The Identification of Dynamic Gene-Protein Networks

WESTRA, Ronald; HOLLANDERS, Goele; BEX, Geert Jan; GYSSENS, Marc &
TUYLS, Karl (2007) The Identification of Dynamic Gene-Protein Networks. In:
Knowledge Discovery and Emergent Complexity in Bioinformatics, First International
Workshop, KDECB 2006, Ghent, Belgium, May 2006, Revised Selected Papers. p. 157-170.

DOI: 10.1007/978-3-540-71037-0_11
Handle: http://hdl.handle.net/1942/7972

The Identification of Dynamic Gene-Protein Networks

Ronald L. Westra[1], Goele Hollanders[2], Geert Jan Bex[2], Marc Gyssens[2], Karl Tuyls[1]

[1] Department of Mathematics and Computer Science,
Maastricht University and Transnational University of Limburg,
Maastricht, The Netherlands
[2] Department of Mathematics, Physics, and Computer Science,
Hasselt University and Transnational University of Limburg,
Hasselt, Belgium
E-mail to: westra@math.unimaas.nl

Abstract. In this study we will focus on piecewise 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 (also known as 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 model, robust identification, hierarchical networks, microarrays, gene regulatory networks.

1 Introduction and problem statement

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. 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* [9, 10], or the cell cycle in *Schizosaccharomyces* (fission yeast) [12]. Others try to provide a general platform for modeling the interactions between genes and proteins. For a thorough overview, consult de Jong in [2], Bower in [1], and others [5, 11]. We will focus on dynamical models, and not discuss static models where the relations between genes are considered fixed in time. A dynamical model can be described using continuous time, or discrete events (or time). Given the discrete nature of the data we have at our disposal to derive the models, a discrete event model seems most appropriate. 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].

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

A frequent approach to modeling the dynamical interactions amongst genes and proteins is to consider them as biochemical reactions, and thus represent them as 'rate equations', i.e. as 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. 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 [4, 16, 18, 19]. A practical problem is that the precise details of most reactions are unknown, and therefore cannot be modeled as rate equations. This could be compensated by a well-defined parametrized generic form of the interactions, in which the parameters can be estimated from sufficient empirical data. A generic form based on rational positive functions is proposed by J. van Schuppen [21]. However, in the few cases where parts of such interaction networks have been described from experimental analysis, like the circadian rhythms in certain amoeba [9], or the cell cycle in fission yeast [12], it is clear that such forms have a too extensive syntax to be of any practical use.

Let us for now ignore these problems, and consider the dynamics of gene-RNA-protein networks. When we assume a stochastic differential equation as a 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 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; and $\xi(t)$ denotes a stochastic Gaussian white noise term. This expression involves a parameter vector θ 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 N-dimensional space of all possible configurations. In the formalism of dynamic systems theory, x will eventually enter an area of attraction, and become subject to the influence of an attractor. An attractor here can be a uniform convergent attractor, a limit cycle, or a 'strange attractor'. We can understand the entire space as being partitioned into cells, with each cell having an attractor or a repeller. 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 its specific (un)stable equilibrium point. In that case, it is possible to approximate the dynamics of Equation (1) in cell ℓ—for x near the ℓ-th equilibrium $x_{\mathrm{eq}}^{(\ell)}$ and small u—(except the noise term) as:

$$\dot{x}(t) \approx \frac{\partial f(x_{eq}^{(\ell)}, u)}{\partial x}(x - x_{eq}^{(\ell)}) + \frac{\partial f(x_{eq}^{(\ell)}, u)}{\partial u} u \equiv A_\ell x(t) + B_\ell u(t) + c_\ell \qquad (2)$$

Thus, the qualitative behavioral dynamics of gene-protein interactions is characterized 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 'checkpoints'. Within each such checkpoint, the system is in relative equilibrium, see [20]. We will follow the view of *piecewise linear behavior* (PWL). This approach corresponds to the piecewise linear models introduced by Glass and Kauffman [8], and the qualitative piecewise linear models described by de Jong et al. [2, 3].

3 Identification of dynamic networks using *piecewise linear* models

Next, we will be concerned with the identification (also known as *reverse engineering*) of piecewise linear gene regulatory systems from microarray data. We consider the case where time series of genome-wide expression data are available. The nature of our problem—few microarray experiments and lots of genes—implies that we are dealing with *poor 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. A different perspective is offered by including some characteristics of the biological problem, such as the hierarchy and sparsity of the networks. The case of the identification of a *simple* linear system with sparse and hierarchic interactions is discussed by Peeters and Westra [14, 23], and Yeung et al. [24]. In realistic situations, this model is too simple however. As was pointed out by Øyehaug et al. [13], such systems tend to behave in a switch-like manner, and they determine the switching timepoints using complex biological modeling. In contrast, we will determine the switching timepoints by identifying sparse *piecewise* linear systems. As a consequence, our focus is on modeling the subsystems between the switching points rather than on the dynamics of the switching points themselves, as, e.g., in Plahte et al. [15]. More concretely, our main aim is to obtain the local gene-gene interaction matrices A_ℓ, that directly relate to the graph of the gene regulatory network. Additionally, the matrices B_ℓ provide information on the coupling of genes to specific inputs.

3.1 General dynamics of switching subsystems

In what follows, let us assume a dynamical input-output system Σ that switches irregularly between K linear time-invariant subsystems $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_K\}$.

Let $S = \{s_1, s_2, \ldots, s_{K-1}\}$ denote the set of—unknown—switching times, i.e., the time instants $t = s_\ell$ when the system switches from subsystem Σ_ℓ, to $\Sigma_{\ell+1}$. Similarly as with the simple linear networks, we assume empirical data $X = (x[1], \ldots, x[M])$, $U = (u[1], \ldots, u[M])$, and $\dot{X} = (\dot{x}[1], \ldots, \dot{x}[M])$ at M sampling times $T = \{t_1, t_2, \ldots, t_M\}$, representing full observations of the N states and P inputs, and $x[k] \equiv x(t_k)$. The interval between two sample instants is denoted as $\tau_k = t_{k+1} - t_k$. Here, we assume that the system is sampled on regular time intervals, i.e., that the sample intervals are equal to τ. Within one subsystem $\varSigma_\ell$, the effect of the inputs $u(t)$ is represented as a state-space system of first-order differential (for continuous time systems) or difference equations (for discrete time systems), using an internal vector $x(t)$ spanning the so-called subspace. In our case, this represents the observed gene expressions. In the case of continuous time and in the absence of noise, this system can be written as:

$$\dot{x}(t) = A_\ell x(t) + \tilde{B}_\ell \tilde{u}(t), \tag{3}$$

with $\tilde{B}_\ell = (B_\ell | - A_\ell e_\ell)$, $\tilde{u}^T = (u^T,\ 1)$, where e_ℓ indicates the equilibrium point of the ℓ-th subsystem and A_ℓ and B_ℓ refer to Equation (2). We will use this linear expression, and from here on drop the *tilde*. A general disadvantage is that the time evolution of the different genes, i.e., $x_\nu(t)$, $\nu = 1, \ldots, n$, will strongly correlate, thus obscuring their true relation. This can be avoided by using Equation (3) with time series of triplets $\xi[k] \equiv (x[k], u[k], \dot{x}[k])$ with a sufficient amount of statistically independent and varying inputs $u(t_k)$. Practically, this opens the way to combining distinct empirical sets. However, a practical disadvantage of Equation 3 is that the derivative $\dot{x}(t_k)$ can only be approximated from the measurements, such as $\dot{x}[k] \approx (x[k] - x[k-1])/(t_k - t_{k-1})$.

We furthermore assume that the system matrices in these equations are constant during intervals $[s_\ell, s_{\ell+1}[$, and abruptly change at the transition between the intervals at $t = s_{\ell+1}$. We assume that on the time scale τ, 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, e.g., see [22]. This is accomplished by defining *weights* $w_{k,\ell}$ as the degree to which observation k belongs to subsystem $\varSigma_\ell$. If observation $\xi[k]$ belongs to system $\varSigma_\ell$ then $w_{k\ell} = 1$. Non-integer values in [0,1] can be interpreted as the fuzzy membership of observation k to system $\varSigma_\ell$. Since we assume that the subsystems $\{\varSigma_1, \varSigma_2, .., \varSigma_K\}$ act disjointly and subsequently, the result can be improved by matching the weights to a block function structure; i.e., $w_{kl} = 1$ 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. These considerations lead to the constraints $\mathcal{C}_{MK}$ on w:

$$\mathcal{C}_{MK}(w) : \begin{cases} w_{1,1} = 1, w_{M,K} = 1, \\ \forall_{k,\ell} w_{k,\ell} \in [0,1], \\ \forall_k \sum_l w_{k,\ell} = 1, \\ \forall_\ell \sum_{k=1}^{M-1} |w_{k+1,1} - w_{k,1}| = 1, \\ \forall_\ell \sum_{k=1}^{M-1} |w_{k+1,\ell} - w_{k,\ell}| = 2, \\ \forall_\ell \sum_{k=1}^{M-1} |w_{k+1,K} - w_{k,K}| = 1. \end{cases} \tag{4}$$

3.2 Combining the system matrices $\{A, B\}$ with the subsystem weightmatrix W

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 metabolical processes.

When a sufficiently accurate record of estimates of the state derivatives $\dot{X}$ 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 [14]. 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

$$\dot{X} = H_1 X + H_2 U. \tag{5}$$

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

$$\mathtt{vec}(H_1) = W \cdot \mathtt{vec}(A), \tag{6}$$

$$\mathtt{vec}(H_2) = W \cdot \mathtt{vec}(B). \tag{7}$$

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

$$A = \begin{pmatrix} A_1 \\ \vdots \\ A_K \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ \vdots \\ B_K \end{pmatrix}, \quad W = w \otimes I_{N^2} = \begin{pmatrix} w_{1,1} I_{N^2} & \cdots & w_{1,K} I_{N^2} \\ \vdots & \vdots & \vdots \\ w_{M,1} I_{N^2} & \cdots & w_{M,K} I_{N^2} \end{pmatrix}, \tag{8}$$

where $\otimes$ is the Kronecker-product, and I_{N^2} is the $N^2 \times N^2$ identity matrix. Note that Equation (5) 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 [14]. 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, the problem has now shifted to solving two additional non-linear equations:

$$W \diamond A = H_1, \tag{9}$$

$$W \diamond B = H_2. \tag{10}$$

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 (6) and (7) explicit. This is an underdetermined set of equations that can only be solved by additional information, such as assuming sparsity for A, and a block structure for W, as defined in Equation (4).

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

We will now focus on the general case, that the genome wide expressions X, their derivatives $\dot{X}$, and the external inputs U are available as empirical data $\mathcal{D}$. In this case, the objective of system identification is to compute concurrently the system parameters A, B, *and* weights W (and hence the switching times S). Equation (5) provides us with the general state space equations for a PWL system.

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

$$\mathcal{E}_{sys}(A, B, w|\mathcal{D}) = \sum_{k,l} w_{kl}||A_l x[k] + B_l u[k] - \dot{x}[k]||_p \tag{11}$$

Here, (A, B) represent the set of system parameters, and $\mathcal{D} \equiv \{X, U, \dot{X}\}$ the observed data, i.e., the measured genome-wide expressions X, their fluxes $\dot{X}$, and the external inputs U. The criterion furthermore involves the relation between the k-th observation and the ℓ-th subsystem Σ_ℓ; namely the *weight* $w_{k\ell}$ and the *distance* $d_{k\ell}$ between observed and the expected value of observation k relative to subsystem model Σ_ℓ.

In order to handle the underdetermined character of the problem, we furthermore employ the sparsity and the hierarchy of the underlying biology. This means that the matrices A_ℓ and B_ℓ are *row-sparse,* but not necessarily collum-sparse, as some genes—called the master-genes or source-genes—control a large part of the entire genome. Under a wide range of conditions, this problem is equal to minimization of the L_1-norm of the rows of A_ℓ and B_ℓ as argued by J. J. Fuchs [7]. This implies a global minimization such as

$$\mathcal{E}_{sparse}(A|\mathcal{D}) = \sum_{\ell} ||\mathtt{vec}(A_\ell^T)||_1 \equiv ||A||_1 \tag{12}$$

under the constraints that $\{X, U, \dot{X}\}$ satisfy Equation (5).

The problem of estimating the system parameters can thus formally be defined as the search for the vectors A^*, B^* and w^* that globally minimize $\mathcal{E}$. This can be formulated as a quadratic programming problem, as follows:

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

$$\begin{aligned} &(A^*, B^*, w^*) = \arg\min\nolimits_{(A,B)\in\mathbb{R}^{N(P+N)}, w\in\mathbb{R}^{KM}} \mathcal{E}(A, B, w|\mathcal{D}) \\ &\text{subject to:} \\ &\quad \mathcal{E}(A, B|\mathcal{D}) = \lambda_1\mathcal{E}_{sys}(A, B, w|\mathcal{D}) + \lambda_2\mathcal{E}_{sparse}(A|\mathcal{D}) + \lambda_3\mathcal{E}_{sparse}(B|\mathcal{D}), \\ &\quad C_{MK}(w). \end{aligned} \tag{13}$$

for selected λ's with: $\lambda_1 + \lambda_2 + \lambda_3 = 1$, and the constraints $C_{MK}(w)$ in Equation (4). This is a regularized (or scalarized) convex quadratic optimization problem that is 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 study the following two coupled linear programming problems associated to the original QP:

LP1: given the weight matrix $\tilde{w}$ compute the system matrices (A^*, B^*):

$$\begin{aligned} &(A^*, B^*) = \arg\min\nolimits_{(A,B)\in\mathbb{R}^{N(P+N)}} \mathcal{E}(A, B, \tilde{w}|\mathcal{D}) \\ &\text{subject to:} \\ &\quad \mathcal{E}(A, B|\mathcal{D}) = \lambda_1\mathcal{E}_{sys}(A, B, w|\mathcal{D}) + \lambda_2\mathcal{E}_{sparse}(A|\mathcal{D}) + \lambda_3\mathcal{E}_{sparse}(B|\mathcal{D}) \end{aligned} \tag{14}$$

LP2: given system matrices $(\tilde{A}_\ell, \tilde{B}_\ell)$ apply the L_1-norm $\tilde{d}_{kl} = \|\dot{x}[k] - \tilde{A}_\ell x[k] - \tilde{B}_\ell u[k]\|_1$ to compute the weight matrix w^*:

$$w^* = \arg\min_{w \in \mathbb{R}^{KM}} \mathcal{E}_{sys}(\tilde{A}, \tilde{B}, w|\mathcal{D}) = \textstyle\sum_{k=1}^{M-1} \sum_{l=1}^{K} \tilde{d}_{kl} w_{kl} \tag{15}$$

subject to:

$$C_{MK}(w).$$

LP1 is a regularized optimization. J. J. Fuchs [6, 7] has described conditions under which the regularization drives the optimization problem towards the global solution. Though these conditions do not strictly apply here, we find that this approach succeeds in numerical simulations. Both LP-problems can be solved efficiently with a partial dual simplex method as in [14], or by using large-scale or interior-points methods. The algorithm to estimate the system parameters $\{A, B\}$ and w consists of iteratively solving the two optimizations LP1 and LP2 subsequently, until the criterion has sufficiently converged. Though the solution of the original quadratic programming problem QP in Equation (13) is also the global solution of the two coupled LP-problems LP1 and LP2, there can also exist local solutions to the couple {LP1,LP2}, unfortunately.

3.4 Construction and control of the subsystem weightmatrix

For small values of the regularization terms in $\mathcal{E}$ in LP1 (Equation (14)), i.e., $\lambda_2, \lambda_3 \ll \lambda_1$, and a simultaneous, extreme under-determined system, i.e., $\#\Sigma_\ell \ll N$, the tandem {LP1,LP2} proposed above, runs into problems. The problem amounts to the degree of freedom that formulation LP1 offers to match empirical data $\mathcal{D}$ with system $\Sigma = (A, B)$ in order to minimize the distance to the model space $d(\mathcal{D}, \Sigma)$. It is well-known that at least $M_{\min} \propto log(N)$ measurements are required for a good reconstruction of sparse matrices A and B, see for instance [6, 7, 24]. Therefore, when $\#\Sigma_\ell \ll M_{\min}$, the heavily under-determined system has a high degree of freedom to match the data with the model. This will cause the tandem {LP1,LP2} to halt as the criterion $d(\mathcal{D}, \Sigma) \approx 0$ has been reached.

Avoiding this problem requires (i) the restriction of the maximum number of subsystems to $K < M/log(N)$, and (ii) the careful control of the weight matrix w during the iteration, such that each subsystem Σ_ℓ has at least $M_{\min}$ elements, i.e., $\#\Sigma_\ell \geq M_{\min}$. For this reason, the following iteration is performed for initializing the weight matrix:

1. Assign the *current measurement* k to 1, and the *current system* ℓ to 1. Initianlize w to the $M \times K$ null matrix: $w = 0$.
2. The first $M_{\min}$ measurements are assigned to the current—i.e., first—subsystem: $w(11 = 1, \ldots, w_{M_{\min},1} = 1$. Now the current measurement k is set to $M_{\min} + 1$.
3. The current measurement, $\xi_k = (x[k], u[k], \dot{x}[k])$, belongs to the current subsystem Σ_ℓ if $d(\xi_k, \Sigma_\jmath)$ is minimized by $\jmath = \ell$. In that case: (i) it is assigned to the current system by setting $w_{k\ell} = 1$, and (ii) the next measurement is considered, i.e., k is increased, and step 3 is repeated.
4. If another system $\Sigma_\jmath$ is closer to ξ_k, then this system is assigned to the current system: $\ell = \jmath$, and measurement k is considered as the first of $M_{\min}$ measurements assigned directly to this subsystem, i.e., $w_{k\ell} = 1, \ldots, w_{k+M_{\min}-1,\ell} = 1$, k is set to $k + M_{\min}$, and step 3 is repeated.

This iteration process is continued as long as there are unassigned measurements. When the final subsystem has less then $M_{\min}$ elements, these are discarded. Finally, all measurements will belong to some subsystem, while w obeys all constraints defined in Equation (4). One of the advantages of this *matching* algorithm is that it requires no advance knowledge of the number of subsystems.

3.5 A tandem for network reconstruction using the subsystem weight matrix

The procedure for constructing and managing the subsystem weight matrix w, defined in Section 3.4, allows for an efficient tandem approach to solving the identification problem.

The non-linear problem $\dot{X} = H_1X + H_2U$, defined in Equation (5), can 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 not well-posed quadratic problem. For these reasons, we again split the problem and follow a tandem approach as discussed in Section 3.2. However, in the present tandem the construction of the subsystem weight matrix w is performed by the matching approach defined above, rather than by the LP2 defined in Equation (15). Both amount to a solution obeying the weight constraints in Equation (4), but the matching algorithm will prevent too underdetermined systems that will prematurely halt the iteration as they generate a fictitious match with the model. The computation of the system matrices (A, B) is again performed by the robust L_1 identification in LP1, with $\lambda_1 = 0$, and $\lambda_2 = \lambda_3$. The tandem is controlled by the distance between the data and the model: $d(\mathcal{D}, \Sigma)) = \mathcal{E}_{sys}(A, B, W|\mathcal{D})$, defined in Equation (11). If this quantity has converged below a pre-specified threshold, the iteration is terminated.

4 Numerical experiments and performance of the approach.

The approach described in the previous section resulted in an efficient and fast algorithm that is able to estimate accurately the gene-gene coupling matrix based on several genome-wide measurements, and that is robust towards measurement noise.

All experiments were performed on a PC with an Intel Pentium M processor of 1.73 GHz and 1 GB RAM memory under Windows XP Professional, using Matlab 6.5 Release 13 including the Optimization Toolbox. The latter's routine `linprog` was used to solve LP problems; its default solution method is a primal-dual interior point method, but an active set method can optionally be used, too. For larger problems, it turned out to be essential for obtaining reasonable computation times that the LP problems were solved by application of the active set method on the dual problem formulation. Therefore, this method was adopted throughout all the experiments.

Since results can depend on the particularities of given data and the original system that generated it, all experiments have been performed on a number of independent runs on randomly selected data and systems. Hence they convey the behavior of our approach "on average". The number of independent runs is 50 for each of the experiments described below.

In line with the definitions above, we use the parameters N, M, K to quantify the size and complexity of the input. In addition, the sparsity of the local interaction matrix A is measured by the number of non-zero entries per row and denoted by k (which should be much smaller than N). To complete the system's data set, some stochastic Gaussian white noise is added to the input data set. It is normally distributed with zero mean and some standard deviation σ that determines the noise level. To quantify the quality of the resulting approximation A_{est} of A^*, a performance measure is introduced: the number of errors N_e.

These errors are generated in the reconstruction by the failure of the algorithm to identify the true non-zero elements of the original sparse matrix A^*. These errors stem from false positives and false negatives in the reconstructed matrix A_{est}. Their numbers are added up to produce the total number of errors N_e.

The success of the algorithm depends on different factors. First, for a certain number of genes, a sufficient number of measurements has to be available. Therefore, the minimal number of measurements required for a certain number of genes, denoted by M_{min}, has been determined. This is the number of measurements so that the total system error, N_e, is acceptably small. Figure 1 represents the values for M_{min} as a function of the number of genes.

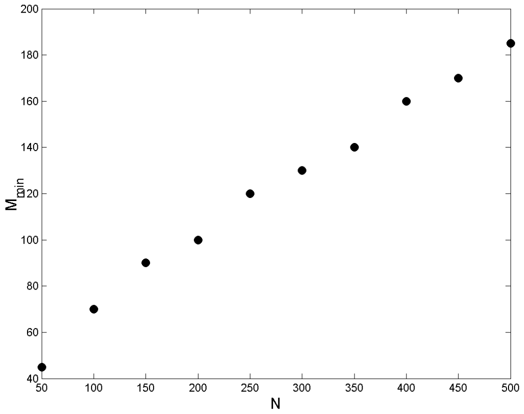

Fig. 1. Minimal number of required measurements M_{min} as a function of the number of genes N.

For comparability reasons, the number of genes in all the following experiments has been fixed to $N = 150$. Consequently, the associated minimal number of measurements has been fixed to $M_{\text{min}} = 90$ (see Figure 1).

Second, the number of errors N_e depends on the noise level σ. Figure 2 shows how this noise level influences the error rate in our approach. As to be expected, the error increases if the noise level increases, and vice versa.

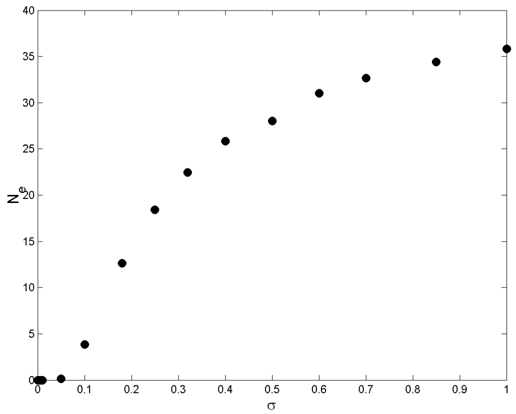

Fig. 2. Number of errors N_e as a function of the noise level σ, with $N = 150$, $M = M_{\min}$ and $k = 1$.

The numerical experiments consist of the comparison of the reconstructed network with the—known—original network structure, and they clearly reveal the range where the approach is effective.

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, k. If k rises above a certain threshold, the performance of the approach is abruptly and severely affected (see Figure 3).

For relatively moderate noise levels and a high degree of sparsity—i.e., a small number k of non-zero elements in the rows of matrix A^*—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^* with rows of 150 components of which all but 3 are equal to zero, can be efficiently reconstructed from just 90 independent measurements (Figure 4).

Figure 4 shows an initial increase, followed by a decrease. Finally, N_e jumps abruptly to zero above a certain threshold value for M. To explain this phenomenon, remember that the number of errors N_e is the sum of the false positives and the false negatives in the gene interaction matrix. The false positives correspond to the non-zero values in the matrix A_{est} that should be zero, and vice-versa for the false negatives. Turning back to Figure 4, the initial increase is caused by false positives. Indeed, as long as $M < M'_{\min}$, where $M'_{\min}$ is the minimal number of required measurements *in the case of a single row*, $k \approx M'_{\min}$. As soon as M reaches $M'_{\min}$, the system becomes completely determined, whence k drops to its proper value. Observe that $M'_{\min} < M_{\min}$ due to the absence of effects related to the composition of rows. Notice that the false negatives decrease monotonously over the entire range of M.

Finally, some experiments concerning multiple subsystems were performed. Figure 5 shows the accuracy of the partioning of the available measurements into different

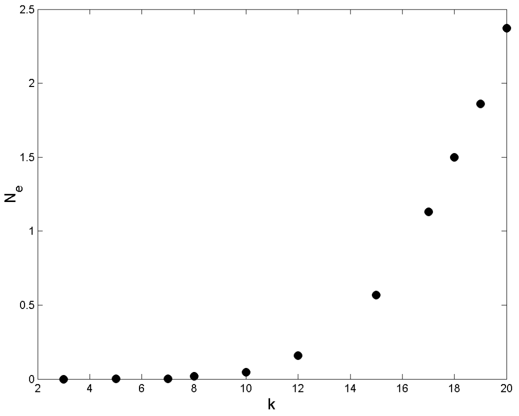

Fig. 3. Number of errors N_e as a function of the number of non-zero elements per row k for a single subsystem ($K = 1$), with $N = 150$ and $M = M_{\min}$.

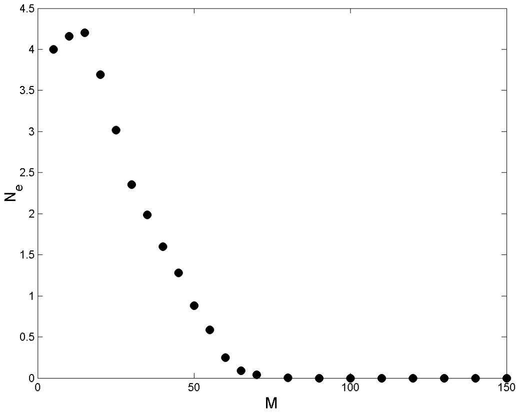

Fig. 4. Number of errors N_e as a function of the number of measurements M, with $N = 150$ and $k = 1$.

subsystems. The error measure δ shown in Figure 5 is defined as the cumulative distance in terms of time stamps between erroneously classified measurements and the switching point of the class they belong to, relative to the total number of measurements. In the experiment illustrated by Figure 5, two subsystems were identified.

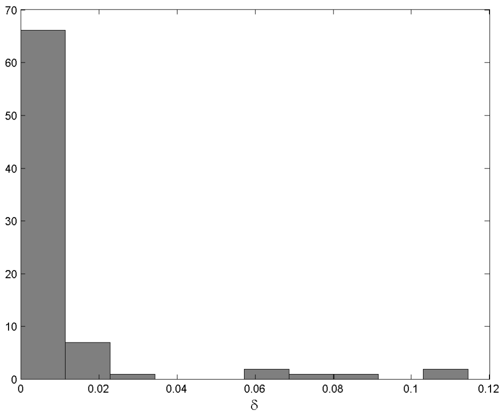

Fig. 5. The distribution of the error measure δ for partioning $M = 200$ measurements into subsystems, with $N = 150$. Two subsystems were identified.

5 Discussion

In this work, we have presented an approach for modeling and identifying 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 linear programming problem. This approach resulted in an efficient and fast algorithm that is able to accurately estimate the gene-gene coupling matrix for a large number of genes based on only several hundred genome-wide measurements, and that is robust towards measurement noise. Figure 6 shows the CPU time used by the algorithm as a function of the number of genes N.

In future work, a few difficulties with regard to the system identifiability of this approach, i.e., the potential to reconstruct the interaction network from empirical data, will have to be addressed.

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.

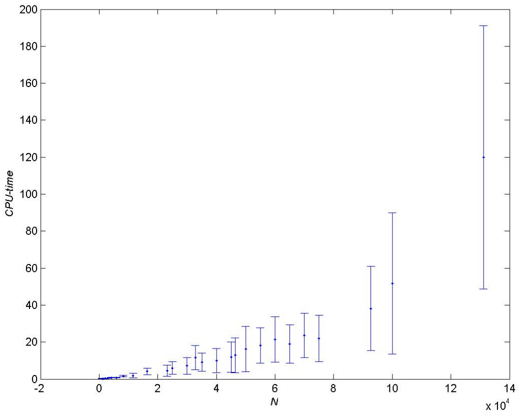

Fig. 6. The CPU time used by the reconstruction algorithm (in seconds) as a function of the number of genes N.

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. Effects of stochastic fluctuations on genes with low transcription factors are severe and will obscure their true dependencies.
5. Because the identification techniques 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.

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, such 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.

2. de Jong H., Modeling and Simulation of Genetic RegulatorySystems: 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 usingpiecewise-linear models, Bull Math Biol. 2004 Mar;66(2): pp 301–40.
4. 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.
5. Endy, D, Brent, R. (2001) Modeling Cellular Behavior, *Nature* 2001 Jan 18; 409(6818):391-5.
6. 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.
7. Fuchs J.J. (2004), On sparse representations in arbitrary redundant bases, IEEE Trans. on IT, June 2004.
8. 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
9. Goldbeter A (2002) Computational approaches to cellular rhythms. Nature 420, 238-45
10. 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.
11. 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.
12. Novak B, Tyson JJ (1997) Modeling the control of DNA replication in fission yeast, PNAS, USA, Vol. 94, pp. 9147-9152, August 1997.
13. Leiv Øyehaug, Erik Plahte, Stig W. Omholt, Targeted reduction of complex models with time scale hierarchy–a case study, *Mathematical Biosciences*, 185, 123-152, 2003.
14. 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
15. Plahte E, Mestl T, Omholt SW, A methodological basis for description and analysis of systems with complex switch-like interactions, *Journal of Mathematical Biology*, 36, 321-348, 1998.
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. van Schuppen J.H. (2004), System theory of rational positive systems for cell reaction networks, CWI Report MAS-E0421, December 2004, ISSN 1386-3703
22. 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.
23. Westra R.L.,(2005*a*), Piecewise Linear Dynamic Modeling and Identification of Gene-Protein Interaction Networks, Nisis/JCB Workshop reverse engineering, Jena, June 10, 2005.

24. 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.