Inference of complex biological networks: distinguishability issues and optimization-based solutions

Gábor Szederkényi*[1,2], Julio R. Banga*[1] and Antonio A. Alonso[1]

[1](Bio)Process Engineering Group, IIM-CSIC, Spanish National Research Council, C/Eduardo Cabello, 6, 36208 Vigo, Spain
[2]Process Control Research Group, MTA SZTAKI, Kende u. 13-17, H-1111 Budapest, Hungary

Email: Gábor Szederkényi*- szeder@scl.sztaki.hu; Julio R. Banga*- julio@iim.csic.es; Antonio A. Alonso - antonio@iim.csic.es;

*Corresponding author

Abstract

Background: The inference of biological networks from high-throughput data has received huge attention during the last decade and can be considered an important problem class in systems biology. However, it has been recognized that reliable network inference remains an unsolved problem. Most authors have identified lack of data and deficiencies in the inference algorithms as the main reasons for this situation.

Results: We claim that another major difficulty for solving these inference problems is the frequent lack of uniqueness of many of these networks, especially when prior assumptions have not been taken properly into account. Our contributions aid the distinguishability analysis of chemical reaction network (CRN) models with mass action dynamics. The novel methods are based on linear programming (LP) , therefore they allow the efficient analysis of CRNs containing several hundred complexes and reactions. Using these new tools and also previously published ones to obtain the network structure of biological systems from the literature, we find that, often, a unique topology cannot be determined, even if the structure of the corresponding mathematical model is assumed to be known and all dynamical variables are measurable. In other words, certain mechanisms may remain undetected (or they are falsely detected) while the inferred model is fully consistent with the measured data. It is also shown that sparsity enforcing approaches for determining 'true' reaction structures are generally not enough without additional prior information.

Conclusions: The inference of biological networks can be an extremely challenging problem even in the utopian case of perfect experimental information. Unfortunately, the practical situation is often more complex than that, since the measurements are typically incomplete, noisy and sometimes dynamically not rich enough, introducing further obstacles to the structure/parameter estimation process. In this paper, we show how the structural uniqueness and identifiability of the models can be guaranteed by carefully adding extra constraints, and that these important properties can be checked through appropriate computation methods.

Background

During the last decade, the wide availability of high-throughput biological data has made it possible to produce new knowledge via a systems biology approach [1–3]. The inference of biochemical networks (i.e. the mathematical mapping of the molecular interactions in the cell) is therefore a question of key importance in the field. During the last decade, many methods have been developed to solve the network-inference (sometimes called reverse-engineering [4]) problems arising in e.g. gene expression [5–13], signal transduction [14–17] and metabolic networks [18–25].

In this context, it is particularly worth mentioning the DREAM initiative (Dialogue for Reverse Engineering Assessments and Methods) [26], which targeted the problems of cellular network inference and quantitative model building in systems biology. DREAM tries to address two fundamental questions: (i) how can we assess how well we are describing the networks of interacting molecules that underlie biological systems? and (ii) how can we know how well we are predicting the outcome of previously unseen experiments from our models? Interestingly, one of the main conclusions of the DREAM3 event was that the vast majority of the teams' predictions were statistically equivalent to random guesses. Moreover, even for particular problem instances like gene regulation network inference, there was no one-size-fits-all algorithm [27].

The use of a performance profiling framework with the DREAM3 benchmark problems revealed that current inference methods are affected by different types of systematic prediction errors [6]. These authors conclude that reliable network inference from gene expression data remains an unsolved problem. Further, they highlight two major difficulties in the case of gene-network reverse engineering: limited data (which may leave the inference problem underdetermined), and the difficulty of distinguishing direct from indirect regulation. Prill et al [27] further explored the issue of intrinsic impediments to network inference, designating identifiability of certain network edges and systematic false positives as the main barriers.

In this paper, we consider the widely used reaction kinetic formalism, where dynamic models of biological networks are described by a set of ordinary differential equations (see, e.g. [28–30] and the related literature). In particular, we consider the central question of the identifiability of such a network as understood in the systems and control area [31, 32].

Identifiability analysis studies whether there is a theoretical chance of uniquely determining the parameters of a mathematical model assuming perfect noise-free measurements and error-free modeling [33–35]. One of the early approaches for identifiability testing of nonlinear models is based on the Taylor-series expansion of the system output using the fact that the Taylor coefficients are unique [36]. A similar but more general method uses the generating series or Volterra-series coefficients of the system which is the nonlinear generalization of the Laplace-transform method used for linear systems [37]. In [38] a similarity transformation approach is proposed that gives necessary and sufficient conditions on local and global identifiability through the checking of nonlinear controllability and observability conditions. The appearance of differential algebra methods in systems and control theory [39, 40] opened the possibility for new types of identifiability tests that have gained significant popularity [41–43]. Further theoretical developments in the field include the identifiability conditions of rational function state space models [43], the possible effect of initial conditions on identifiability [44], and the application of Lie-algebras [45]. While identifiability is the property of a certain parameterized model, a related notion called distinguishability addresses the problem whether two or more parameterized models (with the same or with different structure) can produce the same output for any allowed input [46–48]. The literature about identifiability and distinguishability of biological and chemical system models is relatively wide: Compartmental systems (that form a special subclass of general mass-action networks) are studied in [38, 49, 50]. The authors treat general nonlinear CRNs in [51, 52] and [53] where it is shown that for thermodynamically meaningful models, nonlinearity reduces the chance of indistinguishability compared to the linear case [54]. Geometric conditions for the indistinguishability of CRNs are given in [55] with a related comment in [56]. Computer algebra tools can be successfully used for the symbolic computations needed for identifiability and distinguishability testing of complex models [57–60] .

The importance of identifiability has been recognized previously in systems biology, too [14, 61–64]. However, and despite a number of works illustrating ways to test the structural and practical identifiability of models [65–67], a significant portion of modeling studies in systems biology continue to ignore this key property.

It has been known for long that chemical reaction networks with different structure and/or parametrization may produce the same dynamical models describing the time-evolution of species concentrations [28, 55]. A related problem, namely the non-unique structure of Petri nets associated to reaction network dynamics, is studied in [68]. Additionally, the value of prior information in biological network inference was clearly shown in [69, 70] by applying Bayesian network models. However, a constructive optimization-based approach for the study of dynamically equivalent (or similar) reaction networks is a recent development [71–74], which we further extend in this paper.

As a novelty, we present in this paper the definition and a computational method to find the so-called core reactions that are present in any dynamically equivalent reaction network if the set of complexes is given a priori. Moreover, a computationally improved method is introduced for the computation of dense realizations of CRNs together with a modified algorithm to check the uniqueness of a constrained reaction network structure. Structural non-uniqueness and the use of the proposed computational methods will be illustrated with the help of biological models known from the literature.

The structure of the paper is the following. The 'Methods' section introduces the notions of chemical reaction networks, structural identifiability and distinguishability of dynamical models. Moreover, it contains the procedures to obtain core reactions of a network and its sparse and dense representations, which rely on standard methods of linear programming (LP) and mixed integer linear programming (MILP) [75–78]. The analysis of four biological system models can be found in the 'Results and discussion' section, followed by the conclusions.

Methods

The model class considered in this paper is of the following form

$$\begin{aligned} \dot{x} &= f(x, u, \theta), \quad x(0) = x_0 \\ y &= h(x, u, \theta), \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$ is the state vector, $y \in \mathbb{R}^m$ is the output, $u \in \mathbb{R}^k$ is the input, and $\theta \in \mathbb{R}^d$ denotes the parameter vector. We assume that the functions f and h are polynomial in the variables x, u and θ. Clearly, mass action type CRNs described in the following subsection (where θ is typically the set of reaction rate coefficients), and simple deterministic models of gene regulation such as the one in Example 4 belong to this model class.

Basic notions and known results related to mass-action models

In this subsection, the basic definitions for the description of CRNs will be given together with the already published results on finding dynamically equivalent network realizations with certain prescribed properties.

Structural and dynamical description of mass-action networks

Following [79] and several other works, we will characterize CRNs with the following three sets.

1. $\mathcal{S} = \{X_1, \dots, X_n\}$ is the set of *species* or chemical substances.

2. $\mathcal{C} = \{C_1, \dots, C_m\}$ is the set of *complexes*. Formally, the complexes are represented as linear combinations of the species, i.e.

$$C_i = \sum_{j=1}^{n} \alpha_{ij} X_j, \quad i = 1, \dots, m, \tag{2}$$

   where α_{ij} are nonnegative integers and are called the *stoichiometric coefficients*.

3. $\mathcal{R} = \{(C_i, C_j) \mid C_i, C_j \in \mathcal{C}, \text{ and } C_i \text{ is transformed to } C_j \text{ in the CRN}\}$ is the set of *reactions*. The relation $(C_i, C_j) \in \mathcal{R}$ will be denoted as $C_i \to C_j$. Moreover, a nonnegative weight, the *reaction rate coefficient* denoted by k_{ij} is assigned to each reaction $C_i \to C_j$. Naturally, if the reaction $C_i \to C_j$ is not present in the CRN then $k_{ij} = 0$.

The above characterization naturally gives rise to the following graph structure (often called 'Feinberg-Horn-Jackson graph' or simply reaction graph) of a CRN [29]. The weighted directed graph $G = (V, E)$ of a CRN consists of a finite nonempty set V of vertices and a finite set E of ordered pairs of distinct vertices called directed edges. The vertices correspond to the complexes, i.e. $V = \{C_1, C_2, \dots C_m\}$, while the directed edges represent the reactions, i.e. $(C_i, C_j) \in E$ if complex C_i is transformed to C_j in the CRN. The positive reaction rate coefficients k_{ij} are assigned as weights to the corresponding directed edges $C_i \to C_j$ in the graph. (Edges corresponding to zero rate coefficients are not drawn in the reaction graph.) A set of complexes $\{C_1, \dots, C_k\}$ is called a *linkage class* of a CRN, if the complexes of the set are linked to each other in the reaction graph but not to any other complex. It is remarked that loops (i.e. directed edges that start and end at the same vertex) are not allowed in reaction graphs.

Assuming mass-action kinetics, the following dynamical description will be used to describe the time-evolution of species concentrations [29, 79]:

$$\dot{x} = Y \cdot A_k \cdot \psi(x), \tag{3}$$

where x_i denotes the concentration of species X_i. Let us denote the (i,j)th element of an arbitrary matrix W by $W_{i,j}$, where i is the row index and j is the column index. The jth column of Y contains the composition of complex C_j, i.e. $Y_{i,j} = \alpha_{ji}$. The structure and parameters of the reaction graph are stored in the column conservation matrix A_k (also called the *Kirchhoff matrix* of the CRN) as follows

$$[A_k]_{i,j} = \begin{cases} -\sum_{l=1,l\neq i}^{m} k_{il}, & \text{if} \quad i = j \\ k_{ji}, & \text{if} \quad i \neq j. \end{cases} \tag{4}$$

Finally, $\psi : \mathbb{R}^n \mapsto \mathbb{R}^m$ is a monomial-type vector mapping defined by

$$\psi_j(x) = \prod_{i=1}^{n} x_i^{Y_{i,j}}, \quad j = 1, \ldots, m. \tag{5}$$

Dynamical equivalence of mass-action networks

As it is known even from the early literature [28], CRNs with different structures and/or parametrization can give rise to the same kinetic differential equations. Therefore, we will call two CRNs given by the matrix pairs $(Y^{(1)}, A_k^{(1)})$ and $(Y^{(2)}, A_k^{(2)})$ *dynamically equivalent*, if

$$Y^{(1)} A_k^{(1)} \psi^{(1)}(x) = Y^{(2)} A_k^{(2)} \psi^{(2)}(x) = f(x), \quad \forall x \in \bar{\mathbb{R}}_+^n, \tag{6}$$

where for $i = 1, 2$, $Y^{(i)} \in \mathbb{R}^{n \times m_i}$ have nonnegative integer entries, $A_k^{(i)}$ are valid Kirchhoff matrices, and

$$\psi_j^{(i)}(x) = \prod_{k=1}^{n} x_k^{[Y^{(i)}]_{k,j}}, \quad i = 1, 2, \; j = 1, \ldots, m_i. \tag{7}$$

In this case, $(Y^{(i)} A_k^{(i)})$ for $i = 1, 2$ are called *realization*s of a kinetic vector field f (see, e.g. [80] for more details). It is also appropriate to call $(Y^{(1)}, A_k^{(1)})$ a *realization* of $(Y^{(2)}, A_k^{(2)})$ and vice versa.

We will assume throughout the paper that the set of complexes (i.e. the stoichiometric matrix Y) is fixed and known before the computations. In this case, the condition (6) for dynamical equivalence can be written as

$$Y \cdot A_k^{(1)} = Y \cdot A_k^{(2)} =: M, \tag{8}$$

where $A_k^{(1)}$ and $A_k^{(2)}$ are valid Kirchhoff matrices and M is the invariant matrix containing the coefficients of the monomials.

Among the dynamically equivalent realizations, it is important to recall the following characteristic ones described in [71, 72]. A *sparse realization* contains the minimal number of reactions that is needed for the exact description of the corresponding dynamics (3). A *dense realization* contains the maximal number of reactions among dynamically equivalent realizations with a fixed complex set given by Y. While sparse realizations are generally structurally non-unique (as it will be illustrated for the constrained case, too, in Example 1), the structure of dense realizations with a given complex set is unique, and it contains every possible dynamically equivalent structure as a proper subgraph (i.e. a dense realization is a kind of super-structure) [71].

Known computation approaches for finding preferred CRN realizations

Here we briefly summarize the already published results corresponding to the computation of preferred dynamically equivalent CRN realizations (more details of these methods can be found in the publications [71–73, 81]). The computation of dense and sparse realizations can be traced back to mixed integer linear programming (MILP) where the decision variables are the non-diagonal elements of A_k, the linear constraints encode the kinetic properties of the model, and the objective function contains integer variables for minimizing/maximizing the number of nonzero reaction rate coefficients [72]. It is remarked that the computation of sparse realizations is an NP-hard problem, where generally mixed integer linear programming cannot be avoided [82]. There exist certain conditions under which the problem can be solved in polynomial time [83] but these are often not fulfilled in the case of CRNs. Moreover, there are effective heuristics to address the problem [84], but convergence to one of the truly sparsest structures is not guaranteed. Luckily, the MILP-based computation of sparse CRN realizations can be parallelized effectively thus allowing a larger number of complexes to be treated. The computation of realizations having the minimal/maximal number of complexes or the reversibility property can also be solved in the MILP framework [71]. Moreover, it was shown in [73] that finding detailed balanced and complex balanced realizations of CRNs is a simple linear programming (LP) problem. Finally, weakly reversible dynamically equivalent CRN realizations can also be determined (if they exist) using MILP [85].

Constrained realizations of CRNs and testing their structural uniqueness

The following is a straightforward extension of the results published in [71]. To prove the uniqueness of a CRN structure given a set of simple constraints, we have to extend the notions of dense and sparse realizations. The constraint set denoted by $\mathcal{K}$ will be used for the exclusion of selected reactions from the CRN, i.e. it is of the form:

$$\mathcal{K} = \{[A_k]_{i_1,j_1} = 0, \ \ldots, \ [A_k]_{i_s,j_s} = 0\}, \tag{9}$$

where s is the number of individual constraints, and $i_k \neq j_k$ for $k = 1, \ldots, s$. Now we can introduce the following definitions. A dynamically equivalent *constrained realization* of a CRN (Y, A_k) is a reaction network (Y, A'_k) such that $Y \cdot A_k = Y \cdot A'_k$ and the prescribed constraints $\mathcal{K}$ in the form of eq. (9) are fulfilled for A'_k. A dynamically equivalent *constrained dense realization* of a CRN (Y, A_k) is a constrained realization that contains the maximal number of nonzero elements in A'_k. Similarly, the *constrained sparse realization* is a constrained realization with the minimal number of nonzeros in A'_k. To characterize constrained dense/sparse realizations, the results of [71] can be adapted easily as follows.

P1 Given a CRN (Y, A_k) and a constraint set $\mathcal{K}$, the unweighted reaction graph of any constrained realization is the subgraph of the unweighted reaction graph of the constrained dense realization.

P2 If the sets of complexes and constraints are fixed, then for any CRN, the structure of the constrained dense realization is unique.

P3 The reaction graph structure of a CRN with given sets of complexes and constraints is unique if and only if the unweighted directed graphs of its constrained dense and sparse realizations are identical.

The proofs of **P1**, **P2** and **P3** follow similar (although not completely identical) lines that were published in [71], and they are given for convenience in subsection A.1 of the Appendix at the end of the paper.

New concepts and computation results related to dynamically equivalent networks

This subsection contains new methodological contributions that extend the previously published results.

Making the computation of dense realizations more efficient

Computing dense realizations is treated originally also in a MILP-framework in [72]. However, using the structural uniqueness of such realizations given by **P1**, it is easy to give a polynomial-time algorithm based on a finite series of linear programming (LP) optimization steps. The idea of the improved algorithm is simple: the reaction $C_i \to C_j$ belongs to the (constrained) dense realization if and only if there exists any dynamically equivalent (constrained) realization where $[A_k]_{j,i} > 0$. The result directly follows from the fact that the unweighted reaction graphs of (constrained) dense realizations give a unique super-structure. This allows us to formulate a polynomial-time method based on pure LP to determine (constrained) dense realizations. The details of the computations corresponding to this improved method are described in subsection A.2 of the Appendix.

Using the notion and described properties of constrained realizations, we are now able to test the structural uniqueness of given CRNs. To accomplish this, only the (constrained) dense and sparse realizations have to be computed and compared (see **P3**). This method will be illustrated in Example 2.

Definition and computation of core and non-core reactions

We will call a reaction a *core reaction*, if it is present in any dynamically equivalent realization of a CRN with a given complex set (and possibly an additional constraint set). Other reactions, the rate coefficient of which can be zero in certain realizations, are called *non-core reactions*. It clearly follows from the definition, but is remarked separately that the set of core reactions is generally not identical to the set of reactions of a sparse realization. The identification of core reactions of a CRN has not been published yet, therefore we give the outline of the corresponding computation method. Firstly, a dense realization of the network has to be computed to get all the mathematically possible reactions. Then, for each reaction $C_p \to C_q$ in the dense realization, the feasibility of the following constraint set has to be checked:

$$Y \cdot A_k = M \tag{10}$$

$$\sum_{i=1}^{m} [A_k]_{i,j} = 0, \quad j = 1, \ldots, m \tag{11}$$

$$[A_k]_{i,j} \geq 0, \quad i, j = 1, \ldots, m, \quad i \neq j, \quad (i, j) \neq (q, p) \tag{12}$$

$$[A_k]_{i,i} \leq 0, \quad i = 1, \ldots, m \tag{13}$$

$$[A_k]_{q,p} = 0, \tag{14}$$

where the matrix A_k contains the decision variables, and the known matrices are Y and M. It is well-known that this task is equivalent to an LP problem where the objective function is an arbitrary linear function of the elements of A_k [76]. Then, reaction $C_p \to C_q$ is a core-reaction if and only if the set defined by (10)-(14) is empty (i.e. the corresponding LP problem is infeasible), because in this case there is no dynamically equivalent CRN realization where $C_p \to C_q$ is not present. We remark here that the presented procedures for determining constrained dense realizations and computing core reactions are parallel in their original forms since the individual LP steps are independent of each other. Therefore the proposed methods can be very effectively implemented in a grid or multi-core hardware environment [86].

Basic concepts on structural identifiability and distinguishability

Let us recall eq. (1). Shortly speaking, global structural identifiability means that

$$\hat{y}(t|\theta') \equiv \hat{y}(t|\theta'') \Rightarrow \theta' = \theta'', \tag{15}$$

where

$$\hat{y}(t|\theta) = h(x(t, \theta), u(t), \theta), \tag{16}$$

and $x(t,\theta)$ denotes the solution of (1) with parameter vector θ. According to (15), a structurally non-identifiable model can produce exactly the same observed output with different parametrization. This is clearly a fundamental obstacle of determining the true model parameters from measurements even if the selected model structure is considered to be correct.

Let us denote two parameterized models with possibly different structure by $\mathcal{M}_1(\theta_1)$ and $\mathcal{M}_2(\theta_2)$, respectively, where θ_i denote the parameter vector. Then $\mathcal{M}_1$ is called *distinguishable* from $\mathcal{M}_2$ if for any θ_1 (possibly except for a finite number of values) there is no θ_2 such that the input-output behaviour of $\mathcal{M}_1$ and $\mathcal{M}_2$ is the same [47]. Clearly, if $\mathcal{M}_1$ and $\mathcal{M}_2$ are indistinguishable and both model structures are feasible in a certain application, then there is no way to decide from input-output measurements to which one corresponds to the true model that generated the data.

In the case of CRNs, we will assume that all species concentrations are measured (i.e. $y = x$), the input is zero (i.e. we study autonomous systems), and that the set of possible chemical complexes is given. Generally, the model parameter vector θ is the set of reaction rate coefficients which are the off-diagonal elements of A_k. Clearly, if a CRN has several different dynamically equivalent realizations, then these realizations are not distinguishable without additional constraints, and the model cannot be identifiable if all the rate coefficients are to be determined [55]. This situation can be improved by using prior knowledge in the form of adding further constraints on the model parameters such as the simple ones given by eq. (9). This way, the number of parameters to be estimated can be reduced and/or their feasibility region can be shrinked. It is important to note that although the structural uniqueness of a CRN definitely reduces the degree of non-identifiability (since zero and non-zero parameters are separated), it does not necessarily imply structural identifiability [53], and this latter property has to be checked by further numerical methods. [31, 32].

Results and Discussion

In this section, the application of the previously mentioned methods for finding different dynamically equivalent structures will be illustrated using biological models taken from the literature. The detailed numerical data corresponding to Examples 1-3 are contained in a standard spreadsheet form with brief explanations in Additional file 1: CRN_data.xls.

Example 1: a positive feedback motif

The first example is a positive feedback motif shown in Fig. 1.a and taken from [87] containing 5 species, 11 complexes and 9 reactions. This basic motif is also discussed in [88]. The network contains a gene that promotes its own transcription and translation after dimerization. In the model, X_1 and X_2 denote the concentrations of protein monomers and dimers, respectively. X_3 and X_4 are the concentrations of unoccupied and occupied promoters, respectively, and X_5 corresponds to the mRNA. The degradation of dimers is ignored. The roles of the reaction rate coefficients are the following: k_1 and k_2 are the dimerization and re-dimerization rates, respectively. k_3 and k_4 are the binding and dissociation rates of the dimer to the promoter, while k_5 and k_6 denote the activated and basal transcription rates, respectively. k_7 is the degradation rate of the mRNA, k_8 is the degradation rate of the monomer, and k_9 denotes the translation rate. The time-evolution of the species-concentrations is described by the following ODEs:

$$\dot{x}_1 = -2k_1x_1^2 + 2k_2x_2 + k_9x_5 - k_8x_1 \tag{17}$$

$$\dot{x}_2 = k_1x_1^2 - k_2x_2 - k_3x_2x_3 + k_4x_4 \tag{18}$$

$$\dot{x}_3 = -k_3x_2x_3 + k_4x_4 \tag{19}$$

$$\dot{x}_4 = k_3x_2x_3 - k_4x_4 \tag{20}$$

$$\dot{x}_5 = k_5x_4 + k_6x_3 - k_7x_5. \tag{21}$$

Our starting point is that we have a dynamic model of the process in the standard polynomial form of (17)-(21), the parameters of which are known from the results of identification and/or from literature. As we will see below, without well-defined constraints on the possible set of complexes and reactions, exactly the same dynamics can be realized in principle by a wide range of mechanisms.

The matrices characterizing the stoichiometry and graph structure of the system are the following (indicating only the nonzero non-diagonal elements of A_k):

$$Y = \begin{bmatrix} 2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \tag{22}$$

$$[A_k]_{2,1} = k_1,\ [A_k]_{1,2} = k_2,\ [A_k]_{4,3} = k_3,\ [A_k]_{3,4} = k_4,\ [A_k]_{5,4} = k_5, \tag{23}$$

$$[A_k]_{7,6} = k_6,\ [A_k]_{9,8} = k_7,\ [A_k]_{9,10} = k_8,\ [A_k]_{11,8} = k_9. \tag{24}$$

We used the following parameter values that were taken from the Appendix of [87].

$$k_1 = k_2 = k_3 = k_4 = 10^7,\ k_5 = 1.7,\ k_6 = 0.025,\ k_7 = 0.1,\ k_8 = 0.05,\ k_9 = 0.5, \tag{25}$$

where the units of measure are [M^{-1}] for $k_1, \ldots, k_4$, and [min^{-1}] for $k_5, \ldots, k_9$. The dynamically equivalent dense realization of the network is shown in Fig. 1.b, where the 8 core and 4 non-core reactions are indicated separately. The three different sparse structures are shown in the subplots of Fig. 2. The first subplot is identical to the original structure shown in Fig. 1.a. This means that the mechanism cannot be described exactly with less than 9 reactions. It turns out from the second and third subplots that (at least mathematically), the degradation of mRNA is dynamically not a necessary element of the model. However, the biological plausibility of the mathematically possible structures and reactions always has to be carefully examined.

As it is expected, the possible structures of sparse/dense realizations and the corresponding core and non-core reactions can change with the modification of parameter values. This is illustrated in Fig. 3.a, where the following randomly generated parameter values were used:

$$k_1 = 18.9,\ k_2 = 7.1,\ k_3 = 15.4,\ k_4 = 12.7,\ k_5 = 10.6,\ k_6 = 3.5,\ k_7 = 11.3,\ k_8 = 9.1,\ k_9 = 4.0. \tag{26}$$

It is visible that the structure of the dense realization is the same as in Fig. 1.b but the core reactions are different from the ones shown there. Here the degradation of mRNA is described by a core reaction but interestingly, the reaction corresponding to translation is not a core one. Naturally, this implies that the possible sparse realization structures with the second parametrization are different from the ones shown in Fig. 2. Note that here the only goal was to illustrate the possible change of core and non-core reactions, and therefore the biological relevance of the parameter values in eq. (26) is not assumed in this case.

In the next step, let us assume that another complex, namely $X_2 + X_4$ is allowed in the model (again not necessarily assuming biological meaningfulness in this particular case). With the addition of this new complex, the stoichiometric matrix of the system can be written as

$$Y' = \begin{bmatrix} 2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}. \tag{27}$$

The dense CRN realization of the dynamics (17)-(21) with the updated Y' matrix given in eq. (27) using the original parameters described in (25) is shown in Fig. 3.b, where the core and non-core reactions are again indicated. It is apparent that now there are only 5 core reactions, and none of the remaining 12 reactions are essential to represent the dynamics (17)-(21). This means that the introduction of a new complex increased the flexibility of the network (i.e. mathematically, the majority of the reactions can be substituted by other ones and the network still maintains its original dynamics). Of course, not any combination of the non-core reactions can be omitted from the network, because the sparse realizations show that at least 9 reactions are needed to keep dynamical equivalence. It can be computed easily that the theoretical maximum number of sparse realizations with different structures is $\binom{12}{17-9} = 495$. However, as the numerical experiments show, majority of these structures do not give a practically feasible dynamically equivalent realization.

The above results clearly show that certain mechanisms may remain undetectable (or they are falsely detected) even if we have complete species concentration measurements and full information about possible complex formation, that are not very realistic assumptions. Moreover, the sparsest dynamically equivalent structure of mass-action models is not unique, therefore sparsity enforcing approaches for determining 'true' reaction structures are not enough in themselves without the necessary amount of prior information given in the form of additional constraints. The practical situation is most often even worse than that, since the measurements are typically incomplete, noisy and sometimes dynamically not rich enough, that may introduce further obstacles to the structure/parameter estimation process [66, 89].

Example 2: a biochemical switch in yeast cells

The following example is taken from [90] and it describes a 'switching device' in yeast cycle regulation. The detailed system description can be found in [90] and in the accompanying supporting information document. The order of state variables, corresponding to concentrations, is the same as in the original article, and is shown below:

x_1: [Sic1], x_2: [Sic1P], x_3: [Clb], x_4: [Clb·Sic1], x_5: [Clb·Sic1P], x_6: [Cdc14], x_7: [Sic1P·Cdc14], x_8: [Clb·Sic1P·Cdc14], x_9: [Clb·Sic1·Clb]. The original structure with 18 reactions is shown in Fig. 4.a. The Y matrix of the network is given by

$$Y = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{28}$$

The non-zero off-diagonal elements of A_k are (the diagonal ones can be computed using the column conservation property):

$$\begin{aligned}
&[A_k]_{2,1} = k_3,\ [A_k]_{2,3} = k_2,\ [A_k]_{3,2} = k_1,\ [A_k]_{4,5} = k_5,\ [A_k]_{5,4} = k_4,\ [A_k]_{6,5} = k_6,\\
&[A_k]_{6,8} = k_9,\ [A_k]_{7,8} = k_8,\ [A_k]_{8,7} = k_7,\ [A_k]_{9,10} = k_{11},\ [A_k]_{10,9} = k_{10},\ [A_k]_{11,10} = k_{12},\\
&[A_k]_{12,13} = k_{14},\ [A_k]_{13,12} = k_{13},\ [A_k]_{14,13} = k_{15},\ [A_k]_{15,16} = k_{17},\ [A_k]_{16,15} = k_{16},\ [A_k]_{17,16} = k_{18}.
\end{aligned} \tag{29}$$

Since there are no parameter values published in [90], we used the following randomly selected rate coefficients:

$$k = [4.1\ 3.2\ 6.7\ 7.3\ 3.8\ 2.4\ 4.5\ 5.1\ 6.2\ 7.7\ 8.6\ 9.5\ 2.4\ 4.9\ 5.8\ 10.2\ 6.3\ 8.5]^T. \tag{30}$$

The structure of the dense realization indicating the 12 core and 16 non-core reactions can be seen in Fig. 4.b.

It can be shown using the computational methods described in the 'Methods' section that the only possible sparse realization structure is identical to that of the original network. Therefore in this special case, there is only one possible reaction structure containing the minimal number of reactions. A straightforward approach to ensure the structural uniqueness of the whole network is to exclude all reactions that are not meaningful from the examined application's point of view or that are contradictory to modeling assumptions. For the current example, the removal of an unexpectedly low number of reactions is enough to obtain a unique structure. It can be shown by computing the corresponding constrained dense and sparse realizations, that excluding the reactions $X_5 \to X_3 + X_5$, $X_4 \to X_3 + X_4$, $X_2 + X_3 \to X_3 + X_5$, and $X_3 + X_1 \to X_3 + X_4$ is enough to make the reaction structure unique that is identical to the original structure shown in Figure 4. In other words, the exclusion of 4 well-selected reactions leads to the removal of an additional 6 reactions leaving only 18.

Example 3: a repressilator structure with 5 nodes and auto-activation

Consider the repressilator model shown in Fig. 5 with 5 nodes where also auto-activation is assumed. Similarly to [91], we make the following assumptions: cooperative regulator binding, genes are present in constant amounts, transcription and translation are modeled by single-step kinetics, and finally, proteins are degraded by first order reactions. We note that complex dynamic phenomena such as multiple steady states or oscillations have been shown with a wide range of parameters in similar systems, especially in the case when the number of genes is odd [91]. We also assume that there is some protein production (leakage) when both the activator and the repressor are bound to the genes (although this assumption does not affect the main results of the forthcoming analysis). It is clearly shown in [92] that kinetic models with simple mass-action kinetics very effectively describe complex dynamics in genetic regulatory networks, therefore we follow the same modeling methodology. Using the assumptions listed above, the CRN describing the system is the following:

$$G_i + P_i \underset{k_{i,2}}{\overset{k_{i,1}}{\rightleftarrows}} G_i^A \quad \text{(auto-activation 1)} \tag{31}$$

$$G_i^A \overset{k_{i,3}}{\rightarrow} G_i^A + P_i \quad \text{(protein production 1)} \tag{32}$$

$$G_i + P_j \underset{k_{i,5}}{\overset{k_{i,4}}{\rightleftarrows}} G_i^R \quad \text{(repression 1)} \tag{33}$$

$$G_i^R + P_i \underset{k_{i,7}}{\overset{k_{i,6}}{\rightleftarrows}} G_i^{AR} \quad \text{(auto-activation 2)} \tag{34}$$

$$G_i^A + P_j \underset{k_{i,9}}{\overset{k_{i,8}}{\rightleftarrows}} G_i^{AR} \quad \text{(repression 2)} \tag{35}$$

$$G_i^{AR} \overset{k_{i,10}}{\rightarrow} G_i^{AR} + P_i \quad \text{(protein production 2)} \tag{36}$$

$$P_i \overset{k_{i,11}}{\rightarrow} 0 \quad \text{(protein degradation)} \tag{37}$$

for the index pairs $(i, j) \in \{(1, 5), (2, 1), (3, 2), (4, 3), (5, 4)\}$. In eqs. (31)-(37), G_i and P_i represent the ith gene and protein, respectively. For the genes, superscripts A and R refer to activated and repressed states, respectively. Let us denote with $r_{i,k}$ the reaction with rate coefficient $k_{i,k}$ in eqs. (31)-(37).

Two cases with different sets of randomly selected rate coefficients were studied, and the structures of the obtained results were the same. The numerical details can be found on the 3rd sheet of Additional File 1: CRN_data.xls. The total number of reactions for the repressilator model is 55 that is equal to the number of reactions in the sparse realization. The dense realization contains 70 reactions which means that there are a maximum of 15 more mathematically possible reactions while maintaining exactly the same dynamics as the original biological model. These additional reactions are the following:

$$G_i^{AR} \to G_i^R, \quad P_i + G_i^R \to G_i^R, \quad P_i + G_i^R \to P_i + G_i^{AR}, \text{ for } i = 1, \dots, 5. \tag{38}$$

The number of core reactions in the model are 45. The set of non-core reactions (that, in principle can be substituted by other reactions) is given by

$$G_i^{AR} \leftrightarrows G_i^R + P_i, \quad i = 1, \dots, 5. \tag{39}$$

In particular, it is easy to show (see also Additional File 1: CRN_data.xls) that reactions $G_i^{AR} \to G_i^R + P_i$ and $G_i^{AR} \to G_i^R$ are always indistinguishable. Similarly, the reaction $G_i^R + P_i \to G_i^{AR}$ can be substituted with the combination of reactions $G_i^R + P_i \to G_i^R$ and $G_i^R + P_i \to G_i^{AR} + P_i$. It can be seen from these results that in order to have a model with unique structure, it is very important to a priori exclude all reactions that are not meaningful for the particular application.

Example 4: sparse linear gene regulation network models

For structural identification, gene regulation networks are often modeled as linear time-invariant systems [84, 93] of the form

$$\dot{x} = Ax + Bu, \tag{40}$$

where $A \in \mathbb{R}^{n\times n}$ contains the connectivity information of the network. $A_{i,j} > 0$ indicates activation from node j to node i, while $A_{i,j} < 0$ means repression, diagonal elements of A represent auto-activation or auto-repression depending on their sign. $x \in \mathbb{R}^n$ is the fully or partially measurable state of the system describing the time evolution of concentrations, and the input part Bu represents experimental perturbation (e.g. activation) of the genes. It is also a common assumption that the network is 'sparse' which means that there are only a limited number of activation or repression links between the nodes (i.e. the matrix A is 'sparse', too). But assuming sparsity can be a serious obstacle to identifiability as it will be shown.

First, consider the 'true' genetic network structure that was simulated and inferred in [93] and that is redrawn in Fig. 6.a. From the figure, we can reconstruct the structure of the corresponding A matrix as follows (the exact parameter values are not described in the paper, but the investigated structural properties do not depend on the individual parameter values)

$$\begin{bmatrix}
* & 0 & 0 & 0 & 0 & 0 & 0 & 0 & + & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & + & + & 0 \\
0 & 0 & * & 0 & 0 & - & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & + & 0 & 0 & + \\
0 & 0 & 0 & + & 0 & 0 & 0 & - & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & * & 0 & 0 & + & 0 \\
0 & 0 & + & + & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & - & 0 & - & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & - & + & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & + & - & 0 & 0
\end{bmatrix}, \tag{41}$$

where '+', '-' and '*' represent positive, negative and nonzero (but otherwise undefined) parameter values, respectively. If there are no prior assumptions about the structure of the interconnection matrix or about the relations between certain parameters, we can easily test the structural non-identifiability of the model by checking whether all nodes are reachable from the perturbed node on a directed path in the interconnection graph or not [33]. The reachability of nodes can be tested by several methods, e.g. a depth- or breadth-first-search (DFS or BFS) of the corresponding directed graphs that are fast polynomial-time algorithms [94]. To give a very simple example, it is clear from Fig. 6.a, that if nodes 1 or 2 were excited by an input signal, then the connections between the other nodes (3-10) would be undetectable by any method, however sophisticated it is. To examine whether situations like this one are common, we generated 10000 random state space models using the same method, and assuming zero initial conditions as in [93]. The connectivity of the corresponding directed graphs was tested using DFS. For 10 nodes and 2 nonzero elements in each row of A (i.e. $N = 10$, $K = 2$), we obtained that 73.38% of the generated models are structurally non-identifiable. The histogram showing the number of reachable states is shown in Fig. 6.b. The situation is dramatically improving if K is increased to 3 as shown in Fig. 6.c. In this case, around 17% of the models are structurally non-identifiable. When we have 20 nodes and 5 nonzero elements in each row of A (the second case investigated in [93]), then only 1.6% of the generated models are structurally non-identifiable. The results show that ‘sparsity’ has a clearly negative effect on structural identifiability because of limited information transmission between nodes. And finally, we did not speak at all about practical identifiability which is known to be a challenging issue even if the required structural properties are fulfilled [66].

Relation between high level networks and CRN structure

As shown in Example 3, the various possible dynamically equivalent CRN structures do not correspond to a different GRN structure, if all species concentration measurements are available and the mapping described in [92] is used for transforming the models into each other. Hence, exact matching of the dynamics of different GRN structures may generally be a too severe restriction. To extend this line of research, the relaxation of dynamical equivalence to ‘close dynamical similarity’ seems to be more meaningful but the corresponding definitions and computational methods are much more complex than in the case of dynamical equivalence. One promising recent approach to assess dynamical similarity of CRNs (that also adds more degrees of freedom to the computations) is the concept of ‘linear conjugacy’ [74]. However, it might happen that dynamically completely equivalent GRN structures will be shown in the future.

Conclusions

It has been shown in this paper using illustrative examples that biological network structures modeled by CRNs often cannot be uniquely determined even if the structure of the corresponding mathematical model is assumed to be known and all dynamical variables are measurable. The structural uniqueness and identifiability of the models often require additional constraints.

The main new contributions of the paper are the following. Firstly, core reactions present in any dynamically equivalent CRN realizations with a given complex set have been defined and a simple procedure with polynomial time-complexity has been given for determining them. Clearly, the core reactions are mandatory elements of every dynamically equivalent CRN realization assuming a fixed complex set. Secondly, a polynomial-time method based on linear programming for computing dense realizations has been outlined that is more scalable and therefore presents a clear improvement over the previously used MILP-based method. As an additional minor extension of previous results, constrained realizations of CRNs have been defined, and a computational method has been proposed to check the uniqueness of constrained realizations.

The presented concepts and algorithms were illustrated on previously published models describing biological processes. It was shown that the set of core reactions may change with the modification of the complex set. The examples also show that the frequently applied sparsity assumption alone is not enough for structural uniqueness of CRNs. Moreover, in the case of simple linear genetic network models, too sparse structures can degrade identifiability properties. The results further support the fact that as much prior information as possible should be incorporated in structural and parametric inference problems.

Competing interests

The authors declare that they have no competing interests.

Authors contributions

All authors contributed to the conception and design of the work. JRB and GS selected and evaluated the examples. GS performed the numerical computations. All authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was financially supported by project CAFE (Computer Aided Process for Food Engineering) FP7-KBBE-2007-1 (Grant no: 212754). The authors acknowledge the support by the Spanish government, MICINN project 'MultiSysBio' (ref. DPI2008-06880-C03-02), and by CSIC intramural project 'BioREDES' (ref. PIE-201170E018). GS acknowledges the partial support of the Hungarian Scientific Research Fund through grant no. K83440. The authors thank Dr. Irene Otero Muras (Dept. of Computational Systems Biology, ETH Zurich) for the fruitful discussions. The authors wish to thank the anonymous reviewers for their helpful comments.

Appendix

A.1 Proofs of P1, P2 and P3

Proof of **P1**. Let us denote the ith column of any matrix W by $W_{\cdot,i}$. The proof is based on the following well known fact of linear algebra. Consider an inhomogeneous set of linear equations:

$$Ax = b \tag{42}$$

If $x = p$ is any particular solution of (42) then the entire solution set for (42) can be characterized as

$$\{p + v \mid v \text{ is any solution of } Ax = 0\} \tag{43}$$

The matrix equation $Y \cdot A_k = M$ (see eqs. (3) and (8)) obviously defines m sets of linear equations of the form

$$Y \cdot [A_k]_{\cdot,i} = M_{\cdot,i}, \quad i = 1, \dots, m \tag{44}$$

Let us choose any i indexing the sets of equations in (44). For simplicity, let $p = [A_k]_{\cdot,i}$, $b = M_{\cdot,i}$. Let us assume that there are z elements of the constraint set (9) where $j_k = i$ for $k = 1, \dots, s$. (If z is 0, then we get the earlier result proved in [71].) These constraints can be expressed by further linear equations of the form:

$$[A_k]_{h,i} = 0, \quad h = 1, \dots, z \tag{45}$$

The equation sets (44) and (45) can be written into a single set of equations as

$$\bar{Y} \cdot p = \bar{b} \tag{46}$$

where $\bar{Y} \in \mathbb{R}^{(n+z)\times m}$ and $\bar{b} \in \mathbb{R}^{n+z}$. Let us assume now that p is a dense solution for (46), i.e. it contains the maximal possible number of nonzero elements. If p has no zero elements, then the result to be proved is trivially satisfied. Therefore, without the loss of generality we can assume that the first $l < m$ elements of p are nonzero, while the rest are zero, i.e. $p_j \neq 0$ for $j = 1, \ldots, l$, and $p_j = 0$ for $j = l+1, \ldots, m$. This can always be achieved by the appropriate reordering of the elements of p. Assume now that $p' \in \mathbb{R}^m$ is also a solution for (46), but $p'_c \neq 0$ for some $c \in \mathbb{Z}$, $l+1 \leq c \leq m$. Then $p' = p + v$, where $\bar{Y} \cdot v = 0$, and $v_c \neq 0$. In this case, $p'' = p + \lambda \cdot v$ is also a solution for (46) for any $\lambda \in \mathbb{R}$ and λ can always be chosen so that $p''_j \neq 0$ for $j = 1, \ldots, l$, and there is at least one index $l+1 \leq c \leq m$ for which $p''_c \neq 0$. However, this contradicts to the assumption that p is a dense solution for (46). □

Proof of **P2**. This is a straightforward consequence of **P1**, since the unweighted directed graphs of all constrained dense realizations must be identical. □

Proof of **P3**. If the graph structure of the constrained realization is unique, then it trivially implies that the structures of the constrained dense and sparse realizations are identical, since there exists only one possible constrained reaction structure. If the structures of the constrained dense and sparse realizations are identical, then the number of nonzero reaction rates is the same in any constrained realizations including the constrained dense ones. Then it follows from **P1** that the constrained reaction structure is unique. □

A.2 Details of the improved computation method to find dense realizations

The task of determining which reactions of a CRN belong to the dense realization can be effectively solved through the following problem set consisting of $m(m-1)$ LP computation steps, where m is the number of complexes in the CRN.

$$
\begin{aligned}
&\text{for each } \; p,q = 1,\ldots,m, \;\; p \neq q \text{ do:}\\
&\quad \text{maximize } f_{pq} = [A_k]_{p,q}\\
&\quad \text{subject to :}\\
&\qquad Y \cdot A_k = M,\\
&\qquad \sum_{i=1}^{m} [A_k]_{i,j} = 0, \quad j = 1,\ldots,m,\\
&\qquad 0 \leq [A_k]_{i,j} \leq U_{ij}, \quad i,j = 1,\ldots,m, \quad i \neq j,\\
&\qquad [A_k]_{i,i} \leq 0, \quad i = 1,\ldots,m,
\end{aligned}
\tag{47}
$$

where the decision variables are the off-diagonal entries of A_k, and U_{ij} are appropriately large positive upper bounds for $[A_k]_{i,j}$ to exclude the possibility of unbounded feasible solutions. The reaction $C_q \to C_p$ is in the dense realization if and only if the maximal objective function value for f_{pq} in (47) is positive. Let us denote the solution of (47) corresponding to (p, q), $p \neq q$ by $\bar{A}_k^{pq}$. Since the linear equality and inequality constraints in (47) are trivially convex, we will use the average of the obtained solutions $\bar{A}_k^{pq}$ as a lower bound to compute a possible dense realization in the final optimization step. For this, we define

$$\epsilon_{ij} = \left[\frac{1}{m(m-1)} \sum_{\substack{p,q=1 \\ p \neq q}}^{m} \bar{A}_k^{pq} \right]_{i,j}, \quad i \neq j. \tag{48}$$

By construction, $\epsilon_{ij} \geq 0 \ \forall i \neq j$, and $\epsilon_{ij} > 0$ if and only if the reaction $\mathcal{C}_j \to \mathcal{C}_i$ is in the dense realization. Then the actual dense realization can be determined by solving the following LP feasibility problem for A_k (with arbitrary linear objective function):

$$\begin{aligned}
&Y \cdot A_k = M, \\
&\sum_{i=1}^{m} [A_k]_{i,j} = 0, \quad j = 1, \dots, m, \\
&\epsilon_{ij} \leq [A_k]_{i,j} \leq U_{ij}, \quad i, j = 1, \dots, m, \quad i \neq j, \\
&[A_k]_{i,i} \leq 0, \quad i = 1, \dots, m.
\end{aligned} \tag{49}$$

It is important to remark that the definition of ϵ_{ij} in the form of (48) guarantees the solvability of (49). Naturally, the above described method can also be used for determining constrained dense realizations by adding constraints of the form (9) to the LP problems (47) and (49).

References

1. Wolkenhauer O: **Systems biology: The reincarnation of systems theory applied in biology?** *Briefings in Bioinformatics* 2001, **2**:258–270.

2. Stelling J: **Mathematical models in microbial systems biology**. *Current Opinion in Microbiology* 2004, **7**:513–518.

3. Kitano H: **Computational systems biology**. *Nature* 2002, **420**:206–210.

4. Csete M, Doyle J: **Reverse engineering of biological complexity**. *Science* 2002, **295**:1664–1669.

5. De Jong H: **Modeling and simulation of genetic regulatory systems: a literature review**. *Journal of Computational Biology* 2002, **9**:67–103.

6. Marbach D, Prill R, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G: **Revealing strengths and weaknesses of methods for gene network inference**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**:6286–6291.

7. Ay A, Arnosti D: **Mathematical modeling of gene expression: a guide for the perplexed biologist**. *Critical Reviews in Biochemistry and Molecular Biology* 2011, **46**:137–151.

8. Hecker M, Lambeck S, Toepfer S, Van Someren E, Guthke R: **Gene regulatory network inference: data integration in dynamic models – a review**. *Biosystems* 2009, **96**:86–103.

9. Bansal M, Belcastro V, Ambesi-Impiombato A, Di Bernardo D: **How to infer gene networks from expression profiles**. *Molecular Systems Biology* 2007, **3**:article number 78.

10. Tegner J, Yeung M, Hasty J, Collins J: **Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:5944–5949.

11. de la Fuente A, Brazhnik P, Mendes P: **Linking the genes: inferring quantitative gene networks from microarray data**. *Trends in Genetics* 2002, **18**:395–398.

12. Ronen M, Rosenberg R, Shraiman B, Alon U: **Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:10555–10560.

13. Thomas R, Paredes CJ, Mehrotra S, Hatzimanikatis V, Papoutsakis ET: **A model-based optimization framework for the inference of regulatory interactions using time-course DNA microarray expression data**. *BMC Bioinformatics* 2007, **8**:228–239.

14. Schaber J, Klipp E: **Model-based inference of biochemical parameters and dynamic properties of microbial signal transduction networks**. *Current Opinion in Biotechnology* 2010, **22**:109–116.

15. Li S, Assmann S, Albert R: **Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling**. *PLoS Biology* 2006, **4**:1732–1748.

16. Saez-Rodriguez J, Kremling A, Conzelmann H, Bettenbrock K, Gilles E: **Modular analysis of signal transduction networks**. *IEEE Control Systems Magazine* 2004, **24**:35–52.

17. Klamt S, Saez-Rodriguez J, Lindquist J, Simeoni L, Gilles E: **A methodology for the structural and functional analysis of signaling and regulatory networks**. *BMC Bioinformatics* 2006, **7**:56 (pp. 1–26).

18. Arkin A, Shen P, Ross J: **A test case of correlation metric construction of a reaction pathway from measurements**. *Science* 1997, **277**:1275–1279.

19. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles E: **Metabolic network structure determines key aspects of functionality and regulation**. *Nature* 2002, **420**:190–193.

20. Jeong H, Tombor B, Albert R, Oltvai Z, Barabási A: **The large-scale organization of metabolic networks**. *Nature* 2000, **407**:651–654.

21. Forster J, Famili I, Fu P, Palsson B, Nielsen J: **Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network**. *Genome Research* 2003, **13**:244.

22. Kell D: **Metabolomics and systems biology: making sense of the soup**. *Current Opinion in Microbiology* 2004, **7**:296–307.

23. Sauer U: **Metabolic networks in motion: 13C-based flux analysis**. *Molecular Systems Biology* 2006, **2**:article number 62.

24. Crampin E, Schnell S, McSharry P: **Mathematical and computational techniques to deduce complex biochemical reaction mechanisms**. *Progress in Biophysics and Molecular Biology* 2004, **86**:77–112.

25. Yuan Y, Stan GB, Warnick S, Goncalves J: **Robust dynamical network structure reconstruction**. *Automatica* 2011, **47**:1230–1235.

26. Stolovitzky G, Monroe D, Califano A: **Dialogue on Reverse-Engineering Assessment and Methods**. *Annals of the New York Academy of Sciences* 2007, **1115**:1–22.

27. Prill R, Marbach D, Saez-Rodriguez J, Sorger P, Alexopoulos L, Xue X, Clarke N, Altan-Bonnet G, Stolovitzky G: **Towards a rigorous assessment of systems biology models: the DREAM3 challenges**. *PLoS ONE* 2010, **5**:e9202 (pp. 1–18).

28. Horn F, Jackson R: **General mass action kinetics**. *Archive for Rational Mechanics and Analysis* 1972, **47**:81–116.

29. Feinberg M: **Chemical reaction network structure and the stability of complex isothermal reactors - I. The deficiency zero and deficiency one theorems**. *Chemical Engineering Science* 1987, **42**:2229–2268.

30. Érdi P, Tóth J: *Mathematical Models of Chemical Reactions. Theory and Applications of Deterministic and Stochastic Models.* Manchester, Princeton: Manchester University Press, Princeton University Press 1989.

31. Ljung L: *System Identification: Theory for the User, 2nd edition.* Prentice Hall 1999.

32. Walter E, Pronzato L: *Identification of Parametric Models.* Springer 1997.

33. Bellmann R, Aström KJ: **On structural identifiability**. *Mathematical Biosciences* 1970, **7**:329–339.

34. Walter E: *Identification of State Space Models.* Springer 1982.

35. Walter E: *Identifiability of Parametric models.* Pergamon Press, Oxford 1987.

36. Pohjanpalo H: **System identifiability based on the power series expansion of the solution**. *Mathematical Biosciences* 1978, **41**:21–33.

37. Walter E, Lecourtier Y: **Global approaches to identifiability testing for linear and nonlinear state space models**. *Mathematics and Computers in Simulation* 1982, **24**:472–482.

38. Vajda S, Godfrey K, Rabitz H: **Similarity transformation approach to identifiability analysis of nonlinear compartmental models**. *Mathematical Biosciences* 1989, **93**:217–248.

39. Diop S, Fliess M: **On nonlinear observability**. In *First European Control Conference, ECC'91*, Grenoble 1991:152–157.

40. Fliess M, Glad T: **An algebraic approach to linear and nonlinear control**. In *Essays on Control: Perspectives in the Theory and its Applications.* Edited by Treutelman HL, Willeuis JC, Boston: Birkhauser 1993:223–267.

41. Ljung L, Glad T: **On global identifiability of arbitrary model parametrizations**. *Automatica* 1994, **30**:265–276.

42. Margaria G, Riccomagno E, White LJ: **Structural identifiability analysis of some highly structured families of statespace models using differential algebra**. *Journal of Mathematical Biology* 2004, **49**:433–454.

43. Margaria G, Riccomagno E, Chappell MJ, Wynn HP: **Differential algebra methods for the study of the structural identifiability of rational function state-space models in the biosciences**. *Mathematical Biosciences* 2001, **174**:1–26.

44. Saccomani M, Audoly S, D'Angio L: **Parameter identifiability of nonlinear systems: the role of initial conditions**. *Automatica* 2003, **39**:619–632.

45. Yates J, Evans N, Chappell M: **Structural identifiability analysis via symmetries of differential equations**. *Automatica* 2009, **45**:2585–2591.

46. Walter E, Lecourtier Y, Happel J: **On the structural output distinguishability of parametric models, and its relations with structural identifiability**. *IEEE Transactions on Automatic Control* 1984, **AC-29**:56–57.

47. Walter E, Pronzato L: **On the identifiability and distinguishability of nonlinear parametric models**. *Mathematics and Computers in Simulation* 1996, **42**:125–134.

48. Evans ND, Chappell MJ, Chapman MJ, Godfrey KR: **Structural indistinguishability between uncontrolled (autonomous) nonlinear analytic systems**. *Automatica* 2004, **40**:1947–1953.

49. Pohjanpalo H, Wahlström B: **On the uniqueness of linear compartmental systems**. *International Journal of Systems Science* 1977, **8**:619–632.

50. Yates J, Jones R, Walker M, Cheung S: **Structural identifiability and indistinguishability of compartmental models**. *Expert Opinion on Drug Metabolism and Toxicology* 2009, **5**:295–302.

51. Godfrey K, Chapman M, Vajda S: **Identifiability and indistinguishability of nonlinear pharmacokinetic models**. *Journal of Pharmacokinetics and Pharmacodynamics* 1994, **22**:229–251.

52. Davidescu F, Jorgensen S: **Structural parameter identifiability analysis for dynamic reaction networks**. *Chemical Engineering Science* 2008, **63**:4754–4762.

53. Vajda S, Rabitz H: **Identifiability and distinguishability of general reaction systems**. *Journal of Physical Chemistry* 1994, **98**:5265–5271.

54. Vajda S, Rabitz H: **Identifiability and distinguishability of first-order reaction systems**. *Journal of Physical Chemistry* 1988, **92**:701–707.

55. Craciun G, Pantea C: **Identifiability of chemical reaction networks**. *Journal of Mathematical Chemistry* 2008, **44**:244–259.

56. Szederkényi G: **Comment on "Identifiability of chemical reaction networks" by G. Craciun and C. Pantea**. *Journal of Mathematical Chemistry* 2009, **45**:1172–1174.

57. Pohjanpalo H, Wahlström B: **Software for solving identification and identifiability problems, e.g. in compartmental systems**. *Mathematics and Computers in Simulation* 1982, **24**:490–493.

58. Raksányi A, Lecourtier Y, Walter E, Venot A: **Identifiability and distinguishability testing via computer algebra**. *Mathematical Biosciences* 1985, **77**:245–266.

59. Bellu G, Saccomani M, Audoly S, D'Angio L: **DAISY: A new software tool to test global identifiability of biological and physiological systems**. *Computer Methods and Programs in Biomedicine* 2007, **88**:52–61.

60. Saccomani M, Audoly S, Bellu G, D'Angio L: **Examples of testing global identifiability of biological and biomedical models with the DAISY software**. *Computers in Biology and Medicine* 2010, **40**:402–407.

61. Liao J, Boscolo R, Yang Y, Tran L, Sabatti C, Roychowdhury V: **Network component analysis: reconstruction of regulatory signals in biological systems**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:15522–15527.

62. Yue H, Brown M, Knowles J, Wang H, Broomhead D, Kell D: **Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: a case study of an NF-κB signalling pathway**. *Molecular Biosystems* 2006, **2**:640–649.

63. Zak D, Gonye G, Schwaber J, Doyle F: **Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network**. *Genome Research* 2003, **13**:2396–2405.

64. Chen WW, Niepel M, Sorger PK: **Classic and contemporary approaches to modeling biochemical reactions**. *Genes & Development* 2010, **24**:1861–1875.

65. Jaqaman K, Danuser G: **Linking data to models: data regression**. *Nature Reviews Molecular Cell Biology* 2006, **7**:813–819.

66. Banga JR, Balsa-Canto E: **Parameter estimation and optimal experimental design**. *Essays in Biochemistry* 2008, **45**:195–209.

67. Ashyraliyev M, Fomekong-Nanfack Y, Kaandorp J, Blom J: **Systems biology: parameter estimation for biochemical models**. *FEBS Journal* 2009, **276**:886–902.

68. Soliman S, Heiner M: **A unique transformation from ordinary differential equations to reaction networks**. *PLoS ONE* 2010, **5**:e14284 (pp. 1–6).

69. Werhli AV, Husmeier D: **Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge**. *Statistical Applications in Genetics and Molecular Biology* 2007, **6**:article no. 15.

70. Mukherjee S, Speed TP: **Network inference using informative priors**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**:14313–14318.

71. Szederkényi G, Hangos KM, Péni T: **Maximal and minimal realizations of reaction kinetic systems: computation and properties**. *MATCH Communications in Mathematical and in Computer Chemistry* 2011, **65**:309–332.

72. Szederkényi G: **Computing sparse and dense realizations of reaction kinetic systems**. *Journal of Mathematical Chemistry* 2009, **47**:551–568.

73. Szederkényi G, Hangos KM: **Finding complex balanced and detailed balanced realizations of chemical reaction networks**. *Journal of Mathematical Chemistry* 2011, **49**:1163–1179.

74. Johnston MD, Siegel D: **Linear conjugacy of chemical reaction networks**. *Journal of Mathematical Chemistry* 2011, **49**:1263–1282.

75. Wang Y, Zhang XS, Chen L: **Optimization meets systems biology**. *BMC Systems Biology* 2010, **4(Suppl 2)**:S1 (pp. 1–4).

76. Dantzig GB, Thapa MN: *Linear Programming 1: Introduction*. Springer-Verlag 1997.

77. Floudas C: *Nonlinear and Mixed-Integer Optimization*. Oxford University Press 1995.

78. Raman R, Grossmann IE: **Integration of logic and heuristic knowledge in MINLP optimization for process synthesis**. *Computers and Chemical Engineering* 1992, **16**:155–171.

79. Feinberg M: *Lectures on chemical reaction networks*. Notes of lectures given at the Mathematics Research Center, University of Wisconsin 1979. [http://www.che.eng.ohio-state.edu/~feinberg/LecturesOnReactionNetworks/].

80. Hárs V, Tóth J: **On the inverse problem of reaction kinetics**. In *Qualitative Theory of Differential Equations*, *Volume 30*. Edited by Farkas M, Hatvani L, North-Holland, Amsterdam 1981:363–379.

81. Hangos KM, Szederkényi G: **Mass action realizations of reaction kinetic system models on various time scales**. *Journal of Physics: Conference Series (5th International Workshop on Multi-Rate Processes and Hysteresis)* 2011, **268**:012009.

82. Jokar S, Pfetsch ME: **Exact and approximate sparse solutions of underdetermined linear equations**. *SIAM Journal on Scientific Computing* 2008, **31**:23–44.

83. Donoho DL: **For most large undetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution**. *Communications on Pure and Applied Mathematics* 2006, **59**:903–934.

84. Zavlanos MM, Julius AA, Boyd SP, Pappas GJ: **Inferring stable genetic networks from steady-state data**. *Automatica* 2011, **47**:1113–1122.

85. Szederkényi G, Hangos KM, Tuza Z: **Finding Weakly Reversible Realizations of Chemical Reaction Networks Using Optimization**. *MATCH Communications in Mathematical and in Computer Chemistry* 2012, **67**:193–212.

86. Kim H, Bond R: **Multicore software technologies: a survey**. *IEEE Signal Processing Magazine* 2009, **26**:80–89.

87. Mileyko Y, Joh RI, Weitz JS: **Small-scale copy number variation and large-scale changes in gene expression**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**:16659–16664.

88. Alon U: *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall / CRC Mathematical & Computational Biology 2007.

89. Ljung L: **Perspectives on system identification**. *Annual Reviews in Control* 2010, **34**:1–12.

90. Conradi C, Flockerzi D, Raisch J, Stelling J: **Subnetwork analysis reveals dynamic features of complex (bio)chemical networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:19175–19180.

91. Müller S, Hofbauer J, Endler L, Flamm C, Widder S, Schuster P: **A generalized model of the repressilator**. *Journal of Mathematical Biology* 2006, **53**:905–937.

92. Dilao R, Muraro D: **A software tool to model genetic regulatory networks. Applications to the modeling of threshold phenomena and of spatial patterning in drosophila**. *PLoS ONE* 2010, **5**:e10743 (pp. 1–10).

93. Bansal M, di Bernardo D: **Inference of gene networks from temporal gene expression profiles**. *IET Systems Biology* 2007, **1**:306–312.

94. Bang-Jensen J, Gutin G: *Digraphs: Theory, Algorithms and Applications*. Springer 2001.

Figure titles and captions

Figure 1 - Positive feedback motif: original reaction graph and dense realization structure

(a) This subfigure shows the reaction graph of a gene regulation network model with positive feedback described originally in [87] and used in Example 1. (b) This subfigure shows all the mathematically possible reactions that can result in the same dynamical behaviour as the original biologically meaningful network shown in Fig. 1. The core-reactions in the dense realization are shown with solid arrows, while the non-core reactions are indicated by dashed arrows.

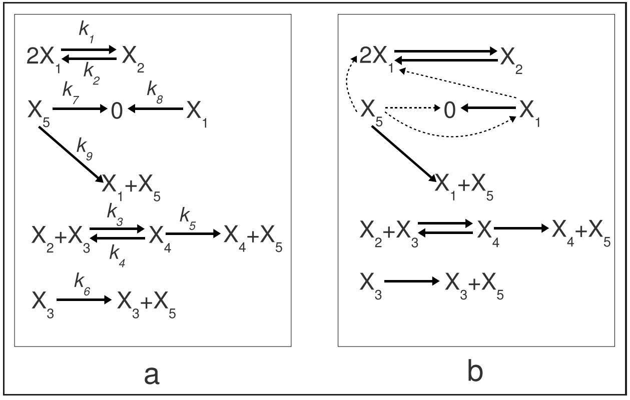


Figure 2 - Sparse realization structures for the positive feedback motif

Three different dynamically equivalent structures can be given for the positive feedback motif with the minimal number of reactions. The core and non-core reactions are indicated in the same way as in Fig. 1.b.

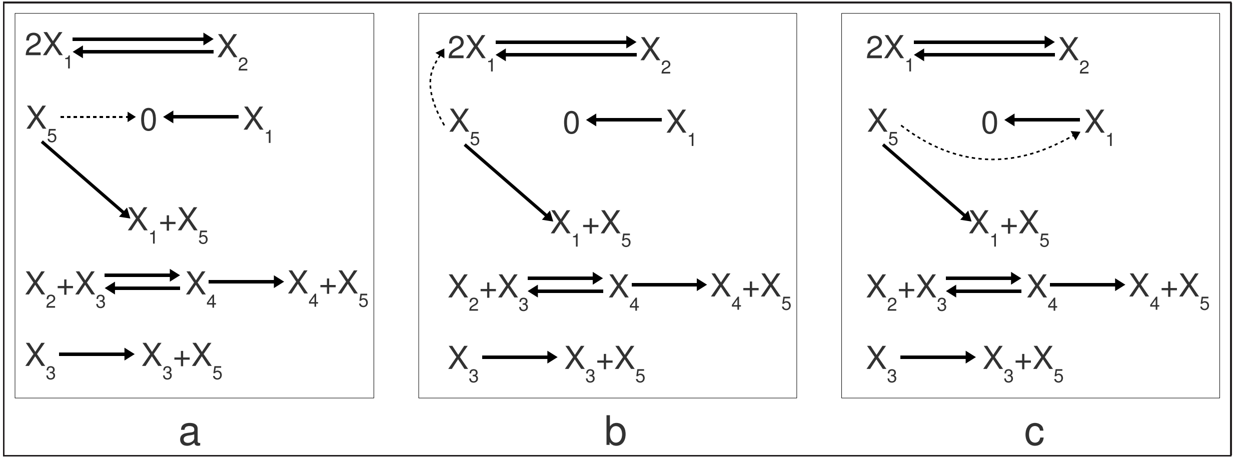


Figure 3 - The effect of modifying the complex set and the parameters

(a) The core and non-core reactions of the dense realization of the positive feedback motif are shown in this subfigure with a randomly selected parametrization that is different from the one given in [87]. (b) The core and non-core reactions of the dense realization of the positive feedback motif can be seen in this subfigure when an additional complex $X_2 + X_4$ is involved into the model.

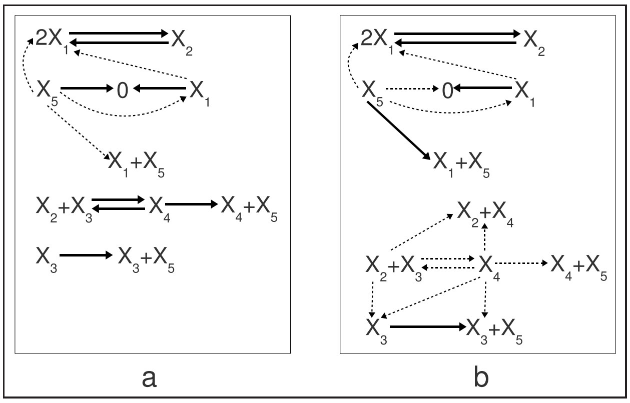


Figure 4 - Model of a biochemical switch in yeast cells

(a) The subfigure shows the original structure of a CRN describing a biochemical switch published in [90]. The numbering of species and rate coefficients is identical to the description in the original paper. (b) The dense realization of the network is depicted in this subfigure and contains 28 reactions, out of which only 12 belong to the set of core reactions.

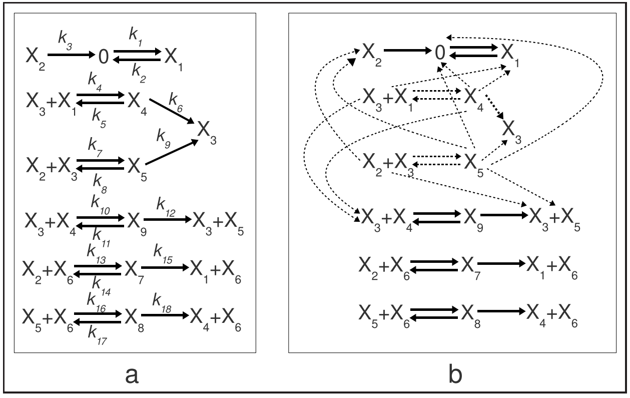


Figure 5 - A standard repressilator structure

A repressilator structure with 5 nodes and auto-activation is shown in the figure. The mass-action type CRN model of this structure contains 51 distinct complexes and 55 reactions.

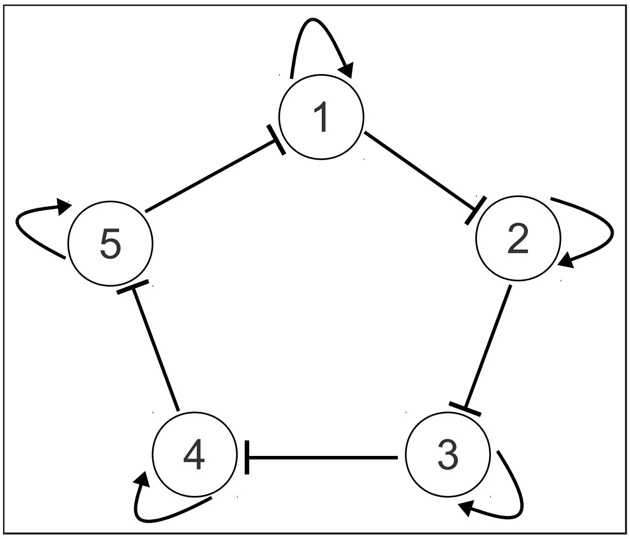
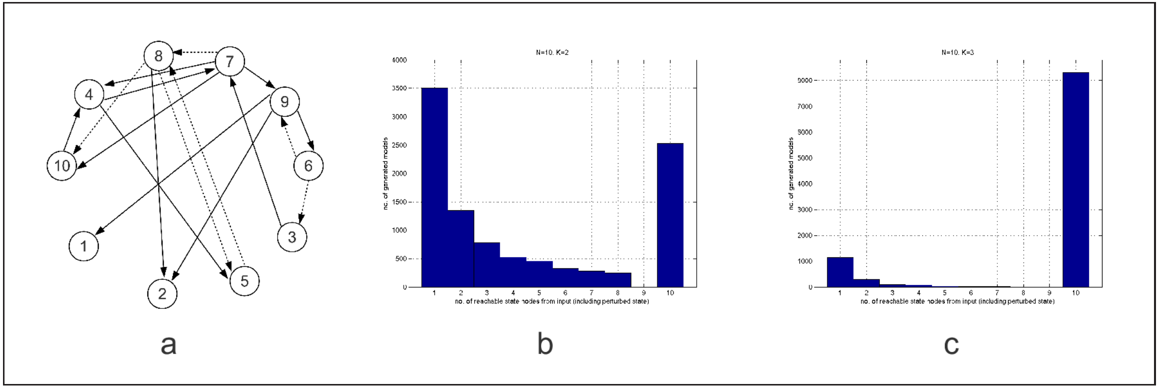


Figure 6 - A sparse gene regulation network and their structural identifiability properties

(a) This subfigure is the reproduction of one of the sparse gene regulation networks used for structural identification in [93]. The network has 11 activation (solid edges), 6 suppression (dashed edges), and 3 autoregulation links (at nodes 1, 3 and 6) with undefined sign. (b) The subfigure shows that the vast majority of the randomly generated models in the case when $N = 2$ and $K = 2$ are structurally unidentifiable, because not all nodes of the network are reachable from the perturbed one. (c) As the network becomes less sparse ($N = 10$, $K = 3$), the structural identifiability properties are quickly improving. In this case, more than 80% of the randomly generated models are structurally identifiable.



Additional Material

Additional file 1: CRN_data.xls

Excel file. **Detailed numerical data of the CRNs shown in Examples 1-3.** This file contains the detailed data (i.e. stoichiometric matrices and reaction rate coefficients) of the dynamically equivalent reaction networks studied in Examples 1,2 and 3. The individual sheets correspond to the different examples.