BMC Systems Biology

Research article

Open Access

Efficient, sparse biological network determination

Elias August* and Antonis Papachristodoulou

Address: Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK

Email: Elias August* - elias_august@hotmail.com; Antonis Papachristodoulou - antonis@eng.ox.ac.uk

* Corresponding author

Published: 23 February 2009

BMC Systems Biology 2009, **3**:25 doi:10.1186/1752-0509-3-25

Received: 29 July 2008
Accepted: 23 February 2009

This article is available from: http://www.biomedcentral.com/1752-0509/3/25

Abstract

Background: Determining the interaction topology of biological systems is a topic that currently attracts significant research interest. Typical models for such systems take the form of differential equations that involve polynomial and rational functions. Such nonlinear models make the problem of determining the connectivity of biochemical networks from time-series experimental data much harder. The use of linear dynamics and linearization techniques that have been proposed in the past can circumvent this, but the general problem of developing efficient algorithms for models that provide more accurate system descriptions remains open.

Results: We present a network determination algorithm that can treat model descriptions with polynomial and rational functions and which does not make use of linearization. For this purpose, we make use of the observation that biochemical networks are in general 'sparse' and minimize the 1-norm of the decision variables (sum of weighted network connections) while constraints keep the error between data and the network dynamics small. The emphasis of our methodology is on determining the interconnection topology rather than the specific reaction constants and it takes into account the necessary properties that a chemical reaction network should have – something that techniques based on linearization can not. The problem can be formulated as a Linear Program, a convex optimization problem, for which efficient algorithms are available that can treat large data sets efficiently and uncertainties in data or model parameters.

Conclusion: The presented methodology is able to predict with accuracy and efficiency the connectivity structure of a chemical reaction network with mass action kinetics and of a gene regulatory network from simulation data even if the dynamics of these systems are non-polynomial (rational) and uncertainties in the data are taken into account. It also produces a network structure that can explain the real experimental data of *L. lactis* and is similar to the one found in the literature. Numerical methods based on Linear Programming can therefore help determine efficiently the network structure of biological systems from large data sets. The overall objective of this work is to provide methods to increase our understanding of complex biochemical systems, particularly through their interconnection and their non-equilibrium behavior.

Background

Determining the interaction topology in large-scale biological systems has been an active area of research for some time now. Most methodologies that deal with high-throughput experimental data make use of information about the behavior of the system locally around the steady-state [1-3]. For example, a class of techniques that fall under the rubric of 'stationary state Jacobian Matrix Elements' analyzes the system behavior when it is perturbed locally from steady-state and looks at whether the concentration of one species is increased or decreased when another species' concentration is increased. In [4] and [5], the authors have built on this approach and determined the functional interactions in cellular signaling and gene networks by taking into account the 'modular' structure of the networks in question. Alternatively, inferences about the topology of the network can be made by introducing pulse changes in concentration of a chemical species in the network, and observing the network's response, concluding causal chemical connectivities [6]. In [3], a linear dynamical system was considered to represent a gene regulatory network, and an approach, based on Linear Programming, was proposed in order to obtain the sparsest network structure from genetic perturbation experiments.

However, most information of the system dynamics is in its transient behavior and, more importantly, many reactions that the living cell requires are actually for most of the time far from steady state [7]. The problem of determining the network structure in the case where transient time-series data for non-equilibrium behavior are available is much harder and this has been an active area of research for over a decade. The reason is that while such data are often abundant due to the development of high-throughput, effective experimental techniques, at the same time, a high computational effort is required to extract information about the network structure using traditional techniques. A recent review of available techniques can be found in [8] or [9], but earlier articles, such as [10], also list several approaches to this network identification problem.

In actual fact, identifying the interconnection topology in biological and biochemical systems is of greater importance than extracting the parameters (the rates of the various reactions) that best fit the particular time series data. There are several reasons for this: first, the parameters are often collected under noisy experimental conditions and are sensitive to the laboratory environment. Second, as is often the case with large networks, persistence of observed phenomena is robust to a large range of most parameter values and therefore identifying these parameters exactly is not of great interest. Indeed, *chemical reaction networks* often have the same functionality in the neighborhood of most of the nominal reaction rates. But most importantly, networks are rarely robust to the random rewiring of the underlying interconnection structure and hence determining the network connectivity is much more important than determining the kinetic parameters that fit the particular data.

In this paper, we first consider chemical reaction networks with *mass action kinetics* (see references [11] and [12]) and seek to identify the chemical pathways and mechanisms, that is, how the chemical complexes interact within the chemical network. Chemical reaction networks are dynamical systems of the form

$$\dot{x} = Af(x), x \in \mathbb{R}^n, A \in \mathbb{R}^{n \times m}, \quad (1)$$

where the unknown matrix encompassing the connectivity structure is A and the vector of functions f: $\mathbb{R}^n \to \mathbb{R}^m$ (which need to satisfy appropriate smoothness conditions to ensure local existence and uniqueness of solutions) is known. This makes (1) linear in the unknown parameters. Our main objective is to provide a procedure to identify the interconnection topology that is encapsulated in A, given experimental time-series data.

An important property of the network given by A is sparseness, i.e., it has much less edges than the full graph with the same vertex set. In this paper, we extend the results in [13] that focus on obtaining sparse interconnection networks from experimental data to general and large-scale networks. We apply the presented methodology in order to reconstruct a biochemical network from mock-up experimental data obtained through simulations. More importantly, we show its validity in determining the glycolytic pathway of *Lactococcus lactis* from real experimental data. Although this pathway has been investigated in great detail (see for example, [14-16]) and is the test object of many system identification approaches as a recent paper fittingly notes in its title, it is 'an unfinished systems biological case study' [14].

Finally, we suggest how the method of identifying connectivity for systems of the form (1) can be adjusted to determine the structure of *gene regulatory networks*, in which the unknown parameters do not enter the system dynamics in an affine way. We then apply the methodology in order to reconstruct a gene regulatory network from mock-up experimental data obtained through simulations.

Notation

$\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{m \times n}$ is the set of all real numbers, real vectors of length n, $m \times n$ real matrices

A_{ij} (i, j)th is the (i, j) entry of matrix $A \in \mathbb{R}^{m \times n}$

$\mathbb{R}^n_+, \bar{\mathbb{R}}^n_+$:$\{x \in \mathbb{R}^n: x_i > 0, i = 1, ..., n\}$, $\{x \in \mathbb{R}^n: x_i \geq 0, i = 1, ..., n\}$

vec(A) is a vector which contains the columns of A stacked one below each other

$e = [1, 1, \cdots, 1]^{\mathrm{T}}$

$\mathrm{diag}(A)$, $A \in \mathbb{R}^{n \times n}$ is a vector of length n, where $(\mathrm{diag}(A))_i = A_{ii}$

$\mathrm{diag}(x)$, $x \in \mathbb{R}^n$ is a matrix in $\mathbb{R}^{n \times n}$, where $(\mathrm{diag}(x))_{ii} = x_i$ and $(\mathrm{diag}(x))_{ij} = 0$ if $j \neq i$

Methods

Chemical reaction networks

Chemical reaction networks are used to describe and understand biological processes. An illustrative example is the following reaction network proposed by Michaelis and Menten for chemical reactions involving enzymes,

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} E + P. \tag{2}$$

Here, S denotes the substrate, E the enzyme, ES the enzyme-substrate complex and P the final product. They are the so called *species* that participate in the reactions. The positive constants k_1, k_{-1}, k_2 and k_{-2} are the reaction *rate constants*, *edges* represent reactions and *vertices* represent *complexes* (the objects that appear before and after the reaction arrows).

In *chemical kinetics*, it is common to assume that the dynamics of the chemical reaction network (such as the one given by (2)) can be described by the following set of nonlinear ODEs [17]:

$$\frac{dx}{dt} \triangleq \dot{x} = Nv(x), \tag{3}$$

where $v(x)$ is the *rate vector* (or *flux vector*), x is the concentration vector of the different species and N is the *stoichiometric matrix*. If p molecules of species i appear before the reaction arrow in reaction j then $N_{ij} = -p$ and if they appear after then $N_{ij} = p$.

The description given by (3) hides the underlying chemical network structure, which we aim to determine in this paper. Hence, in the following, we introduce the notation used in *chemical reaction network theory*, which decomposes N and $v(x)$ into: the so called *bookkeeping matrix Y*, which maps the space of complexes into the space of species; the concentration vector of the different complexes $\Psi(x)$; and matrix A_κ, which defines the network structure. For the Michaelis-Menten reaction, the vectors of species and complexes are given by

$$x = \begin{bmatrix} [E] \\ [S] \\ [ES] \\ [P] \end{bmatrix} \text{ and } \Psi(x) = \begin{bmatrix} [E][S] \\ [ES] \\ [E][P] \end{bmatrix},$$

respectively. The elements of the ith row of matrix Y tell us in which complexes species i appears and how often; or, equivalently, the entries to the jth column tell us of how much of each species complex j is made of. For (2),

$$Y = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Matrix K is the transpose of the *weighted adjacency matrix* of the *digraph* representing the chemical reaction network; that is, entry K_{ij} is nonnegative and corresponds to the rate constant associated with the reaction from complex j to i. The so called *kinetic matrix* A_κ is given by $A_\kappa = K - \mathrm{diag}(K^{\mathrm{T}}e)$. For (2),

$$K = \begin{bmatrix} 0 & k_{-1} & 0 \\ k_1 & 0 & k_{-2} \\ 0 & k_2 & 0 \end{bmatrix} \text{ and } A_\kappa = \begin{bmatrix} -k_1 & k_{-1} & 0 \\ k_1 & -(k_{-1} + k_2) & k - 2 \\ 0 & k_2 & -k_{-2} \end{bmatrix}.$$

In chemical reaction network theory, it is common to assume mass action kinetics. The law of mass action assumes that if reactions take place at constant temperature in a homogenous and well mixed solution then the probability of a collision between molecules is proportional to the product of their concentrations. This means that $\ln \Psi(x) = Y^{\mathrm{T}} \ln x$, and one reformulates the set of nonlinear ODEs given by (3) as [18]:

$$\dot{x} = YA_\kappa \Psi(x). \tag{4}$$

In general, we assume that a chemical reaction network has n species and m complexes. Thus, in (4): $x \in \bar{\mathbb{R}}_+^n$, $\Psi(x) \in \bar{\mathbb{R}}_+^m$, $A_\kappa \in \mathbb{R}^{m \times m}$, and $Y \in \bar{\mathbb{R}}_+^{n \times m}$. Experimental data is stacked in vector $\Psi(x)$ and often matrix Y is known such that we can explicitly search for the network structure given by A_κ. Finally, for clarity, we provide the expanded ODE representation of the Michaelis-Menten reaction given by (2):

$$\begin{aligned}
[\dot{E}] &= -k_1[E][S] + (k_{-1} + k_2)[ES] - k_{-2}[E][P],\\
[\dot{S}] &= -k_1[E][S] + k_{-1}[ES],\\
[\dot{ES}] &= k_1[E][S] - (k_{-1} + k_2)[ES] + k_{-2}[E][P],\\
[\dot{P}] &= k_2[ES] - k_{-2}[E][P].
\end{aligned} \tag{5}$$

Determining affine and sparse interconnections in dynamical systems

Consider a dynamical system of the following form:

$$\dot{x} = Af(x), x \in \mathbb{R}^n, A \in \mathbb{R}^{n \times m}, \tag{6}$$

where $f(\cdot) \in \mathbb{R}^m$ is a vector of known functions, which satisfy appropriate smoothness conditions to ensure local existence and uniqueness of solutions. With $A = Y A_\kappa$ and $f(x) = \Psi(x)$, the above description results in a dynamical system of the form given by (6). Note that the unknown parameters (which also encode the network structure) are in A, which enters the system dynamics linearly. Let neither the value of the entries nor the structure of matrix A be known. What we wish to find is the structure and entries in matrix A, given experimental data.

For this purpose, we consider the following discrete-time system:

$$x(t_{k+1}) = x(t_k) + (t_{k+1} - t_k)\, Af(x(t_k)), \tag{7}$$

which is the Euler discretization of (6).

Now, the set of measurements, which we denote by $\hat{x}$, can be used to fit the unknown entries to A such as to minimize the error between the data and the model predictions, which are given by (7). It is popular to solve the minimization problem which results in the least 2-norm of the error between $x_i(t_{k+1})$ and $\hat{x}_i(t_{k+1})$ (least squares minimization problem). We can write such an error metric as:

$$\min||Ma - b||_2 \tag{8}$$

where $a \in \mathbb{R}^{nm}$ is a vector containing A_{ij}, which we treat as decision variables, and $M \in \mathbb{R}^{((p-1) \times n) \times nm}$ and $b \in \mathbb{R}^{(p-1) \times n}$ are defined by 'stacking' all such conditions obtained by manipulating the data as per (7). Here p corresponds to the number of measurements. This problem has the following analytical solution:

$$a^* = M^\dagger b \triangleq (M^T M)^{-1} M^T b \tag{9}$$

There are a few drawbacks of the above least-squares approach. In the presence of extra constraints on the variables A_{ij} (constrained regression), the problem does not have a closed-form solution, in general. Such constrained minimizations, the simplest of which is a *Second Order Cone Problem* (SOCP) [19], may carry a significant computational cost for large problems. The same is true if the data are contaminated with error which needs to be taken into account when producing $A_{\text{least-squares}}$ [20]. Furthermore, the solution to a least-squares problem will not be sparse in general; it will rather result in a full matrix.

In [19] and more recently in [21], the fact was mentioned that a large number of elements of the solution z of a *Linear Program* (LP) such as

$$\min ||z||_1, \tag{10}$$

are zero, that is, (10) produces sparse solutions. For this reason, this is the approach we follow in the paper. In particular, if A is sparse then the following program seeks explicitly to minimize the entries to matrix A and, thus, tries to recover this property of the matrix:

$$\begin{aligned}
\min \quad & \|\text{vec}(A)\|_1\\
\text{s. t.} \quad & -\mu_k^- \le -\hat{x}(t_{k+1}) + \hat{x}(t_k) + (t_{k+1} - t_k)Af(\hat{x}(t_k)) \le \mu_k^+,\\
& \mu_k^+ \ge 0,\ \mu_k^- \ge 0,\ \forall k,\ k = 1, \ldots, p-1.
\end{aligned} \tag{11}$$

By making μ_k^+ and μ_k^- as small as possible for all k, we can ensure that the data are in close Euler-fit with the model making the approximation error as small as possible. The advantage of solving LPs is that the task can be performed using fast, efficient and readily available algorithms. Note also that the number of decision variables in (11) relates directly to the size of A and not of the data, which makes it suitable for the identification of large-scale systems. Support for the validity of above heuristic to obtain a sparse interconnection matrix A are also Theorem 1.1 of [22] and the work presented in [23].

An additional advantage of our approach is also that we may incorporate uncertainties in the measurements with little additional computational complexity. If we model the uncertainty in the measurements as

$$\tilde{x}(t_k) - \epsilon(k) \le \hat{x}(t_k) \le \tilde{x}(t_k) + \epsilon(k), \quad \tilde{f}(t_k) - \delta(k) \le f(\hat{x}(t_k)) \le \tilde{f}(t_k) + \delta(k), \quad \epsilon(k), \delta(k) \ge 0, \tag{12}$$

$\tilde{x}\ (t_k) \ge 0$, $\tilde{f}\ (t_k) \ge 0$, for all k, and $A_{ij} \ge 0$ then we can formulate the robust counterpart to (11) that is still an LP (see also [24,25]). The following LP is a robust formulation of program (11):

$$\begin{aligned} \min \quad & \| \operatorname{vec}(A) \|_1 \\ \text{s. t.} \quad & -\mu_k^- \leq -\tilde{x}(t_{k+1}) - \epsilon(k+1) + \tilde{x}(t_k) - \epsilon(k) + (t_{k+1} - t_k)A\left(\tilde{f}(t_k) - \delta(k)\right), \\ & -\tilde{x}(t_{k+1}) + \epsilon(k+1) + \tilde{x}(t_k) + \epsilon(k) + (t_{k+1} - t_k)A\left(\tilde{f}(t_k) + \delta(k)\right) \leq \mu_k^+, \\ & A_{ij} \geq 0, \forall i, j, \tilde{x}(t_k), \tilde{f}(t_k), \epsilon(k), \delta(k), \mu_k^+, \mu_k^- \geq 0, \forall k, k = 1, \ldots, p-1. \end{aligned} \tag{13}$$

In summary, using the above ideas, we aim to extract from data the sparsity pattern in matrix A, which is related to the connectivity of the underlying graph structure, drawing conclusions on the network interaction topology.

Finally, note that if data points are rare, that is $p \leq m$, and there are not any constraints on matrix A then the error between the data and the model predictions can be made zero and (9) does not have a unique solution. However, the following LP can be used to try to recover the sparsity structure of the matrix:

$$\begin{aligned} \min \quad & \|\operatorname{vec}(A)\|_1 \\ \text{s. t.} \quad & \hat{x}(t_{k+1}) = \hat{x}(t_k) + (t_{k+1} - t_k)Af(\hat{x}(t_k)), \ \forall k, \ k = 1, \ldots, p-1. \end{aligned} \tag{14}$$

Obtaining the structure of a gene regulatory network

Using the same ideas, another class of a networks that can be determined are gene regulatory networks given microarray time-series data. We first briefly explain the function of gene regulatory networks and DNA microarray time-series.

A gene encodes the information necessary to produce a specific protein. The process, in which the information is used to synthesize a protein, is highly controlled and this control allows the cell to vary the level of a particular protein in the cell depending on the cell's need for this protein. The first step of synthesizing a protein from a gene is RNA polymerase transcribing gene information from DNA to mRNA (see Figure 1a). This mRNA is then translated into synthesized proteins by ribosomes. Control can occur at a number of stages including the synthesis of mRNA, subsequent processing of the mRNA, control of the ribosome and control of mRNA stability. Some proteins, called transcription factors, are responsible for the regulation of the initiation of transcription. A transcription factor binds to the DNA, at the promoter site, in order to either inhibit or activate (or alternatively increase the rate of) the transcription of mRNA that is responsible for the synthesis of a specific protein (see Figure 1b). (Note that self regulation is also possible.) The collection of DNA segments which interact with each other in the manner described is called the gene regulatory network.

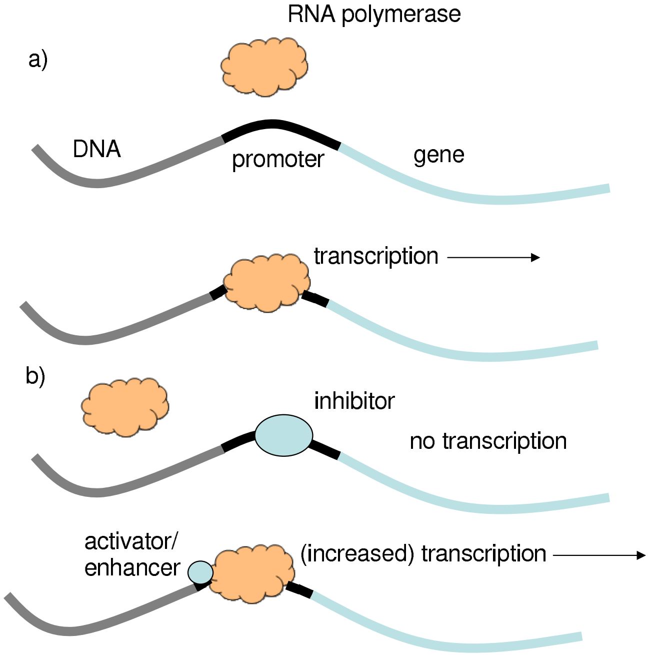


Figure 1
Diagram showing the process of transcription. a) The RNA polymerase binds to the promoter sequence on the DNA and transcribes a gene. b) Transcription can be controlled by inhibitors or activators acting at the promoter sequence.

In order to understand the dynamics and behavior of a gene regulatory network, three levels of information are required:

1. The network interconnection in the form of a *directed graph*;

2. Whether an edge from node *i* to node *j* means that transcription factor *i* is activating (denoted by arrow) or repressing (denoted by 'hammer') *j*;

3. The activation/repression rates for the transcription factors.

Time-series obtained from DNA microarrays [26,27] are extremely helpful to obtain the structure of gene regulatory networks. This is because DNA microarrays allow observation of the presence of specific mRNA within the cell and in particular, time-series data allow measurements on how these change over time after a perturbation, or when following the cell cycle.

Now, consider the model of a gene regulatory network as described in [28] and [29], where nodes represent genes. Knowledge of the three hierarchal levels of information mentioned previously is necessary to describe the network. The first level determines whether there is an interaction between proteins X_1 and X_2. An interaction of the form '$X_1 \to X_2$' means that protein X_1 activates the production of protein X_2 and '$X_1 \dashv X_2$' that X_1 inhibits it. This information belongs to the second level. The activation and repression Hill input functions are given mathematically by (see [28]):

$$\frac{kx_1^n}{1+kx_1^n}, \text{ and } \frac{1}{1+kx_1^n}, \tag{15}$$

respectively, where x_1 is the concentration of X_1. (In [29], the notation $\frac{1}{K}$ is used instead of k. For clarity, we have adopted a 'simpler' notation.) Knowledge about the magnitude of activation or repression coefficient k, $k > 0$, and exponent n, $n > 0$, is part of the third level of information.

If there exists more than one connection to a particular gene/node then all transcription factors associated with the connections could be necessary to fulfill a specific task (activation or inhibition) or it might be that any of them is sufficient to have an effect on the transcription process; more complex combinations are also possible. In [28,30], a generalized input function of the following form is presented, which takes the possible interplay of different transcription factors into account:

$$f_i(x) = \frac{\sum_j b_{ij} x_j^{n_{ij}}}{1+\sum_j k_{ij} x_j^{m_{ij}}}. \tag{16}$$

Here, activation of protein X_i by protein X_j is represented by $n_{ij} = m_{ij} > 0$, and repression by $n_{ij} = 0$, $m_{ij} > 0$. The contribution of the different transcription factors on the transcription rate is denoted by b_{ij}. Putting everything together, the mathematical description of the dynamics of the concentrations of protein X_i of an arbitrarily large gene regulatory network is as follows:

$$\dot{x}_i = \gamma_i + f_i(x) - d_i x_i, \tag{17}$$

where $\gamma_i > 0$ is the basal transcription production rate and $d_i > 0$ is the degradation/dilution rate. In the above model, however, the vector field (right hand side of Equation (17)) is not affine in the unknown parameters and therefore one may think that the method proposed in the previous section can not be extended for this case; we show here how the above can be reformulated and cast in a form that allows identification using Linear Programming.

Let $\Delta t = t_{\ell+1} - t_\ell$. A discrete-time system that approximates (17) is:

$$x_i(t_{\ell+1}) = x_i(t_\ell) + \Delta t(\gamma_i + f_i(x_i(t_\ell)) - d_i x_i(t_\ell)). \tag{18}$$

Indeed, if b_{ij}, k_{ij} and m_{ij} are unknown then (18) is not affine in the unknown parameters as is the case in (7). We rewrite (18) as follows:

$$(x_i(t_\ell)(1-\Delta t d_i) - x_i(t_{\ell+1}) + \Delta t \gamma_i)(1+\sum_j k_{ij} x_j^{m_{ij}}) + \Delta t \sum_j b_{ij} x_j^{\tilde{n}_{ij}} + \Delta t b_i = 0. \tag{19}$$

In (19), if $n_{ij} = 0$ then we denote it by $\bar{n}_{ij}$ and let $b_i = \sum_j b_{ij} x_j^{\bar{n}_{ij}} = \sum_j b_{ij}$. If $n_{ij} > 0$ then we denote it by $\tilde{n}_{ij}$. For all i, j, let an entry to matrix B be b_{ij} for which $n_{ij} > 0$, and let an entry of matrix K be k_{ij}. As before, given a set of measurements, which we denote by $\hat{x}$, this set can be used to approximate the structure of the gene regulatory network determined by b_{ij}, b_i and k_{ij} if the Hill coefficients m_{ij} (and, thus, n_{ij}) are known and the basal production and degradation rates are known or considered uncertain but within a known range. For instance, we can try to recover B, K through a LP. The following LP puts emphasis on minimizing the 1-norm of vec(B), b, and vec(K), which are independent of each other, while we keep the Euler discretisation error, μ, as small as possible.

$$\begin{array}{ll} \min & \|\mathrm{vec}([B\ b\ K])\|_1 \\ \text{s. t.} & -\mu < (\hat{x}_i(t_\ell)(1-\Delta t d_i) - \hat{x}_i(t_{\ell+1}) + \Delta t \gamma_i)(1+\sum_j k_{ij} \hat{x}_j^{n_{ij}}) + \Delta t \sum_j b_{ij} \hat{x}_j^{n_{ij}} + \Delta t b_i < \mu, \\ & \mu > 0,\ b_{ij} \geq 0,\ k_{ij} \geq 0,\ b_i \geq 0,\ \forall i, j, \ell\ (0 \leq \epsilon_{1i} \leq \gamma_i \leq \epsilon_{2i}, 0 \leq \varepsilon_{1i} \leq d_i \leq \varepsilon_{2i}, \forall i). \end{array} \tag{20}$$

(The latter requirements in brackets correspond to the case of uncertain production and degradation rates.) Note that as per (16)

$$k_{ij} = 0 \text{ if and only if } b_{ij} = 0 \text{ or } b_i = 0,\ \forall i, j. \tag{21}$$

The following remark deals with the case when the solution of (20) violates (21). The rationale behind the idea is that by following these rules we can determine unambiguously whether activation or repression takes place.

Remark 1 *Since requirement (21) cannot be implemented in a LP, we deduce the following from the solution of (20) about the connectivity of the network when (21) is violated:*

- if $k_{ij} \neq 0$, $b_{ij} = 0$ and $b_i = 0$ then the production of X_i is not affected by X_j; that is, it is the same case as when $k_{ij} = 0$,

- *if* $b_{ij} \neq 0$ *and* $k_{ij} = 0$ *then* X_j *enhances the production of* X_i*; i.e., it is the same case as when* $k_{ij} \neq 0$,

- *if* $b_i \neq 0$ *and* $k_{ij} = 0$ *for all i then the production of* X_i *is not affected by* X_j*; that is, it is the same case as when* $b_i = 0$.

Results and discussion

Numerical experiments

An artificial chemical reaction network

Consider the network with 5 species $\mathcal{S} = \{A, B, C, D, E\}$ and 5 complexes, $\mathcal{C} = \{A, 2B, A + C, D, B + E\}$ in Figure 2. Suppose we are given time series data for all the species in this system, but we do not know the topology of the interconnection. The first experiment examines the recovered interconnection diagram using the (non-robust) LP (11). Later on, we will consider the same problem with missing data on one species and a robust network determination problem.

We have implemented the network shown in Figure 2 with a K matrix of the form:

$$K = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.8492 \\ 0.3386 & 0 & 0 & 0 & 0.4290 \\ 0.8244 & 0 & 0 & 0.0563 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.7364 & 0 & 0 \end{bmatrix} \tag{22}$$

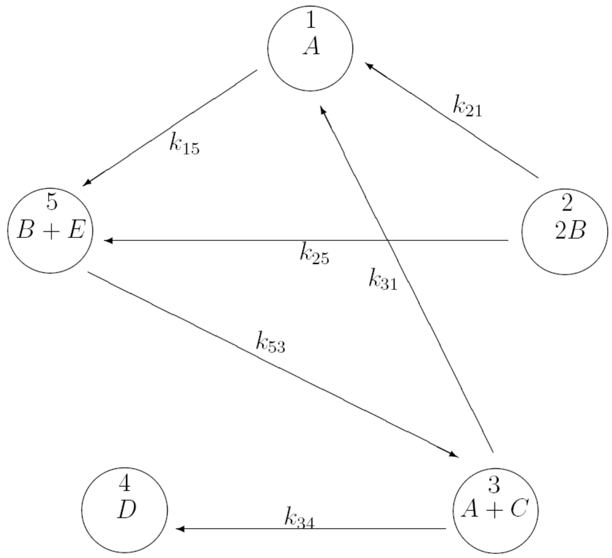


Figure 2
A chemical reaction network.

and have simulated the system with uniformly distributed initial conditions. The data sets were obtained by simulating the above set of nonlinear equations using SIMULINK. Ten such data sets were generated and incorporated in the LP.

Since we do not know how the chemical network is connected, and we cannot even speculate how parts of it may be connected, we need to assume a general structure for it and write the dynamics for the complete network. A least-squares approach, would yield the following structure in matrix K, where non-zero entries denote the fractional occurrences of non-zero k_{ij}'s for the 10 data sets:

$$\begin{bmatrix} 0 & 0.1 & 1 & 0.1 & 1 \\ 1 & 0 & 0.8 & 0.2 & 1 \\ 1 & 0.6 & 0 & 1 & 0.9 \\ 0.1 & 0.8 & 0.9 & 0 & 0 \\ 1 & 0.9 & 1 & 0.9 & 0 \end{bmatrix}$$

Essentially the only zero element predicted is k_{45}. Note that the diagonal of this matrix does not enter into our optimization. We write these entries as zero, but this is merely a convenient notational place holder. The resulting structure of the K matrix from our linear programming approach is given by

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0.5 & 0.4 & 0.9 \\ 1 & 0 & 0 & 1 & 0.2 \\ 0 & 0 & 0.8 & 0 & 0 \\ 0.1 & 0 & 1 & 0.2 & 0 \end{bmatrix}$$

where again non-zero entries denote the fractional occurrences of non-zero entries for the 10 data sets tested. Observe that the second column is equal to zero which implies that the second complex is not the product of any reaction. Having determined this sparse structure for the system, we can repeat the same LP optimization, but now impose the new information about the sparse structure obtained in the new linear program, i.e. that $k_{12} = 0$ etc. Iterating once on this data, we get the following results:

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0.5 & 0.3 & 0.7 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 \end{bmatrix}$$

This experiment reveals that the sparsity structure can be further reduced by an iterative procedure. One could also use the above as a 'probability' lookup table, and investigate sparsity structures, such as setting k_{23} and k_{24} equal to zero. Indeed this solution is also feasible, which reveals additional structure in the matrix K. Working this way, we have found that the following non-zero matrix results in feasible LPs:

$$K_{\text{nom}} = \begin{bmatrix} 0 & 0 & k_{13} & 0 & k_{15} \\ k_{21} & 0 & 0 & 0 & k_{25} \\ k_{31} & 0 & 0 & k_{34} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & k_{53} & 0 & 0 \end{bmatrix}$$

which is the same as the network shown in Figure 2, but for a link between complex 1 and complex 3. Of course, this is not surprising: there is no unique reaction mechanism that can fit a data set; rather, there can be many networks which with some kinetic parameters can represent the same data within experimental error. In fact, we can only hope to *invalidate* a postulated reaction mechanism using data, a point we will return to in the concluding section.

The next experiment we performed was to assume that some of the species could not be observed in the experiments for technical reasons. In particular, we assumed that the concentration of species A could not be measured. This does not pose significant problems, as we can replace the occurrences of the terms in the vector field involving the variable x_1 with a vector of new variables q which we also ask to be 'sparse', through minimization of the sum of q_i. Eight such substitutions need to be made; the result is a matrix of the form:

$$K_{\text{miss}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ k_{21} & 0 & k_{23} & 0 & k_{25} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & k_{43} & 0 & 0 \\ 0 & 0 & k_{53} & k_{54} & 0 \end{bmatrix}$$

and a $q = [q_1, ..., q_8]$ which corresponds to nonzero entries for k_{31}, k_{34}, k_{35} k_{13} and k_{15}. Therefore in this case too, a sparse topology interconnection is obtained, but the matrix in this case is not as sparse as before.

Suppose now that data are uncertain, and we want to search for *robust* sparse structures for the K matrix. We set $\epsilon_i^+ = \epsilon_i^- = 0.0004$ for $i = 1, ..., 5$ and all data points – such uncertainty could be due to roundoff errors (see Equation (12)). A robust LP can be formulated, as discussed earlier, and the obtained optimization results in a network with a richer sparsity structure:

$$K_{\text{rob}} = \begin{bmatrix} 0 & 0 & k_{13} & 0 & k_{15} \\ k_{21} & 0 & k_{23} & 0 & k_{25} \\ k_{31} & 0 & 0 & k_{34} & 0 \\ 0 & 0 & k_{43} & 0 & 0 \\ 0 & 0 & k_{53} & 0 & 0 \end{bmatrix}$$

Finally, we note that once a candidate network is determined, we can perform a least-squares minimization to obtain the best k values for a particular sparsity structure. For example, if we choose K_{nom} as the sparsity structure and fit the least squares error over all 10 experiments, we get the following K matrix:

$$K = \begin{bmatrix} 0 & 0 & 0.0364 & 0 & 0.7721 \\ 0.3295 & 0 & 0 & 0 & 0.3999 \\ 0.7804 & 0 & 0 & 0.0553 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.6668 & 0 & 0 \end{bmatrix} \tag{23}$$

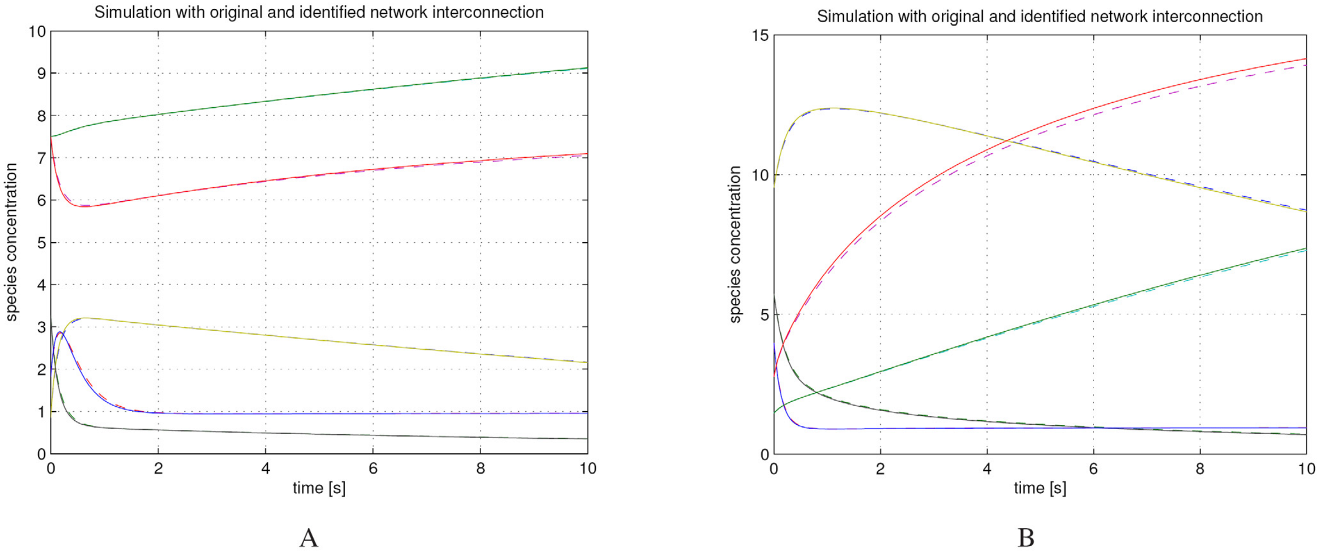
In figures 3A and 3B we show how the nominal system, with the K matrix given by Equation (22) compares in simulation with the K matrix given by Equation (23) for different initial conditions. We see that some initial conditions have better behavior for the two parameter sets than others. There is hope, however, that using other methods and through choice of a particular initial condition we can distinguish between the two network structures; the initial condition in Figure 3B shows some deviation in the dynamics of x_1, and designing an experiment that would yield 'maximum' deviation would allow for differentiability between various models that can explain the same data. More details about this approach can be found in [31].

A sample gene regulatory network

Consider the artificial gene regulatory network modeled through

$$\begin{aligned} \dot{x}_1 &= \gamma_1 - d_1 x_1, \\ \dot{x}_2 &= \gamma_2 + \frac{b_{12}x_1}{1+k_{12}x_1} - d_2 x_2, \\ \dot{x}_3 &= \gamma_3 + \frac{b_{43}x_4 + b_{13}x_1 + b_3}{1+k_{43}x_4 + k_{13}x_1 + k_{53}x_5} - d_3 x_3, \\ \dot{x}_4 &= \gamma_4 + \frac{b_{54}x_5}{1+k_{54}x_5} - d_4 x_4, \\ \dot{x}_5 &= \gamma_5 + \frac{b_{15}x_1 + b_5}{1+k_{15}x_1 + k_{25}x_2} - d_5 x_5, \end{aligned} \tag{24}$$

where



A

B

Figure 3
Simulation of chemical reaction networks. Simulation of the network with reaction rates (22) (solid line) and with reaction rates given by (23) (dashed line) from two initial conditions.

$$B=\begin{bmatrix}0 & 0.51 & 0.87 & 0 & 0.80\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0.20 & 0 & 0\\ 0 & 0 & 0 & 0.22 & 0\end{bmatrix}, K=\begin{bmatrix}0 & 0.31 & 0.87 & 0 & 0.15\\ 0 & 0 & 0 & 0 & 0.77\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0.97 & 0 & 0\\ 0 & 0 & 0.79 & 0.44 & 0\end{bmatrix},$$

$b_3 = 0.71$, $b_5 = 0.80$, $\gamma_i = 0.1$ and $d_i = 1$. The network is depicted in Figure 4, where solid lines with an arrow head denote activation and dash pointed lines with a hammer head denote inhibition.

We assume that d_i is known and $\gamma_i = \gamma$ for all i, where $0.095 \leq \gamma \leq 0.105$. We take 'measurements' every $\Delta t = 0.05$ between $t = 0$ and $t = 5$ (time is in arbitrary units) from four different random initial conditions between 0 and 1 in order to obtain mock-up data. Solving (20) using the solver SeDuMi [32], we obtain the following results for matrices B and K:

$$B=\begin{bmatrix}0 & 0.48 & 0.22 & 0 & 1.15\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0.11 & 0\end{bmatrix}, K=\begin{bmatrix}0 & 0 & 0 & 0 & 0.61\\ 0 & 0 & 0 & 0 & 0.75\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0.32 & 0 & 0\\ 0 & 0 & 0.35 & 0 & 0\end{bmatrix};$$

$b_3 = 0.64$, $b_5 = 0.80$, and all other $b_i = 0$. Following the rules given by Remark 1, we are able to reconstruct the network shown in Figure 4. As the example shows, we are able to determine the interaction network given by (24) through the LP (20) even when degradation rates are considered uncertain.

Reconstructing the glycolytic pathway of Lactococcus lactis

Lactococcus lactis is a bacterium used in the industrial production of cheese and buttermilk as it converts more than

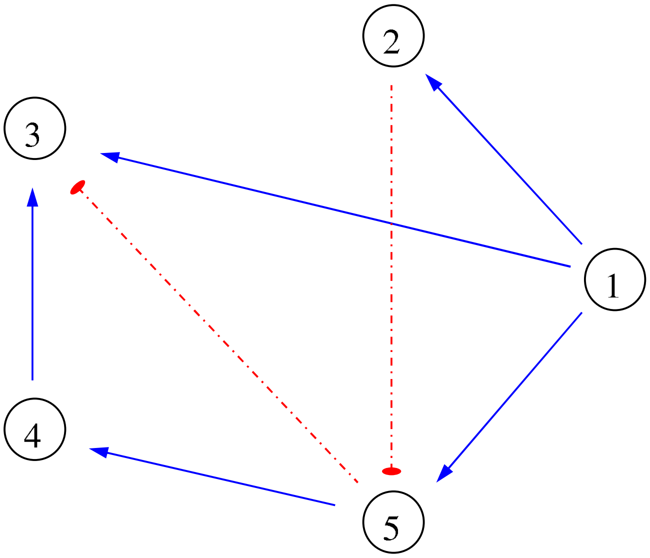


Figure 4
The artificial gene regulatory network modeled through (24). Solid lines with an arrow head denote activation and dash pointed lines with a hammer head denote inhibition.

90% of lactose (milk sugar) to lactic acid [14]. In general, the glycolytic pathway (or glycolysis) consists of chemical reactions that convert glucose into pyruvate. In the first step, glucose is converted into glucose-6-phosphate (G6P). A conversion of G6P into fructose-1,6-bisphosphate (FBP) follows, which is then converted sequentially to glyceraldehyde-3-phosphate (Ga3P), 3-phosphoglyceric acid (3-PGA) and PEP [16]. Additionally, Glucose and PEP are converted directly to pyruvate and G6P. Note that since measurement data for the intermediate Ga3P were unavailable, we include an additional rate denoting depletion of FBP. A simplified description of the pathway from reference [33] is depicted in Figure 5. The relative simplicity of this metabolic network makes *L. lactis* an attractive model for biological systems approaches [14]. A recent paper which presents an approach to determine the connectivity of this system and puts some emphasis on its sparsity is [16]. However, this approach does not take into account the characteristic particulars that make up a chemical reaction network. Here, we first use LP (11) to try to elucidate the glycolytic pathway of *L. lactis* using the same experimental data from [33].

Particularly, we wish to elucidate the glycolytic pathway of *L. lactis* under the assumption that the following complexes participate in the chemical reaction network: Glu, G6P, FBP, 2 × 3PGA, 2 × PEP, 2 × Pyru and Lact. In other words, we wish to obtain interaction topology $A\kappa$ of the chemical reaction network given by $\dot{x} = Y A_\kappa f(x)$, where

$$Y = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, x = \begin{bmatrix} [\mathrm{Glu}] \\ [\mathrm{G6P}] \\ [\mathrm{FBP}] \\ [\mathrm{3PGA}] \\ [\mathrm{PEP}] \\ [\mathrm{Pyru}] \\ [\mathrm{Lact}] \end{bmatrix}, f(x) = \begin{bmatrix} [\mathrm{Glu}] \\ [\mathrm{G6P}] \\ [\mathrm{FBP}] \\ [\mathrm{3PGA}]^2 \\ [\mathrm{PEP}]^2 \\ [\mathrm{Pyru}]^2 \\ [\mathrm{Lact}] \end{bmatrix}.$$

Note that the network topology is completely determined by A_κ. Recall that

$$A_\kappa = K - \mathrm{diag}(K^{\mathrm{T}} e),\ K_{ij} \geq 0\ \forall i, j. \qquad (25)$$

Now, by solving (11) we indeed obtain a sparse chemical reaction topology (Figure 6(a)). However, the error between the model dynamics and experimental data is unreasonably large (Figure 6(b)). Therefore, it is not surprising that this configuration differs greatly from the the proposed pathway of Figure 5.

The ℓ_1 regularized least squares problem, which is called *Lasso* is statistics, considers an objective function to minimize, which consists of the sum of the 1-norm of the vector of unknowns and the least squares of the error:

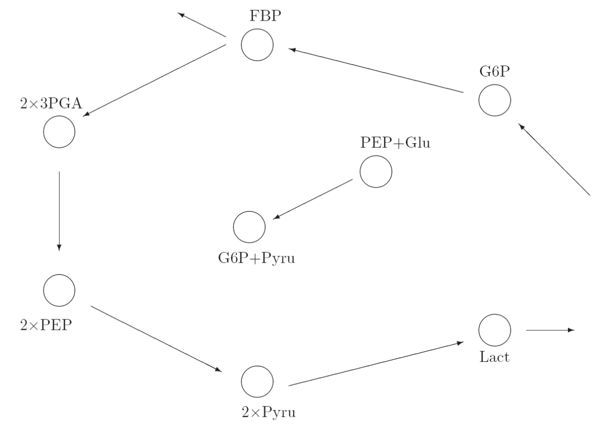


Figure 5
The glycolytic pathway of *Lactococcus lactis*.

$$\begin{array}{ll} \text{given} & Y \\ \min & \left\| \begin{bmatrix} \hat{x}(t_1) - \hat{x}(t_2) + (t_2 - t_1) A f(\hat{x}(t_1)) \\ \vdots \\ \hat{x}(t_{p-1}) - \hat{x}(t_p) + (t_p - t_{p-1}) A f(\hat{x}(t_{p-1})) \end{bmatrix} \right\|_2 + \alpha \left\| \mathrm{vec}(A) \right\|_1 \\ \text{s. t.} & A = Y A_\kappa, \\ & A_{\kappa_{i,j}} \geq 0,\ i \neq j,\ \forall i, j,\ e^{\mathrm{T}} A_\kappa = 0 \text{ (this follows from (25))}, \end{array} \qquad (26)$$

where α is a nonnegative constant that allows us to regulate the weight we put on the sparsity of A explicitly. Note that for α = 0, program (26) minimizes the the error between data and model dynamics solely (Figure 7(b)). This time, the error between the model dynamics and experimental data is considerably smaller. The connection topology is shown in Figure 7(a). Now, we increase α to see whether or which interconnections disappear without altering the system dynamics significantly. This pathway, which remains unaltered for $2 \leq \alpha \leq 10$, is shown in Figure 7(a). The dynamic behavior of this system is indistinguishable from the one shown in Figure 7(b) and, thus, is not shown.

Further increase of α results first in the disappearance of the links between G6P and FBP, and sequential changes do not result in 'sensible' connection topologies. Of course, this is something that in general the investigator does not know. While the pathway depicted in Figure 6(a) might be dismissed because the resulting model behavior compares badly with data, this argument does not hold for the pathway in Figure 7(a).

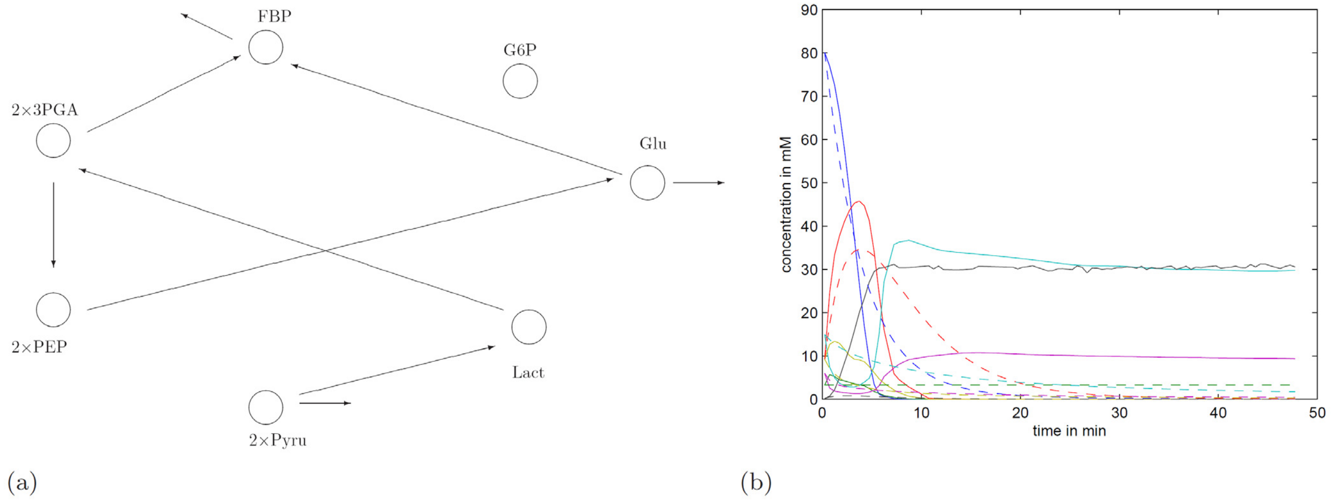


Figure 6
Reaction pathway obtained through (11). a) The reaction pathway obtained through (11). b) The simulated model dynamics defined through the reaction network shown in (a) are shown in dashed lines and solid lines correspond to experimental data.

Now, we exploit the following related approach to try to deduce the interactions of the system by solving the following LP:

$$\begin{array}{ll} \text{given} & Y \\ \min & \left\| \begin{bmatrix} \hat{x}(t_1) - \hat{x}(t_2) + (t_2 - t_1)Af(\hat{x}(t_1)) \\ \vdots \\ \hat{x}(t_{p-1}) - \hat{x}(t_p) + (t_p - t_{p-1})Af(\hat{x}(t_{p-1})) \end{bmatrix} \right\|_1 + \alpha \|\mathrm{vec}(A)\|_1 \\ \text{s. t.} & A = YA_\kappa, \\ & A_{\kappa_{i,j}} \geq 0,\ i \neq j,\ \forall i,j,\ e^{\mathrm{T}} A_\kappa = 0 \text{ (follows from (25))}, \end{array} \qquad (27)$$

We solve (27) for $\alpha = 0$, $\alpha = 2$ and $\alpha = 3$, and obtain the reaction pathway shown in Figure 8(a) which results in a model with the dynamics depicted in Figure 8(b). (Note that for $0 < \alpha \leq 75$, the model dynamics are indistinguishable from the ones shown in Figure 8(b) and are not shown.)

The error between the model dynamics and experimental data is again considerably smaller than the error shown in Figure 6(b). As we can see from Figure 8(a), a relatively sparse reaction topology was obtained.

The pathway $x_1 \to \ldots \to x_7$ was almost reconstructed. A sensible assumption to make is that the degradation of G6P which appears at $\alpha = 3$ corresponds to the conversion into FBP suggested at $\alpha = 2$.

Also, the direct link between glucose and pyruvate was discovered. Finally, with

$$A_\kappa = \begin{bmatrix} -0.3452 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0185 & -0.4105 & 0 & 0 & 0 & 0 & 0 \\ 0.2431 & 0.4105 & -0.3735 & 0 & 0 & 0.0350 & 0 \\ 0 & 0 & 0.0009 & -0.0008 & 0 & 0 & 0.0237 \\ 0 & 0 & 0 & 0.0008 & -0.0106 & 0 & 0.0079 \\ 0.0835 & 0 & 0 & 0 & 0 & -0.0377 & 0 \\ 0 & 0 & 0.1551 & 0 & 0.0105 & 0.0027 & -0.0393 \end{bmatrix}$$

our approach provides a meaningful chemical reaction network of the form (4). (This matrix corresponds to the case when $\alpha = 2$.) Nevertheless, without biochemical information the superiority of this pathway to the pathway in Figure 7(a) cannot be established and it follows that experiments have to be designed to discriminate between several competing models.

Conclusion

We have presented a methodology for determining the interaction topology of biological networks, that are either affine in the unknown parameters or can be transformed to have this property, using time series data collected from experiments. We demonstrated the ability of our method to identify a chemical reaction network structure through several numerical examples. We have also tested our approach by elucidating the glycolytic pathway of the bacterium *Lactococcus lactis*. Our method respects the structural properties that chemical reaction network dynamics should have [11,12].

In the case of gene regulatory networks, more realistic models could be used, but those would include additional parameters, first, by making the Hill coefficient in the acti-

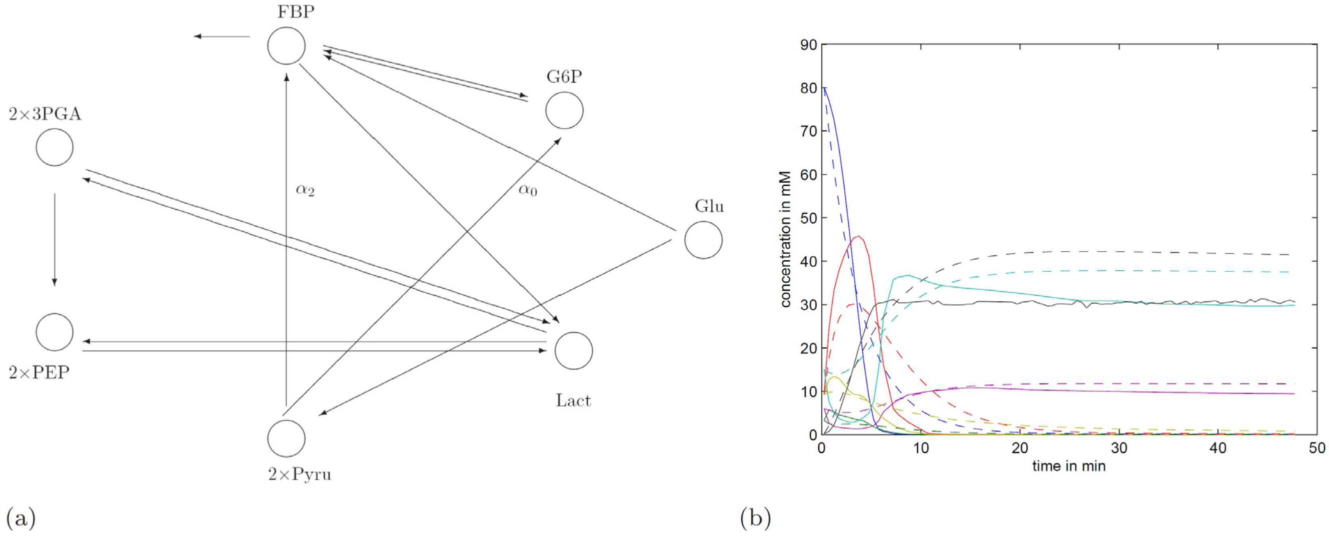


Figure 7
Reaction pathway obtained through (26). a) Reaction pathway obtained through (26) for $\alpha = 0$ and $\alpha = 2$. All reactions participate in both pathways except for two which are marked accordingly. The one reaction that was obtained for $\alpha = 0$ but not for $\alpha = 2$ is marked with α_0 and the one that appears only for $\alpha = 2$ is marked with α_2. b) Here, solid lines correspond to experimental data and dashed lines to the model with the interaction matrix obtained by solving (26) with $\alpha = 0$.

vation and repression terms a free variable; and second, encoding the fact that when two transcription factors act on DNA, either both are required (AND) or any of them is sufficient (OR) for action. Thus, a valuable research direction is to investigate this case and establish whether similar analysis techniques to the ones presented in this paper can be used.

In (27) we introduced a free variable α whose value can change the solution considerably. Hence, it is worthwhile to explore different possible heuristics how to choose the value of this variable. (Here, we kept the balance between increasing α and keeping the model dynamics that followed from the solution of (27) relatively close to experimental data.) An iterative method can also be used, which

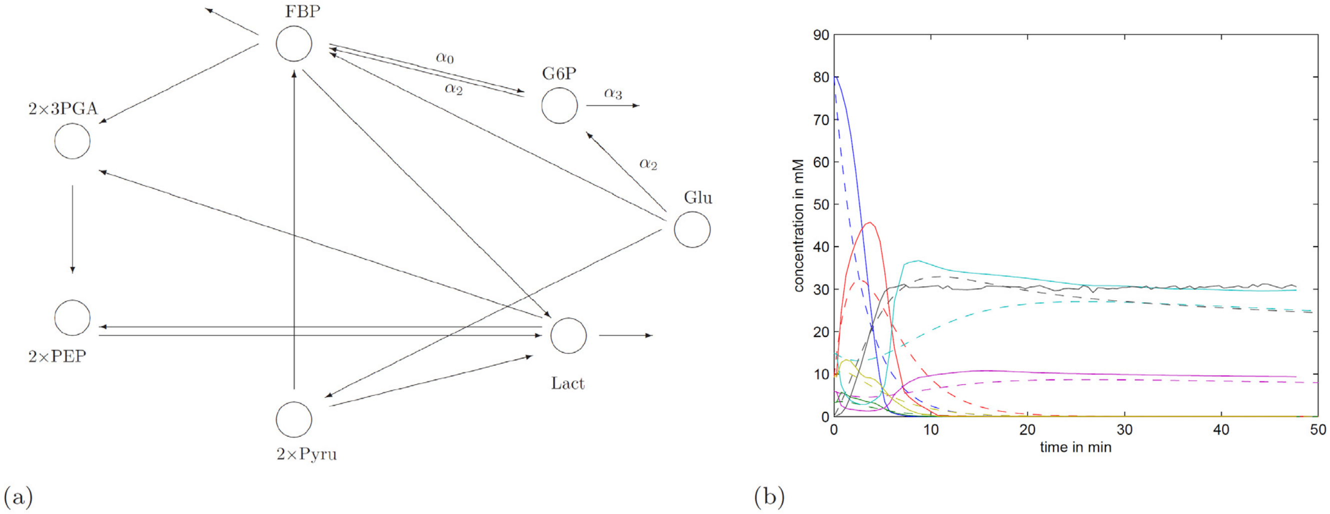


Figure 8
Reaction pathway obtained through (27). a) All reactions participate in both pathways except for four which are marked accordingly. Two reactions that were obtained for $\alpha = 0$ and $\alpha = 2$ but not for $\alpha = 3$ are marked with α_0 and α_2, one that appears only for $\alpha = 0$ is marked with α_0, and one that appears only for $\alpha = 3$ is marked with α_3. (Note that a gradual increase of α, for $3 \leq \alpha \leq 75$, did not change the network structure.) b) Here, solid lines correspond to experimental data and dashed lines to the model dynamics defined through the reaction network shown in Figure 8(a) for $\alpha = 0$.

uses 'live' information from simulations and then iterates with a simple Linear Program to find the network structure that fits best the parameters.

Finally, as shown, different methods or the same one with different constraints provide different models that represent the same data, which is an expected feature of such methods. It follows that experiments have to be designed to discriminate between competing models, in a way that 'closes the loop' between modelling and experiment (see for example [31]).

Authors' contributions

The authors contributed equally to this work. The first author developed the algorithm for determining gene regulatory networks and performed the example of *L. Lactis* while the second author developed the algorithm for biochemical reaction networks and conceived the general idea. Both authors read and approved the final manuscript.

Acknowledgements

This work was financially supported by EPRSC project E05708X. The authors would like to thank Ben Recht and Mark Roberts for useful discussions about this paper.

References

1. Yeung MKS, Tegnér J, Collins JJ: **Reverse engineering gene networks using singular value decomposition and robust regression.** *PNAS* 2002, **99(9):**6163-6168.
2. Tegnér J, Yeung MKS, Hasty J, Collins JJ: **Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling.** *PNAS* 2003, **100(10):**5944-5949.
3. Julius AA, Zavlanos M, Boyd S, Pappas GJ: **Genetic network identification using convex programming.** In *Technical Report MS-CIS-07-20* Department of Computer and Information Science, University of Pennsylvania; 2007.
4. Kholodenko BN, Kiyatkin A, Bruggeman FJ, Sontag ED, Westerhoff HV, Hoek JB: **Untangling the wires: A strategy to trace functional interactions in signalling and gene networks.** *PNAS* 2002, **99(20):**12841-12846.
5. Kholodenko BN, Sontag ED: **Determination of functional network structure from local parameter dependence data.** 2002.
6. Vance W, Arkin A, Ross J: **Determination of causal connectivities of species in random networks.** *PNAS* 2002, **99(9):**5816-5821.
7. Westerhoff HV, Kolodkin A, Conradie R, Wilkinson SJ, Bruggeman FJ, Krab K, van Schuppen JH, Hardin H, Bakker BM, Moné MJ, Rybakova KN, Eijken M, van Leeuwen HJP, Snoep JL: **Systems biology towards life in silico: mathematics of the control of living cells.** *Journal of Mathematical Biology* 2009, **58(1–2):**7-34.
8. Crampin EJ, Schnell S, McSharry PE: **Mathematical and computational techniques to deduce complex biochemical reaction mechanisms.** *Prog Biophys Mol Biol* 2004, **86(1):**77-112.
9. Ross J, Schreiber I, Vlad MO: *Determination of Complex Reaction Mechanisms: Analysis of Chemical, Biological and Genetic Networks* Oxford, UK: Oxford University Press; 2006.
10. Chevalier T, Schreiber I, Ross J: **Toward a systematic determination of complex reaction mechanisms.** *J Phys Chem* 1993, **97(26):**6776-6787.
11. Feinberg M: **Lectures on chemical reaction networks.** 1979 [http://www.che.eng.ohio-state.edu/~FEINBERG/LecturesOnReactionNetworks]. Mathematics Research Centre, University of Wisconsin
12. Feinberg M: **Chemical reaction network structure and the stability of complex isothermal reactors-I. The deficiency zero and deficiency one theorems.** *Chem Eng Sci* 1987, **42(10):**2229-2268.
13. Papachristodoulou A, Recht B: **Determining Interconnections in Chemical Reaction Networks.** In *Proceedings of the 2007 American Control Conference* New York City, USA; 2007:4872-4877.
14. Voit EO, Almeida J, Marino S, Lall R, Goel G, Neves AR, Santos H: **Regulation of glycolysis in *Lactococcus lactis*: An unfinished systems biological case study.** *Syst Biol (Stevenage)* 2006, **153(4):**286-298.
15. del Rosario RCH, Mendoza E, Voit EO: **Challenges in lin-log modelling of glycolysis in *Lactococcus lactis*.** *IET Systems Biology* 2008, **2:**136-149.
16. Srividhy J, Crampin EJ, McSharry PE, Schnell1 S: **Reconstructing biochemical pathways from time course data.** *Proteomics* 2007, **7:**828-838.
17. Gunawardena J: **Notes on Metabolic Control Analysis.** 2002 [http://vcp.med.harvard.edu/papers/mca.pdf]. Bauer Center for Genomics Research, Harvard University, Cambridge, MA 02138, USA
18. Gunawardena J: **Chemical reaction network theory for in-silico biologists.** 2003 [http://vcp.med.harvard.edu/papers/crnt.pdf]. Bauer Center for Genomics Research, Harvard University, Cambridge, MA 02138, USA
19. Boyd S, Vandenberghe L: *Convex Optimization* Cambridge, UK: Cambridge University Press; 2004.
20. El Ghaoui L, Lebret H: **Robust Solutions to Least-Squares Problems with uncertain data.** *SIAM J Matrix Anal Appl* 1997, **18(4):**1035-1064.
21. Zavlanos MM, Julius AA, Boyd SP, Pappas GJ: **Identification of Stable Genetic Networks using Convex Programming.** Proceedings of the American Control Conference, Seattle, Washington, USA; 2008:2755-2760.
22. Candès EJ, Romberg J, Tao T: **Stable signal recovery from incomplete and inaccurate measurements.** *Communications of Pure and Applied Mathematics* 2006, **59:**1207-1223.
23. Donoho DL, Tanner J: **Sparse nonnegative solution of underdetermined linear equations by linear programming.** *PNAS* 2005, **102(27):**9446-9451.
24. Soyster AL: **Convex Programming with set-inclusive constraints and applications to inexact linear programming.** *Operations Research* 1973, **21:**1154-1157.
25. Bertsimas D, Sim M: **The Price of Robustness.** *Operations Research* 2004, **52:**35-53.
26. Filkov V, Skiena S, Zhi J: **Analysis Techniques for Microarray Time-Series Data.** *Journal of Computational Biology* 2002, **9(2):**317-330.
27. Hana S, Yoon Y, Choc KH: **Inferring biomolecular interaction networks based on convex optimization.** *Comput Biol Chem* 2007, **31(5-6):**347-354.
28. Alon U: *An Introduction to Systems Biology: Design Principles of Biological Circuits* Boca Raton, FL: Chapman & Hall/CRC; 2006.
29. Mangan S, Alon U: **Structure and function of the feed-forward loop network motif.** *PNAS* 2003, **100(21):**11980-11985.
30. Setty Y, Mayo AE, Surette MG, Alon U: **Detailed map of a cis-regulatory input function.** *PNAS* 2003, **100(13):**7702-7707.
31. Papachristodoulou A, El-Samad H: **Algorithms for Discriminating Between Biochemical Reaction Network Models: Towards Systematic Experimental Design.** *Proceedings of the 2007 American Control Conference, New York City, USA* 2007:2714-2719.
32. Sturm JF: **Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones.** *Optimization Methods and Software* 1999, **11–12:**625-653 [http://sedumi.ie.lehigh.edu/].
33. Neves AR, Ventura R, ansour N, Shearman C, Gasson MJ, Maycock C, Ramos A, Santos H: **Is the Glycolytic Flux in *Lactococcus lactis* Primarily Controlled by the Redox Charge? Kinetics of NAD$^+$ and NADH Pools Determined *in vivo* by ^{13}C NMR.** *The Journal of Biological Chemistry* 2002, **277(31):**28088-28098.