. The derivation of the Michaelis-Menten equations is discussed in [3]. Here, the transcription term of the negative feedback model is given in Eq. (4), with the parameters $k_M = 0.00616\,\mathrm{nM/s}$ and $K_M = 356\,\mathrm{nM}$.

$$k_{1,eff}\left([P_2]\right) = \frac{k_M}{1 + [P_2]/K_M} \tag{4}$$

A. Implementation and Simulation

To simulate the behavior of the complete cI circuit and the simplified counterpart, we implemented the models in JSim 1.6. JSim is a java-based simulation and animation environment program, distributed by the National Simulation Resource (NSR). For simulation the initial concentrations are set to zero, assuming the genes are switched of at time point zero. A complete parameterization is given in [3].

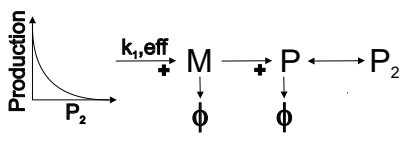

Fig. 3. Simplified kinetic model. Fast reactions have been removed, the production of mRNA (M) has been replaced by an effective transcription rate, with P_2 acting as repressor.

IV. Reconstruction and Prediction of the cI circuit

Measurements of JSim simulations provide data for the reverse engineering. The measurements are taken every 1000 seconds starting at time point 0. Since we do not want to work on a known topology, we incorporate measurement data from other models. These models are as well taken from Bundschuh et. al [3], and are derived from the original model.

We begin the reverse engineering process with a basic linear model. Then we expand it to predict the behavior of the cI circuit.

A. Linear Model

For the reconstruction on the basic linear model, no considerations are taken. Prior information is not needed and the process is free of parameters. A degradation term can not be included since linear self-regulation and degradation are not distinguishable and would collapse to one variable. The initial mRNA value was set to 0.1, since the linear model would be unsolvable otherwise.

B. Simple Non-Linear Model

Previous analysis suggested that there are intrinsic non-linearities in self-regulatory systems [3]. Therefore, we extend the model to capture interaction terms such as Eq. (4) by applying them for h (see Eq. (3)). In contrast to the linear model a first-order degradation term is incorporated. The non-linear interaction term is algebraically distinguishable from first-order degradation. A non-linear activation term g is not incorporated because there is no non-linear counterpart in the simplified model given in Fig. 3.

In this model it is assumed that the interaction can be described by Michaelis-Menten kinetics. To keep the model minimal, we remove the K_M parameter by setting it to 1. The first-order degradation is linear, therefore we have to supply linear and non-linear data to the reconstruction method. Non-linear data is obtained by applying g. Again, this provides a model without any parameters.

C. Expanded Non-Linear Model

So far the models only include mRNA. With one component the model behavior is restricted to quite simple dynamics. To emulate the simplified cI circuit with high agreement to the original model, we introduced P_2 into the reconstruction process as a second component. In order to maintain the linear reconstruction method we have to precompute P_2 from the mRNA measurement data. To accomplish this, we describe P_2 by a simple differential equation, containing first-order production and degradation rates. The production is based on the amount of mRNA and the degradation on the amount of P_2. To fit the parameters of the differential equation, we measure the P_2 abundance during steady state. Now the input data for the mRNA(t) is no longer a non-linear form of

mRNA$(t-1)$, but a non-linear form of P_2, resembling the regulation described by Eq. (4).

For this approach, protein measurement data is necessary, which is not available in most cases. However, here only the steady-state protein abundance must be provided. This makes the modeling approach more realistic, because steady-state protein abundance is more readily available than protein time-series.

V. Results

JSim simulations of the complete and simplified cI circuits are presented. The evaluation of the reconstruction covers the mathematical properties and in a subsequent section the reconstruction of a detailed kinetic model.

A. Properties of the Reconstruction Method

First, we performed evaluations on noise-free data with linear models. The reconstruction method revealed the correct network structure with the probability given in Eq. (2), providing $k+1$ equations. Less than $k+1$ equations lead to arbitrary reconstruction results. This also holds for a simple non-linear model with known non-linearities.

On noisy data with a standard deviation of 5%, the quality of the reconstruction decreases significantly, but remains on a high level. With increased noise, the reconstruction performance decreased in specificity and sensitivity until it reached the level of random guesses. The sensitivity to network size turned out to be very low. 13 measurements turned out to be optimal to reconstruct a network with 40 genes and a standard deviation of 5%. Less measurements, and interestingly also an increased number of measurements, lead to decreased performance. The results are shown in Fig. 4.

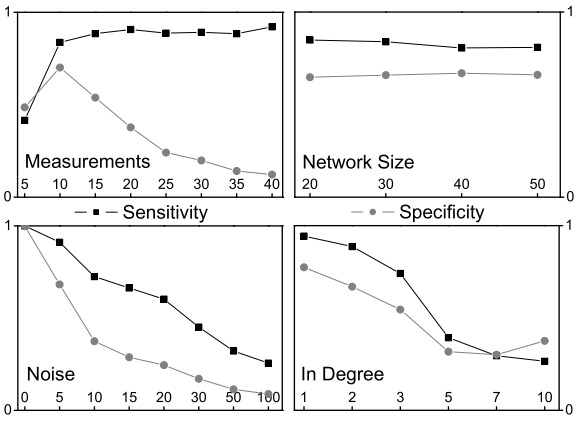

Fig. 4. Reconstruction of artificial networks under different conditions. The networks consist of 40 genes and a maximal in-degree of 3. 10 noisy measurements are provided for reconstruction with a standard deviation of 5%. In every plot, one property is modified while the others are kept constant. The sensitivity and specificity of the reconstructed regulatory connections are shown in the plots.

B. Model and Topology Reconstruction

In this section, we try to reveal the network structure and obtain a predictive model of the cI circuit. The predictive model is evaluated only considering one gene, which implies a topology of self regulation. Revealing the topology is accomplished by adding additional data from two other genes not related to the gene of interest. This provides a network of three genes. The model fit and the topological reconstruction are evaluated independently.

The number of available measurements is of major concern. Therefore, we evaluate the reconstruction under a rich amount of measurements (10) and a minimal amount of measurements (3).

In the extension of the non-linear model, P_2 was introduced as a second component. P_2 was modeled directly from mRNA data leading to a moderately good fit in which the deviation of the simulated and calculated P_2 deviated up to 50% after the fist time step, then they converged to the same steady-state. This model was fitted with the parameters $K_f = 0.01$, $K_b = 0.0001$ and $P_2(0) = 0$.

C. Reconstruction with Three Measurements

To evaluate the reconstruction under a minimal amount of data, three measurements were used for the reconstruction. The linear model had a weight of 0.6764 leading to an enormous growth of mRNA. The concentration at timepoint 20,000 sec is 44.32 mM. Fig. 5 shows the simulation.

The non-linear model has similar dynamics, but due to the reduced number of measurements, the deviation from the kinetic model grows.

The reconstruction of the expanded non-linear model fits well even with a small amount of data. The dynamics are less complex, since the trajectory can be sketched by two lines.

D. Reconstruction with Ten Measurements

The reconstruction with the linear model completely fails to reproduce the original behavior. The gene is positively self regulated with a weight of 0.0748. This leads to a divergent system in all cases, which can be seen in the Fig. 6.

The non-linear model is always reconstructed with a positive value for the regulation term K_M, and a negative term for the degradation. This leads to monotonically increasing mRNA abundance until the steady-state level is reached, where production and degradation compensate each other. The behavior of the model is shown in Fig. 6. It can be seen that the non-linear model increases monotonically, where the cI circuit decreases after reaching a local maximum. Both models converge to steady-state, deviating to some extent.

The reconstruction of the expanded model is slightly improved, where it was already in good agreement with the reference model with three measurements.

E. Topological Reconstruction

The results of the topological reconstruction are presented in Table I. The origin of the supplied data is given in the left column. Thereby, cI is mRNA measurement data from

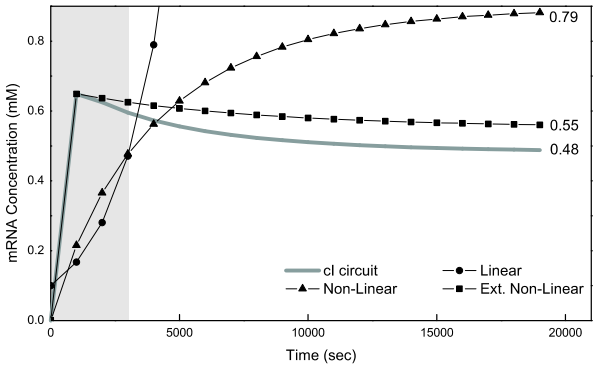

Fig. 5. Simulation of the mRNA concentration by different reconstruction models and the cI circuit. The models are reconstructed on three measurements. The gray background shows the time interval in which the measurements were taken.

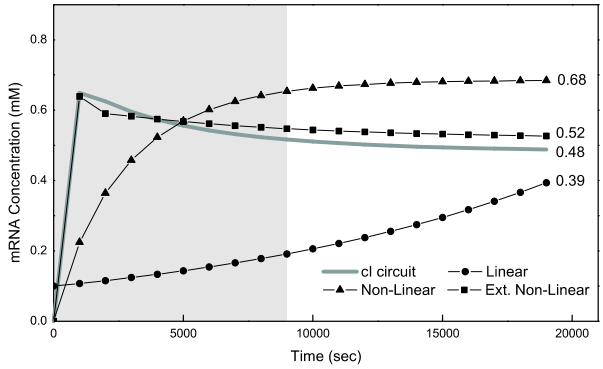

Fig. 6. Simulation of the mRNA concentration by different reconstruction models and the cI circuit. The models are reconstructed on 10 measurements. The gray background shows the time interval in which the measurements were taken.

the cI circuit and h a non-linear Michaelis-Menten term, $g1$ and $g2$ are mRNA measurements from the two other models. The qualitative structure of the target topology is given in the second column. The weights reconstructed with different models are given in the remaining columns. Note that the values are not comparable, since different model structures are employed. Therefore, we only consider the correct reconstruction of the topology, distinguishing activation "+" from repression or degradation "-". The bold numbers indicate a correct resemblance of the target network topology. The "0" values indicate, that these weights were explicitly set to zero by the reduction method.

In case of the linear model the topological reconstruction provided arbitrary connections between unrelated genes. The topological reconstruction with the non-linear and extended non-linear model revealed the correct connections in all cases.

VI. Discussion

We proposed a reconstruction method which can be adjusted to simple linear or different non-linear models. The mathematical properties of the method show to reproduce models of known structure with the minimal amount of $k+1$ equations. This draws an absolute lower bound for the reconstruction of (non)-linear networks, associated with the in-degree. Yeung et al. [23] assumed this bound to be a function of the network size $O(\log n)$. Associating the minimal amount of equations with the in-degree shifts the main complexity from the network size to the in-degree. Although, this only holds under noise-free conditions.

The reconstruction is robust to the network size, but sensitive to the in-degree, confirming the previous statement. The sensitivity to noise is high as well. Overall, our method reveals similar problems as earlier proposed methods [23]. However, it has the favorable property of being an analytical method with a probabilistic step. This provides a reconstruction method that reduces the solution space to determine the minimal network, where other methods optimize objectives only similar to sparsity.

The reconstruction of the kinetic model with a linear network structure appeared to be problematic. We could not see any resemblance of dynamic or steady-state behavior.

Adding a parameter-free Michaelis-Menten term to the model, leads to a better resemblance of the cI circuit. The dynamics were captured in part and the steady-state could be anticipated.

The extended non-linear model needs prior information, including initial protein concentrations, which cannot always be obtained. Although, the gain of adding protein abundance to the model is significant. The dynamics as well as the steady-state behavior of the original system could be simulated with high accuracy, even with a small amount of data. Interestingly, the need for data strongly decreases with an improved adjustment to the kinetic model.

During the modeling phase, we aimed to achieve minimal models. The basic linear model conjoint with sparsity provides this. Non-linearity was introduced with a parameter free Michaelis-Menten interaction term. The extended model determines P_2 with two parameters. To fit these, the steady-state concentration of P_2 was provided. This may render the extended model inapplicable for global-scale reconstruction. However, the parameter-free non-linear model may be well suited for this task.

In this study we investigated one gene circuit and successfully applied a Michaelis-Menten model. Although the Michaelis-Menten model occurs frequently in biochemical reactions, it remains unknown if it can describe a large range of regulatory influences. Genes with complex regulatory patterns may be problematic to model as well, since summing up the regulatory inputs may not provide predictive models.

For our reverse engineering we could not provide a validation on previously unknown data. Genetic networks are poorly understood and a dataset for validation is not available. Therefore, we had to retreat to Occam's razor to guide our modeling. The proposed models accomplish different degrees of minimality. The linear model implies assumptions described in previous works [23]. The main extension incorporated here was the parameter-free Michaelis-Menten term, which can

TABLE I

TOPOLOGICAL RECONSTRUCTION WITH DIFFERENT MODELS AND DATA SETS

		Reconstruction with three measurements			**Reconstruction with ten measurements**		
data	cI circuit	linear model	non-linear model	ext. model	linear model	non-linear model	ext. model
$h(cI)$	$+$	N.A.	**+0.179**	**+0.686**	N.A	**+0.196**	**+0.606**
cI	$-$	0	**-0.133**	**-1.043**	**-0.213**	**-0.195**	**-1.093**
$g1$	$\circ$	+0.111	**+0.013**	**-0.061**	+0.265	**+0.004**	**-0.006**
$g2$	$\circ$	+0.181	**0**	**0**	+0.208	**-0.003**	**+0.001**

be regarded as a general term of interaction in biological systems. The extended model predicted dimer abundance by incorporating two parameters and utilizing protein steady-state measurements. Nonetheless, this model could predict the cI circuit very well.

Overall, the reconstruction method was well suited to reveal the underlying network structure. By elaborate considerations of the network model, we could reduce the amount of data needed significantly. Only a small number of parameters were introduced in case of the extended model, and none for the other models.

VII. FUTURE WORK

In further work we plan to reverse engineer established models [10] with in vitro PT-PCR and microarray measurements. The aim is to give a conceptual validation of the data necessary to reconstruct small to medium size networks.

ACKNOWLEDGEMENT

This work was supported by the National Genome Research Network (NGFN II) of the Federal Ministry of Education and Research in Germany under contract number 0313323.

REFERENCES

[1] A. Arkin, J. Ross, and H. H. McAdams, "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells." *Genetics*, vol. 149, no. 4, pp. 1633–48, Aug 1998.

[2] M. J. Buck and J. D. Lieb, "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments." *Genomics*, vol. 83, no. 3, pp. 349–60, Mar 2004.

[3] R. Bundschuh, F. Hayot, and C. Jayaprakash, "Fluctuations and slow variables in genetic networks." *Biophys J*, vol. 84, no. 3, pp. 1606–15, Mar 2003.

[4] K.-C. Chen, T.-Y. Wang, H.-H. Tseng, C.-Y. F. Huang, and C.-Y. Kao, "A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae." *Bioinformatics*, vol. 21, no. 12, pp. 2883–90, Jun 2005. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bti415

[5] T. Chen, H. L. He, and G. M. Church, "Modeling gene expression with differential equations." *Pac Symp Biocomput*, pp. 29–40, 1999.

[6] H. de Jong, "Modeling and simulation of genetic regulatory systems: a literature review." *J Comput Biol*, vol. 9, no. 1, pp. 67–103, 2002. [Online]. Available: http://dx.doi.org/10.1089/10665270252833208

[7] H. de Jong, M. Page, C. Hernandez, and J. Geiselmann, "Qualitative simulation of genetic regulatory networks: Method and application," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

[8] I. B. Dodd, K. E. Shearwin, and J. B. Egan, "Revisited gene regulation in bacteriophage lambda." *Curr Opin Genet Dev*, vol. 15, no. 2, pp. 145–52, Apr 2005. [Online]. Available: http://dx.doi.org/10.1016/j.gde.2005.02.001

[9] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling." *Science*, vol. 301, no. 5629, pp. 102–5, 05.05 2003. [Online]. Available: http://dx.doi.org/10.1126/science.1081900

[10] A. Gilman and A. P. Arkin, "Genetic "code": representations and dynamical models of genetic components and networks." *Annu Rev Genomics Hum Genet*, vol. 3, pp. 341–69, 05.05 2002. [Online]. Available: http://dx.doi.org/10.1146/annurev.genom.3.030502.111004

[11] D. G. Ginzinger, "Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream." *Exp Hematol*, vol. 30, no. 6, pp. 503–12, Jun 2002.

[12] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita, "Dynamic modeling of genetic networks using genetic algorithm and S-system." *Bioinformatics*, vol. 19, no. 5, pp. 643–50, Mar 2003.

[13] O. Kobiler, A. Rokney, N. Friedman, D. L. Court, J. Stavans, and A. B. Oppenheim, "Quantitative kinetic analysis of the bacteriophage lambda genetic network." *Proc Natl Acad Sci U S A*, vol. 102, no. 12, pp. 4470–5, Mar 2005. [Online]. Available: http://dx.doi.org/10.1073/pnas.0500670102

[14] S. Liang, S. Fuhrman, and R. Somogyi, "Reveal, a general reverse engineering algorithm for inference of genetic network architectures." *Pac Symp Biocomput*, pp. 18–29, 1998.

[15] G. MacBeath, "Protein microarrays and proteomics." *Nat Genet*, vol. 32 Suppl, pp. 526–32, Dec 2002. [Online]. Available: http://dx.doi.org/10.1038/ng1037

[16] W. H. Press, S. S. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C++ : The Art of Scientific Computing*. Cambridge University Press, 2002.

[17] M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon, "Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics." *Proc Natl Acad Sci U S A*, vol. 99, no. 16, pp. 10 555–60, Aug 2002. [Online]. Available: http://dx.doi.org/10.1073/pnas.152046799

[18] A. Schulze and J. Downward, "Navigating gene expression using microarrays–a technology review." *Nat Cell Biol*, vol. 3, no. 8, pp. E190–5, Aug 2001. [Online]. Available: http://dx.doi.org/10.1038/35087138

[19] C. Spieth, F. Streichert, N. Speer, and A. Zell, "A memetic inference method for gene regulatory networks based on s-systems," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2004)*, 2004, pp. 152–157.

[20] D. Thieffry, A. Huerta, E. Perez-Rueda, and J. Collado-Vides, "From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in escherichia coli." *Bioessays*, vol. 20, pp. 433–440, 1998.

[21] E. P. van Someren, L. F. Wessels, and M. J. Reinders, "Linear modeling of genetic networks from experimental data." *Proc Int Conf Intell Syst Mol Biol*, vol. 8, pp. 355–66, 2000.

[22] D. C. Weaver, C. T. Workman, and G. D. Stormo, "Modeling regulatory networks with weight matrices." *Pac Symp Biocomput*, pp. 112–23, 1999.

[23] M. K. S. Yeung, J. Tegnr, and J. J. Collins, "Reverse engineering gene networks using singular value decomposition and robust regression." *Proc Natl Acad Sci U S A*, vol. 99, no. 9, pp. 6163–8, Apr 2002. [Online]. Available: http://dx.doi.org/10.1073/pnas.092576199