

# Robust Identification of Piecewise Linear Gene-Protein Interaction Networks

Ronald L. Westra, Ralf L. Peeters

Dept. Mathematics, Universiteit Maastricht  
P.O. Box 616, 6200 MD Maastricht, The Netherlands  
E-mail: {westra, ralf.peeters}@math.unimaas.nl

## Abstract

In this study we will focus on piece-wise linear state space models for gene-protein interaction networks. We will follow the dynamical systems approach with special interest for partitioned state spaces. From the observation that the dynamics in natural systems tends to punctuated equilibria, we will focus on piecewise linear models and sparse and hierarchic interactions, as for instance described by Glass, Kauffman, and de Jong. Next, the paper is concerned with the identification (a.k.a. reverse engineering and reconstruction) of dynamic genetic networks from microarray data. We will describe exact and robust methods for computing the interaction matrix in the special case of piecewise linear models with sparse and hierarchic interactions from partial observations. Finally, we will analyze and evaluate this approach with regard to its performance and robustness towards intrinsic and extrinsic noise.

Keywords: piecewise linear, robust identification, hierarchical networks, gene expression data, gene regulatory networks.

## 1 Introduction

This paper is concerned with the identification of dynamic gene-protein interaction networks with intrinsic and extrinsic noise from empirical data, such as a set of microarray time series.

Prerequisite for the successful reconstruction of these networks is the way in which the dynamics of their interactions is modeled. The formal mathematical modeling of these interactions is an emerging field where an array of approaches are being attempted, all with their own problems and short-comings. The underlying physical and chemical processes involved are multifarious and hugely complex. This condition contrasts sharply with the modeling of inanimate Nature by physics. While in physics huge quantities of but a small amount of basic types of elementary particles interact in a uniform and deterministic way provided by the fundamental laws of nature, the situation in gene-protein interactions deals with tens of thousands of genes and possibly some million proteins. The quantities thereby involved in the actual interactions are normally very small, as one single protein may be able to (in)activate a specific gene, and thereby change the global state of the system. For this reason, gene regulatory systems are much more prone to stochastic fluctuations than the interactions involved in normal anorganic reactions. Moreover, each of these interactions is different and involves its own peculiar geometrical and electrostatic details. There are different processes involved like transcription, translation and subsequent folding. Therefore, the emergent complexity resulting from gene regulatory networks is much more difficult to comprehend.

In the past few decades a number of different formalisms for modeling the interactions amongst genes and proteins have been presented. Some authors focus on specific detailed processes such as the circadian

rhythms in *Drosophila* and *Neurospora* [10], [11], or the cell cycle in *Schizosaccharomyces* (Fission yeast) [14]. Others try to provide a general platform for modeling the interactions between genes and proteins. For a thorough overview consult de Jong (2002) in [2], Bower (2001) in [1], and others [6], [13].

We will focus on dynamical models, and not discuss static models where the relations between genes are considered fixed in time. In discrete event simulation models the detailed biochemical interactions are studied. Considering a large number of constituents, the approach aims to derive macroscopic quantities. More information on discrete event modeling can be found in [1].

In section 2 we will review formalisms for modeling the interactions amongst genes and proteins. We will shortly discuss how kinetic rate equations relate to (stochastic) differential equations (SDEs), and sketch how from such SDEs we can model the interactions as piecewise linear systems. In section 3 we discuss how these mathematical models can be employed to reconstruct their view on the interactions from empirical data. In section 5 we consider the lessons we can learn from Nature in designing control systems robust towards stochastic fluctuations and chaos.

## 2 Modeling gene-protein interactions as a piecewise linear system

The traditional approach to modeling the dynamical interactions amongst genes and proteins is by considering them as biochemical reactions, and thus representing them as 'rate equations'. The concept of chemical rate equations consists of a set of differential equations, expressing the time derivative of the concentration of each constituent of the reaction as some rational function of the concentrations of all the constituents involved. Though the truth of the underlying biochemical interactions between the constituents is generally accepted, a rate equation is not a fundamental law of Nature, but a statistical average over the entire ensemble of molecular collisions that contribute to an actual chemical reaction [22]. So, rate equations are statistical approximations that – under certain conditions – predict the average number of reactive collisions. The actual observed number will fluctuate around this number, depending on the details of the microscopic processes involved. In case of biochemical interactions between genes and proteins the applicability of the concept of rate equations is valid only for genes with sufficient high transcription rates. This is confirmed by recent experimental findings by Swain and Elowitz [5], [16], [18], [19].

From the above, we may conclude that modeling can only be successful for genes with sufficiently high transcription rates. Even in the optimal case, we would obtain a high-dimensional (reflecting the number of genes, RNAs, and proteins involved – so tens of thousands), non-linear, differential equation, that is subject to substantial stochastic fluctuations. Much more problematic is the fact that the precise details of most reactions are unknown, and therefore cannot be modeled as rate equation. This could be compensated by a well-defined parametrized generic form of the interactions, such that the parameters could be estimated from sufficient empirical data. A generic form based on rational positive functions is proposed by J. van Schuppen [23]. However, in the few cases where parts of such interaction networks have been described from experimental analysis, like the circadian rhythms in certain amoeba [10], or the cell cycle in fission yeast [14], it is clear that such forms have a too extensive syntax to be of any practical use.

Let us for the moment forsake these problems, and consider the dynamics of gene-RNA-protein networks. When we assume a stochastic differential equation as model for the dynamics of the interaction network, the relation can be expressed as:

$$\dot{x} = f(x, u|\theta) + \xi(t) \tag{1}$$

Here  $x(t)$ , called the state-vector, denotes the  $N$  gene expressions and  $M$  RNA/protein densities at time  $t$  – possibly involving higher order time derivatives.  $u(t)$  denotes the  $P$  controlled inputs to the system, such as the timing and concentrations of toxic agents administered to the system observed.  $\xi(t)$  denotes a stochastic Gaussian white noise term. This expression involves a parameter vector  $\theta$ , that contains the

coupling constants between gene expressions and protein densities. We can consider this system as being represented by the state vector  $x(t)$  that wanders through the (at least)  $(N + M)$ -dimensional space of all possible configurations. In the formalism of dynamic systems theory, eventually  $x$  will enter an area of attraction, and become subject to the influence of an attractor. An attractor here can be an uniform convergent attractor, a limit cycle, or a 'strange attractor'. We can understand the entire space as being partitioned into cells, where such attractors – or their antagonists so-called repellers – reign. Thus, the behavior of  $x$  can be described by motion through this collection of cells, swiftly moving through cells of repellers, until they enter the basin of attraction of an attractor. Under the effects of external agents via the vector  $u(t)$  or by stochastic fluctuations via  $\xi(t)$  they can leave this cell, and start wandering again, thereby repeating the process. Now, a vital assumption is that in each cell the behavior is governed by specific (un)stable equilibrium points, and therefore it is possible to make a linear approximation of equation 1 in the cell with index  $l$  as:

$$\dot{x}(t) = F_l x(t) + G_l u(t) \quad (2)$$

In case of a uniform attractor the largest eigen-value of  $F_l$  will be negative, and in case of a uniform repeller the smallest eigen-value will be positive. We can now formalize the qualitative behavioral dynamics of gene-protein interactions as predominantly linear behavior near the stable equilibria – called the steady states, interrupted by abrupt transitions where the system quickly relaxes to a new steady state, either externally induced or by process noise.

In biology such behavior is frequently observed, as for instance in embryonic growth where the organism develops by transitions through a number of well-defined 'check points'. Within each such checkpoint the system is in relative equilibrium. There is an ongoing debate on mathematical modeling of cell division as *checkpoint mechanisms* versus *limit-cycle oscillators*, see [20]. We will follow the view of *piecewise linear behavior* (PWL, also known more appropriately as *piecewise affine* behavior). This approach corresponds to the piecewise linear models introduced by Glass and Kauffman [9], and the qualitative piecewise linear models described by de Jong et al. [2], [3].

### 3 The identification of *piecewise linear networks* by $L_1$ -minimization

Next, we will be concerned with the identification (a.k.a. *reverse engineering* or *reconstruction*) of piecewise linear gene regulatory systems from microarray data. The nature of our problem – few microarray experiments and lots of genes – implies that we are dealing with *poor data* (as opposed to *rich data*), where the number of measurements is *a priori* insufficient to identify all parameters of the system. One standard approach to circumvent this problem is by dimension reduction through the clustering of related genes. We consider the case where time series of genome-wide expression data is available. The case of the identification of a *simple* linear system is discussed in Peeters and Westra [15], [26], and Yeung et al. in [27]. In the following, we will be concerned with the identification of *piecewise* linear systems. Our aim is to obtain the gene-gene interaction matrix. This matrix can be interpreted as a connectivity matrix, and so directly relates to the graph of the gene regulatory network. With this network we are able to make statements like: 'the expression of this gene causes that and that cluster of genes to alter their expression in this and this way'.

Let us in the following assume a dynamical input-output system  $\Sigma$  that switches irregularly between  $K$  linear time-invariant subsystems  $\{\Sigma_1, \Sigma_2, \dots, \Sigma_K\}$ . Let  $S = \{s_1, s_2, \dots, s_{K-1}\}$  denote the set of – possibly unknown – switching times, i.e. the time instants  $t = s_l$  that the system switches from subsystem  $\Sigma_l$  to  $\Sigma_{l+1}$ . Similarly as with the simple linear networks, we assume *Hankel matrices*  $X = (x[1], x[2], \dots, x[M])$ , and  $U = (u[1], u[2], \dots, u[M])$  at  $M$  sampling times  $T = \{t_1, t_2, \dots, t_M\}$ , representing full observations of the  $N$  states and  $P$  inputs. The interval between two sample instants is denoted as  $\tau_k = t_{k+1} - t_k$ . In first instance we assume that the system is sampled on regular time intervals, i.e. that the sample intervals are equal to  $\tau$ .

Within one subsystem  $\Sigma_l$  the relation between the inputs  $u(t)$  and outputs  $y(t)$  is represented as a state-space system of first-order differential (for continuous time systems) or difference equations (for discrete time systems), using an auxiliary vector  $x(t)$  spanning the so-called subspace.

Continuous time:

$$\dot{x}(t) = F_l x(t) + G_l u(t), \quad (3)$$

$$y(t) = H_l x(t) + J_l u(t). \quad (4)$$

Discrete time:

$$x[k + 1] = A_l x[k] + B_l u[k], \quad (5)$$

$$y[k] = C_l x[k] + D_l u[k]. \quad (6)$$

The relation between these is given by:

$$A_l = e^{\tau F_l}, \quad (7)$$

$$B_l = e^{\tau F_l} G_l. \quad (8)$$

with  $x[k] = x(t_k)$ .

### 3.1 Determination of the new state equilibrium points

Moreover, in each new state the new equilibrium point  $\mu_l \in \mathbb{R}^N$  has also to be established. The linearization near  $\mu_l$  can be written as:

$$\frac{\partial}{\partial t}(\mu_l + (x - \mu_l)) = F_l(x - \mu_l) + G_l u + O(\|x - \mu_l\|^2) \quad (9)$$

which can be rewritten as:  $\dot{x} = F_l x + \tilde{G}_l \tilde{u}$ , with:

$$\tilde{G}_l = (G_l - F_l \mu_l), \quad (10)$$

$$\tilde{u} = \begin{pmatrix} u \\ 1 \end{pmatrix}. \quad (11)$$

The reasoning is similar in the discrete case, and we obtain:  $x[k + 1] = A_l x[k] + \tilde{B}_l \tilde{u}[k]$ . Therefore, we can follow the original formulation and, using  $\tilde{u}$  rather than  $u$  as input, estimate  $A_l$  and  $\tilde{B}_l$ , and using:

$$\tilde{B}_l = (B_l - A_l \mu_l), \quad (12)$$

to compute  $\mu_l$  and  $B$ . We will follow this approach, and from here on drop the *tilde*, and simple write  $B_l$  for  $(B_l - A_l \mu_l)$ , and  $u[k]$  for  $\begin{pmatrix} u[k] \\ 1 \end{pmatrix}$ .

### 3.2 General dynamics of switching subsystems

In the context of piecewise linear systems of gene regulatory systems, the dynamics is slightly different to the case of simple linear systems as in [15]. In our context we assume that we observe *all*  $N$  genes, and that there is no direct through-put. This means that  $C_l = I$  and  $D_l = 0$  for all  $l$ . Therefore, we can suffice with equation 5 corrected for the equilibrium point:

$$x[k + 1] = A_l x[k] + B_l u[k]. \quad (13)$$

We furthermore assume that the system matrices in these equations are constant during intervals  $[s_l, s_{l+1}]$ , and abruptly change at the transition between the intervals at  $t = s_{l+1}$ . We assume that on the time scale  $\tau$  the system has relaxed to its new state. This means that we do not observe *mixed states*, which would severely complicate the problem of identification.

Finally, we define the *weights*  $w_{kl}$ , as the membership functions of observation  $k$  to subsystem  $\Sigma_l$ ; if observation  $\{x[k], u[k]\}$  belongs to system  $\Sigma_l$  then  $w_{kl} = 1$ , if  $\{x[k], u[k]\}$  does not belong to  $\Sigma_l$  then  $w_{kl} = 0$ . This definition allows for a *fuzzy* definition of weight, such that  $w_{kl} \in [0, 1]$ . *A priori*, we thus can state two constraints on  $w$ :

$$\forall_{k,l} w_{kl} \in [0, 1], \quad (14)$$

$$\forall_l \sum_l w_{kl} = 1. \quad (15)$$

The challenge in system identification is to estimate the relevant model parameters in piecewise linear dynamics from empirical observations. The success of this approach depends on the amounts of empirical data available – *rich* or *poor*, the validity of the mathematical model, the levels of process noise and measuring noise, and the nature of the sampling process. In case of regular sampling the discrete model 5 can be applied which leads to more straightforward techniques than the continuous model 3 that should be used in case of irregular sampling. In the following sections we will study a number of these conditions in more detail.

### 3.3 Identification of PWL models with *unknown* switching and *regular* sampling from *poor* data

The assumption that the switching times between the linear subsystems are completely known suits various experimental conditions, as for instance when toxic agents are administered. In many biological situations, however, the exact timing between subsystems is not known, as during embryonic growth and in many metabolic processes.

#### 3.3.1 As an extension to the simple linear systems in case state derivatives are available

When a sufficiently accurate record of estimates of the state derivatives  $\dot{X} = \{\dot{x}[1], \dot{x}[2], \dots, \dot{x}[M]\}$  is available, we can simply rewrite this problem as a special case of the method described in the case of a simple linear problem as in [15]. In fact, by exploiting the data  $\mathcal{D} = \{X, U, \dot{X}\}$ , the problem can be stated as a linear equation in terms of new matrices  $H_1$  and  $H_2$  as:

$$\dot{X} = H_1 X + H_2 U. \quad (16)$$

In this equation the matrices  $H_1$  and  $H_2$  relate to the – unknown – system matrices  $\{A_1, B_1, \dots, A_K, B_K\}$  and ditto unknown weights  $\{w_{kl}\}$  as:

$$\text{vec}(H_1) = W \cdot \text{vec}(A), \quad (17)$$

$$\text{vec}(H_2) = W \cdot \text{vec}(B). \quad (18)$$

The matrices  $A$ ,  $B$ , and  $W$  are composed as follows:

$$A = \begin{pmatrix} A_1 \\ \dots \\ A_K \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ \dots \\ B_K \end{pmatrix}, \quad W = w \otimes I_{N^2} = \begin{pmatrix} w_{1,1}I_{N^2} & \dots & w_{1,K}I_{N^2} \\ \dots & \dots & \dots \\ w_{M,1}I_{N^2} & \dots & w_{M,K}I_{N^2} \end{pmatrix}, \quad (19)$$

where  $\otimes$  is the Kronecker-product, and  $I_{N^2}$  is the  $N^2 \times N^2$  identity matrix. Note that equation 16 is not anymore a linear problem, as the unknown matrices  $A$ ,  $B$ , and  $W$  appear in a non-linear way in the equation. This equation is exactly of the type of simple linear networks as in [15]. Therefore, its solution method is fully applicable, so that an efficient and accurate algorithm is available for solving this problem in terms of  $H_1$  and  $H_2$ . However, now the problem has shifted to solving two additional non-linear equations:

$$W \diamond A = H_1, \quad (20)$$

$$W \diamond B = H_2. \quad (21)$$

where  $A$ ,  $B$ , and  $W$  have to be solved from the known – i.e. computed – matrices  $H_1$  and  $H_2$ . The operation  $\diamond$  makes the relations in equations 20 and 18 explicit. This is an underdetermined system that can only be solved by additional information, such as assuming sparsity for  $A$ , and a block structure for  $W$ , such as the two constraints in equations 14 and 15.

This non-linear problem can thus be solved in terms of  $H_1$  and  $H_2$ , but not in terms of  $A$ ,  $B$ , and  $W$ . It is a bilinear problem in terms of  $A$  and  $B$  for fixed  $W$ , otherwise it is a quadratic problem. As a quadratic programming problem this is not a well-posed problem, i.e. it has a nonsingular Jacobian at optimality and is ill-conditioned as the iterates approach optimality. Therefore, we follow a different approach and split the problem in two LP-problems that are well-posed. The approach is as follows: (i) initialize  $A$ ,  $B$ , and  $W$ , (ii) perform the iteration:

1. Compute  $H_1$  and  $H_2$ , using the approach from Peeters and Westra [15] on equation 16,
2. Using fixed values for the weights  $W$ , compute  $A$  and  $B$  using equations 20, and 21,
3. Using fixed values for matrices  $A$  and  $B$ , compute the weights  $W$  using equations 14, 15, 20, and 21,

until: (iii) a cumulative weighted error criterion  $\mathcal{E}$  has converged sufficiently – or a maximum number of iterations has passed. A proper choice for the criterion function is:

$$\mathcal{E}(A, B, W|\mathcal{D}) = \sum_{k,l} w_{kl} \|A_l x[k] + B_l u[k] - \dot{x}[k]\|_2^2 \quad (22)$$

This problem can be solved by minimizing the quadratic  $L_2$ -criterion subject to mentioned constraints, for instance by a gradient descent method. We can, however, formulate a different approach for solving this problem by defining an alternative criterion function  $\mathcal{E}$ , namely as a linear  $L_1$ -criterion:

$$\mathcal{E}_1(A, B, W|\mathcal{D}) = \sum_{k,l} w_{kl} \|A_l x[k] + B_l u[k] - \dot{x}[k]\|_1 \quad (23)$$

This expression allows for an LP-formulation of the problem, in which  $\mathcal{E}_1$  serves as the objective function. Thus, we can split the non-linear optimization problem as two separate LP-formulations that are successively applied in the iteration; (i) an LP-problem  $LP_1$  for obtaining the system matrices  $A$  and  $B$  from minimizing objective function  $\mathcal{E}_1$  with given weights  $w$ , subject to the constraints in equations 20 and 21; and (ii) an LP-problem  $LP_2$  for obtaining the weights  $w$  from minimizing objective function  $\mathcal{E}_1$  with given system matrices  $A$  and  $B$ , subject to the constraints in equations 14, 15, 20, and 21.

We will revisit this philosophy in the next Section, when reviewing the more realistic case when the state derivatives of the gene expressions are *not* available.

### 3.3.2 Using the discrete dynamics of state-space representations

Now we will focus on the general case that only the state vectors  $X$  – i.e. the genome wide expressions – and the external inputs  $U$  are available as empirical data  $\mathcal{D}$ . In this case the objective of system identification

is to concurrently compute the system parameters  $A$ ,  $B$ , and weights  $W$ , and the switching times  $S = \{s_1, \dots, s_{K-1}\}$ <sup>1</sup>. Equation 5 provides us with the state space equations for such a system. In absence of noise the model evolution of the system for  $t_k \in [s_l, s_{l+1}]$  can be represented as:

$$x[k+1] = A_l x[k] + B_l u[k] \quad (24)$$

In practical experimental conditions, white process and measuring noise adds to the righthand side. The fit between the empirical data and the model can be quantified by the weighted difference between observed and expected expression profiles expressed as a linear  $L_1$ -criterion:

$$\mathcal{E}(\xi, w|\mathcal{D}) = \sum_{k,l} w_{kl} \|A_l x[k] + B_l u[k] - x[k+1]\|_1 \quad (25)$$

Here,  $\xi \equiv \begin{pmatrix} \text{vec}(\mathcal{A}) \\ \text{vec}(\mathcal{B}) \end{pmatrix}$  represents the set of system parameters  $\Theta$ , and  $\mathcal{D} \equiv \{X, U, T\}$  the observed Hankel matrices, i.e. the measured genome-wide expressions  $X$ , external inputs  $U$ , and sampling instances  $T$ . The criterion furthermore involves the relation between the  $k$ -th observation and the  $l$ -th subsystem  $\Sigma_l$ ; namely the *weight*  $w_{kl}$  that denotes the membership of  $k$ -th observation to subsystem  $\Sigma_l$ , and the *distance*  $d_{kl}$  between observed and the expected value of observation  $k$  relative to subsystem model  $\Sigma_l$ , as stated in equation 24. The problem of estimating the system parameters can thus formally be defined as the search for the vectors  $\xi^*$  and  $w^*$  that globally minimize  $\mathcal{E}$ .

QP: given the data  $\mathcal{D}$ , compute the system matrices  $\xi$  and the weight matrix  $w$ :

$$(\xi^*, w^*) = \arg \min_{\xi \in \mathbb{R}^{N(P+N)}, w \in \mathbb{R}^{KM}} \mathcal{E}(\xi, w|\mathcal{D}) \quad (26)$$

subject to:

$$\begin{aligned} \mathcal{E}(\xi|\mathcal{D}) &= \sum_{k=1}^{M-1} \sum_{l=1}^K d_{kl} w_{kl} \\ \forall_{k \in \{1, \dots, M-1\}} \forall_l d_{kl} &= \|x[k+1] - A_l x[k] - B_l u[k]\|_1 \\ \forall_{k,l} w_{kl} &\in [0, 1] \\ \forall_l \sum_l w_{kl} &= 1 \end{aligned}$$

This is a convex quadratic optimization problem that is, similar to the case in the previous Section, not well posed because it has a nonsingular Jacobian at the optimum, and becomes ill-conditioned as the iterates approach optimality. Instead of this quadratic programming problem we will therefore again study the two coupled linear programming problems associated to the original.

LP1: given weight matrix  $\tilde{w}$  compute the system matrices  $\xi^*$ :

$$\xi^* = \arg \min_{\xi \in \mathbb{R}^{N(P+N)}} \mathcal{E}(\xi, \tilde{w}|\mathcal{D}) \quad (27)$$

subject to:

$$\begin{aligned} \mathcal{E}(\xi^*, \tilde{w}|\mathcal{D}) &= \sum_{k=1}^{M-1} \sum_{l=1}^K d_{kl} \tilde{w}_{kl} \\ \forall_{k \in \{1, \dots, M-1\}} \forall_l d_{kl} &= \|x[k+1] - A_l x[k] - B_l u[k]\|_1 \end{aligned}$$

LP2: given system matrices  $\tilde{\xi}$  – and therefore the  $L_1$ -distances  $\tilde{d}_{kl} = \|x[k+1] - \tilde{A}_l x[k] - \tilde{B}_l u[k]\|_1$  – compute the weight matrix  $w^*$ :

$$w^* = \arg \min_{w \in \mathbb{R}^{KM}} \mathcal{E}(\tilde{\xi}, w|\mathcal{D}) \quad (28)$$

subject to:

$$\begin{aligned} \mathcal{E}(\tilde{\xi}, w|\mathcal{D}) &= \sum_{k=1}^{M-1} \sum_{l=1}^K \tilde{d}_{kl} w_{kl} \\ \forall_{k,l} w_{kl} &\in [0, 1] \\ \forall_l \sum_l w_{kl} &= 1 \end{aligned}$$

<sup>1</sup>There is a direct relation between switching instants  $S$  and the weights  $w$ , namely that within one time interval  $[s_n, s_{n+1}]$  the weights  $w_{kl}$ , with  $s_n \leq t_k < s_{n+1}$ , are equal.

As we assume that the subsystems  $\{\Sigma_1, \Sigma_2, \dots, \Sigma_K\}$  act disjointly and subsequently, the result can be improved by matching the weights to a block function structure; i.e. that  $w_{kl} = 1$  for  $t_k \in [s_l, s_{l+1}]$  and  $w_{kl} = 0$  elsewhere. This may, however, introduce other problems, for instance if the same subsystem is revisited at different switching intervals.

In this computation we have not yet involved the sparsity and the hierarchy of the interactions, resulting in the row-sparsity of matrices  $\{A_1, A_2, \dots, A_K\}$ . This can be achieved by adding a regularization term to the criterion, containing the  $L_1$ -vector norm of the  $A$ -matrices<sup>2</sup>:  $\|A\|_1 \equiv \sum_{l=1}^K \|A_l\|_1$ . The regularization term is coupled with a small given parameter  $\varepsilon_0$ . J.J. Fuchs [7], [8] has described conditions under which such a regularization term drives the optimization problem towards the global solution. Though these conditions do not strictly apply here, in numerical simulations we find that this approach succeeds. So, we obtain a reformulation of the first LP-problem in equation 27:

LP1\*: given weight matrix  $\tilde{w}$  compute the system matrices  $\xi^*$ :

$$\begin{aligned} \xi^* &= \arg \min_{\xi \in \mathbb{R}^{N(P+N)}} \mathcal{E}(\xi, \tilde{w}|\mathcal{D}) + \varepsilon_0 \|A\|_1 & (29) \\ \text{subject to:} & \\ \mathcal{E}(\xi, \tilde{w}|\mathcal{D}) &= \sum_{k=1}^{M-1} \sum_{l=1}^K d_{kl} \tilde{w}_{kl} \\ \forall_{k \in \{1, \dots, M-1\}} \forall_l & d_{kl} = \|x[k+1] - A_l x[k] - B_l u[k]\|_1 \end{aligned}$$

These two LP-problems can be solved efficiently with a partial dual simplex method as in [15], or by using large scale or interior points methods. The algorithm to estimate the system parameters  $\xi$  and  $w$  consists of iteratively solving the two optimizations LP1\* and LP2 subsequently, until the criterion  $\mathcal{E}(\xi, w|\mathcal{D}) + \varepsilon_0 \|A\|_1$  has sufficiently converged. Though the solution of the original quadratic programming problem QP in equation 26 is also the global solution of the two coupled LP-problems LP1\* and LP2, unfortunately there can exist local solutions to the couple {LP1\*,LP2}. For this reason, the approach has to be performed various times with different starting points.

## 4 Performance of the partial dual linear programming identification approach.

This approach resulted in an efficient and fast algorithm that is able to accurately estimate the gene-gene coupling matrix for tens of thousands of genes based on only several hundred genome wide measurements, and that is robust towards measurement noise. With increasing measurement noise or decreasing number of measurements the approach retains the strongest gene-gene coupling links - i.e. the largest modal value of the coupling matrix  $A$  - longest, see Figure 1.

A basic assumption in the approach is the sparsity of the underlying gene-gene coupling matrix, represented by the number of non-zero entries per row. If this number grows above a certain threshold the performance of the approach is severely affected, see Figure 2b. A number of numerical experiments were performed with this approach. The numerical experiments clearly demonstrate the range where the approach is effective. For relatively moderate noise levels and a high degree of sparsity i.e., a small number  $k$  of nonzero elements in the rows of matrix  $A$  - and not too many external stimuli  $p$ , the approach allows one to reconstruct a sparse matrix with great accuracy from a relative small number of observations  $M \ll N$ . For example, a row of  $A$  with 30,000 components of which all but 10 are equal to zero, can be efficiently reconstructed from just 150 independent measurements, see Figure 4a. The sparsity property of  $A$  fits in nicely with the technique of  $L_1$ -minimization, which automatically will always set many entries of the solution  $A^*$  to zero, whereas  $L_2$ -regression would spread out the error over all components, thus creating many

<sup>2</sup>The  $L_1$ -vector norm of  $A$ :  $\|A\|_{\text{vec},1} \equiv \|\text{vec}(A)\|_1$  differs from the  $L_1$ -matrix norm:  $\|A\|_{\text{mat},1} \equiv \max_{\|x\|=1} \frac{\|Ax\|_1}{\|x\|_1}$

small components. Reconstruction of large networks from this approach is straightforward: each of the rows of the gene-gene interaction matrix can be computed independently from the same set of micro-array experiments.

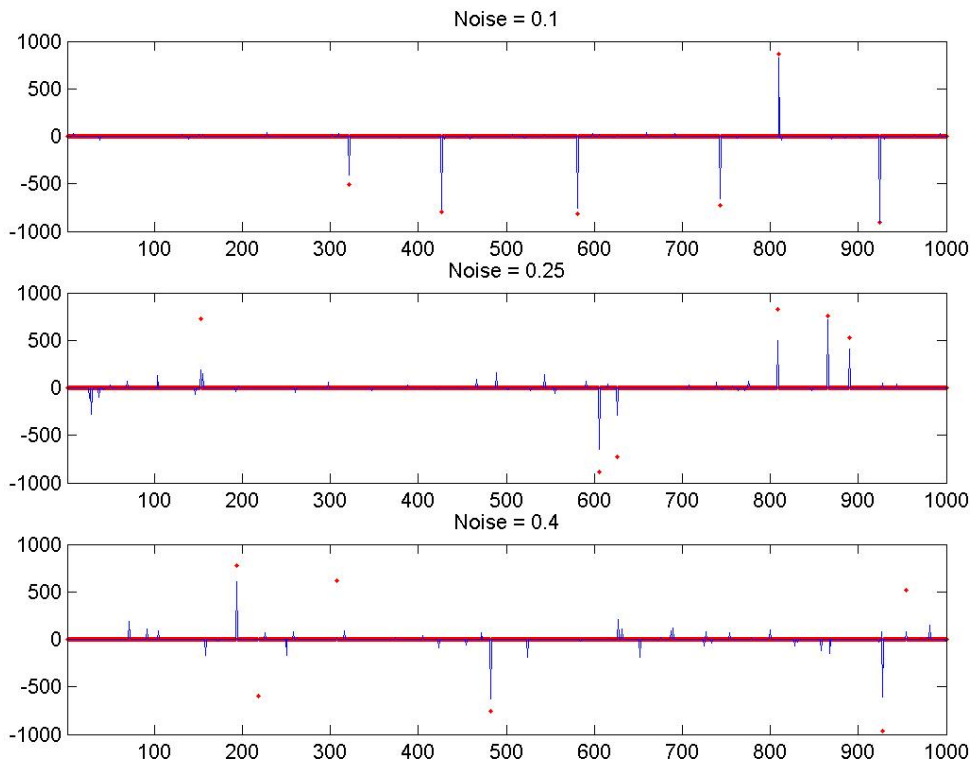


Figure 1: The influence of increasing intrinsic noise on the identifiability. The plot shows the corresponding values of the gene-gene matrix  $a \equiv \text{vec}(A)$ , and increasing zero-mean Gaussian noise added to  $A$ . The red dots indicate the true value of  $a$ , and the blue line the reconstructed values  $a^*$ . For low noise levels, like 0.1, the non-zero values of  $a$  are recovered without exception. At noise level 0.4 only the largest modulus maxima values have a chance to be found.

## 5 Discussion

In this work we have presented an approach for modeling and identification of gene regulatory networks from near genome wide expression profiles with a relative small amount of time instances using a piecewise linear state space model. The state space model is a rich and flexible metaphor from mathematical systems theory that applied to this case allows for hierarchical activation through master genes, representing the effects of multiple external inputs, hidden states such as none-observed genes or protein densities, and the effects of process and measurement noise. For this piecewise linear state space modeling we have presented an identification technique, based on a coupled set of two linear programming problems. This approach resulted in an efficient and fast algorithm that is able to accurately estimate the gene-gene coupling matrix for tens of thousands of genes based on only several hundred genome wide measurements, and that is robust towards measurement noise.

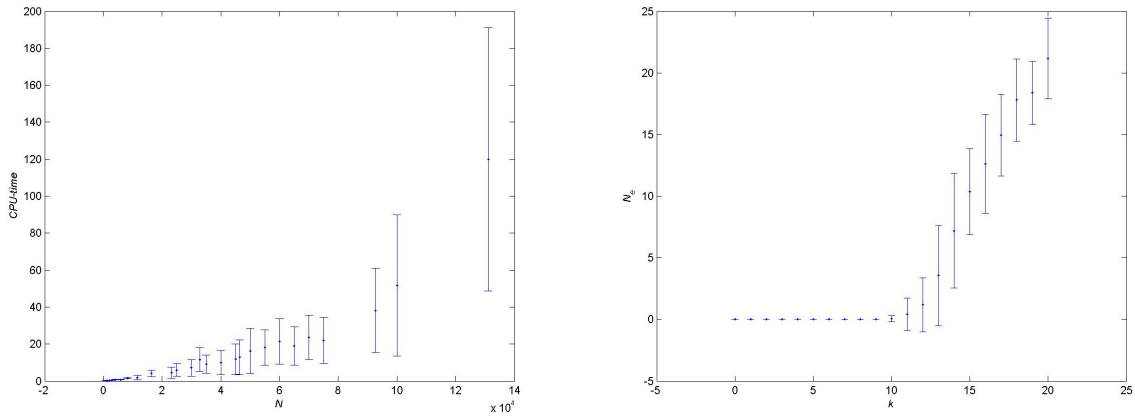


Figure 2: *a*: CPU-time  $T_c$  as a function of the problem size  $N$ , *b*: Number of errors as a function of the number of nonzero entries  $k$  in  $x_0$ , for  $M = 150$ ,  $m = 5$ ,  $N = 50000$ .

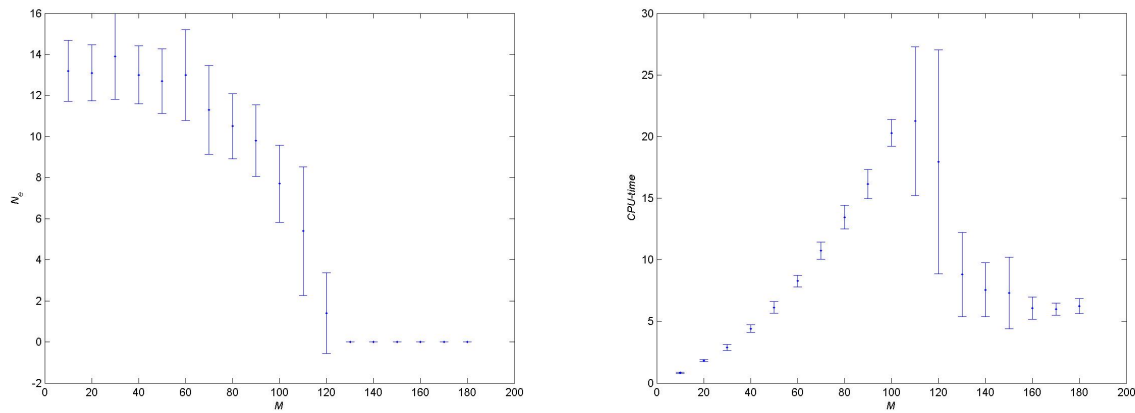


Figure 3: *a*: Number of errors as a function of  $M$  for  $N = 50000$ ,  $k = 10$ ,  $m = 0$ , *b*: Computation time as a function of  $M$ , for  $N = 50000$ ,  $k = 10$ ,  $m = 0$ .

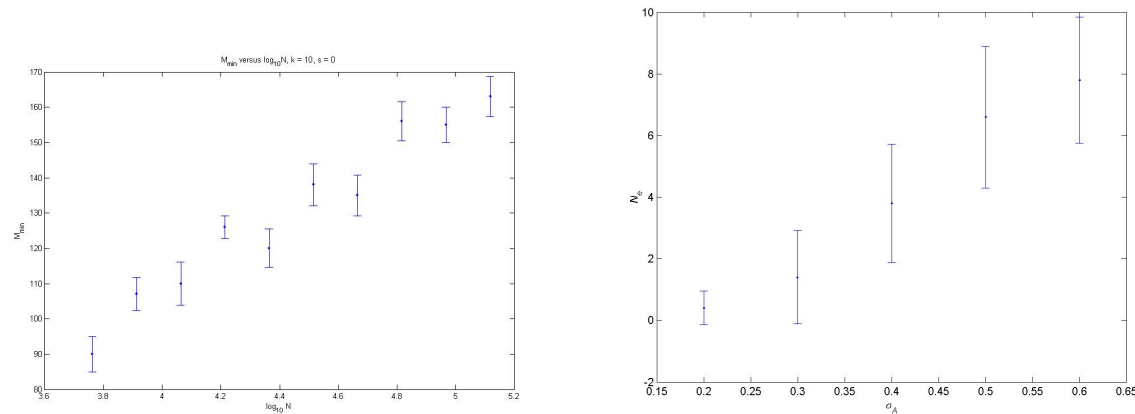


Figure 4: *a*: Dependency of the critical value  $M_{min}$  required to compute the matrix free of error versus the problem size  $N$ , *b*: Number of errors as a function of the intrinsic noise level  $\sigma_A$ , for  $N = 10000$ ,  $k = 10$ ,  $m = 5$ , with  $M = 150$  and measuring noise  $\sigma_B = 0$ .

There remain a number of difficulties with regard to the system identifiability of this approach, i.e. the potential to reconstruct the interaction network from empirical data.

1. Due to the huge costs and efforts involved in the experiments, only a limited number of time points are available in the data. Together with the high dimensionality of the system, this makes the problem severely under-determined.
2. In the time series many genes exhibit strong correlation in their time-evolution, which is not per se indicative for a strong coupling between these genes but rather induced by the over-all dynamics of the ensemble of genes. This can be avoided by persistently exciting inputs.
3. Not all genes are observed in the experiment, and certainly most of the RNAs and proteins are not considered. therefore, there are many *hidden* states.
4. Because the identification techniques proposed below work on the rows, the hierarchical principle does not cause a problem, as the gene-gene interaction matrix is highly row-sparse but not column-sparse. In fact, the method utilizes the sparsity of the matrix as an implicit constraint, namely that the value of the components of the matrix should be zero.
5. Effects of stochastic fluctuations on genes with low transcription factors are severe and will obscure their true dependencies.

With this approach it is possible to reconstruct the steady states and the associated switching times of a metabolic processes from a set of micro-array experiments. In each steady state the gene-gene interaction matrix defines the network topology. The micro-array technique exhibits a strong increase in efficiency and a simultaneous decrease in associated costs. In the near future this will enable the registration of large time series of genome wide expression profiles and associated protein densities. The future availability of such data makes the further development of the mathematical modeling and associated identification of dynamic gene expression, as the approach presented here, an important condition for deducing and understanding the underlying interactions between genes and their environment.

## References

- [1] Bower J.M., Bolouri H.(Editors), Computational Modeling of Genetic and Biochemical Networks, *MIT Press*, 2001. bibitemDavidson1999Davidson E.H. (1999), A View from the Genome: Spatial Control of Transcription in Sea Urchin Development, *Current Opinions in Genetics and Development*, **9**, pp. 530 – 541.
- [2] de Jong H., Modeling and Simulation of Genetic Regulatory Systems: A Literature Review, *Journal of Computational Biology*, 2002, Volume 9, Number 1, pp. 67–103
- [3] de Jong H., Gouze J.L., Hernandez C., Page M., Sari T., Geiselmann J., Qualitative simulation of genetic regulatory networks using piecewise-linear models, *Bull Math Biol.* 2004 Mar;66(2): pp 301–40.
- [4] D’haeseleer P., Liang S., Somogyi R., Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering, *Bioinformatics*, vol. **16**, no. 8, 2000, pp. 707–726.
- [5] Elowitz M.B., Levine A.J., Siggia E.D., Swain P.S., Stochastic gene expression in a single cell, *Science*, vol.**297**, August 16, 2002, pp.1183–1186.

- [6] Endy, D, Brent, R. (2001) Modeling Cellular Behavior, *Nature* 2001 Jan 18; 409(6818):391-5.
- [7] Fuchs J.J. (2003), More on sparse representations in arbitrary bases, in: Proc. 13th IFAC Symp. on System Identification, Sysid 2003, Rotterdam, The Netherlands, August 27-29, 2003, pp. 1357–1362.
- [8] Fuchs J.J. (2004), On sparse representations in arbitrary redundant bases, *IEEE Trans. on IT*, June 2004.
- [9] Glass L., Kauffman S.A. (1973), The Logical Analysis of Continuous Non-linear Biochemical Control Networks, *J.Theor.Biol.*, 1973 Vol. 39(1), pp. 103–129
- [10] Goldbeter A (2002) Computational approaches to cellular rhythms. *Nature* 420, 238-45
- [11] Gonze D, Halloy J, and Goldbeter A (2004) Stochastic models for circadian oscillations : Emergence of a biological rhythm. *Int J Quantum Chem* **98**, pp 228–238.
- [12] Guthke R., Möller U., Hoffmann M., Thies F., Töpfer S., 2004, Dynamic network reconstruction from gene expression data applied to immune response, *Bioinformatics*, 2004, pp 2261
- [13] Hasty J., McMillen D., Isaacs F., Collins J. J., (2001), Computational studies of gene regulatory networks: in numero molecular biology, *Nature Reviews Genetics*, vol. 2, no. 4, pp. 268– 279, 2001.
- [14] Novak B, Tyson JJ (1997) Modeling the control of DNA replication in fission yeast, *PNAS*, USA, Vol. 94, pp. 9147-9152, August 1997.
- [15] Peeters R.L.M., Westra R.L., On the identification of sparse gene regulatory networks, *Proc. of the 16th Intern. Symp. on Mathematical Theory of Networks and Systems (MTNS2004)* Leuven, Belgium July 5-9, 2004
- [16] Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB, Gene regulation at the single-cell level, *Science* 307 (2005) pp 1962.
- [17] Somogyi R., Fuhrman S., Askenazi M., Wuensche A. (1997). The Gene Expression Matrix: Towards the Extraction of Genetic Network Architectures. *Nonlinear Analysis, Proc. of Second World Cong. of Nonlinear Analysis (WCNA96)* 30(3) pp 1815–1824.
- [18] Swain P.S., Efficient attenuation of stochasticity in gene expression through post-transcriptional control, *J Mol Biol* 344 (2004) pp 965.
- [19] Swain P.S., Elowitz MB, Siggia ED, Intrinsic and extrinsic contributions to stochasticity in gene expression, *PNAS* 99 (2002) pp 12795.
- [20] Steuer R. (2004), Effects of stochasticity in models of the cell cycle:from quantized cycle times to noise-induced oscillations, *Journal of Theoretical Biology* 228 (2004) 293-301.
- [21] Tegnér J., Yeung M.K.S., Hasty J., Collins J.J., Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling, *Proc. Nat. Acad. Science*, vol. **100**, no. 10, 2003, pp. 5944–5949.
- [22] van Kampen N. G. (1992), *Stochastic Processes in Physics and Chemistry*, Elsevier ScienceB. V., Amsterdam, (1992).
- [23] van Schuppen J.H. (2004), System theory of rational positive systems for cell reaction networks, *CWI Report MAS-E0421*, December 2004, ISSN 1386-3703

- [24] Verdult V., Verhaegen M., Subspace Identification of Piecewise Linear Systems, In *Proc. 43rd IEEE Conference on Decision and Control (CDC)*, pp 3838–3843, Atlantis, Paradise Island, Bahamas, December 2004.
- [25] Westra R.L., Peeters R.L.M. (2004), Modelling and identification of dynamical gene interactions: presentation, Workshop Intelligent Technologies for Gene Expression Based Individualized Medicine, 14th May 2004, Jena/Germany
- [26] Westra R.L.,(2005a), Piecewise Linear Dynamic Modeling and Identification of Gene-Protein Interaction Networks, Nisis/JCB Workshop reverse engineering, Jena, June 10, 2005.
- [27] Yeung M.K.S., Tegnér J., Collins J.J., Reverse engineering gene networks using singular value decomposition and robust regression, *Proc. Nat. Acad. Science*, vol. **99**, no. 9, 2002, pp. 6163–6168.