BMC Systems Biology

RESEARCH ARTICLE Open Access

Identification of regulatory structure and kinetic parameters of biochemical networks via mixed-integer dynamic optimization

Gonzalo Guillén-Gosálbez[1*], Antoni Miró[1], Rui Alves[2], Albert Sorribas[2] and Laureano Jiménez[2]

Abstract

Background: Recovering the network topology and associated kinetic parameter values from time-series data are central topics in systems biology. Nevertheless, methods that simultaneously do both are few and lack generality.

Results: Here, we present a rigorous approach for simultaneously estimating the parameters and regulatory topology of biochemical networks from time-series data. The parameter estimation task is formulated as a mixed-integer dynamic optimization problem with: (i) binary variables, used to model the existence of regulatory interactions and kinetic effects of metabolites in the network processes; and (ii) continuous variables, denoting metabolites concentrations and kinetic parameters values. The approach simultaneously optimizes the Akaike criterion, which captures the trade-off between complexity (measured by the number of parameters), and accuracy of the fitting. This simultaneous optimization mitigates a possible overfitting that could result from addition of spurious regulatory interactions.

Conclusion: The capabilities of our approach were tested in one benchmark problem. Our algorithm is able to identify a set of plausible network topologies with their associated parameters.

Keywords: Parameter estimation, Structure identification, Akaike criterion, Orthogonal collocation, Dynamic optimization, Biochemical networks

Background

Mathematical models of biochemical systems are becoming essential in systems biology to complement and extract information from time series. This information can be of two types. On the one hand, if the structure of the molecular circuit that executes the process of interest is known, models can be used to infer the numerical parameters that govern the dynamics of the system [1-4]. On the other, models can be used to infer the structure of the system from time series data (see for example [5-7]).

In either case, to obtain a useful model, we face different challenges: (i) defining the system's mass flow structure (*stoichiometry*), (ii), deciding the appropriate mathematical representation (*kinetics*), (iii) estimating the parameters that make the model response consistent with experimental data (*parameter estimation*), and (iv) inferring the system's regulatory structure. In addition, once the model is well defined, it should be able to predict systemic responses under yet untested experimental conditions (*model validation*).

The four challenges described in the previous paragraph are often addressed in independent steps. Current solutions to the first challenge are generally based on compiling information about the system and using that information to create the stoichiometric matrix for the system one wants to analyze (see for instance [8]). To solve the second challenge we need to define kinetic functions that describe the dynamic behavior of the dependent variables of the system. If the kinetic functions are unknown, approximate formalisms that have a solid theoretical support can be used to describe the dynamic behavior of the system within a given accuracy [9,10]. The third challenge is typically formulated as an optimization problem that minimizes the sum of squared residuals between the measured and simulated

* Correspondence: gonzalo.guillen@urv.cat
[1]Departament d'Enginyeria Química, Universitat Rovira i Virgili, Av.Països Catalans 26, 43007 Tarragona, Spain
Full list of author information is available at the end of the article

data (see a review of methods in [1]). The type of optimization problem being faced and the technical challenges to be solved depends upon the biological model of choice, upon the experimental data available, and upon the specific mathematical formalism used [11,12]. In many practical applications, the target biological system is described through nonlinear ordinary differential equations (ODEs). Hence, the parameter estimation task gives rise to dynamic optimization problems that are hard to solve. The fourth challenge could in principle be addressed in the same way as the first. However, despite the enormous amount of biological information available in public databases, regulatory signals are, in general, poorly understood and hardly ever properly characterized *in vivo*. Regulatory signals appear in a model as parameters accounting for the influence that metabolites others than the substrates of a reaction have on its velocity. Hence, parameter fitting can also be used to address the fourth challenge. However, the overwhelming majority of parameter estimation methods assumes a given structure and considers a fix regulatory scheme (see a review in [1]). This simplification is motivated by the difficulty in identifying regulatory effects, a task for which a myriad of alternative kinetic models must be explored [7,13-15].

Traditional methods for the selection of biological systems have mostly applied regression or chi-squared-based criteria (rather than information-theoretic fit criteria) [16]. However, information-theoretic criteria such as the Akaike's Information Criterion (AIC) [17] or the Bayesian Information Criterion (BIC) [18], are now perceived as important measures to assess quality of fit. AIC is often preferred over BIC becaue it has a more immediate connection to the theory of information [19]. AIC captures the trade-off between the complexity (measured by the number of parameters), and accuracy of the fitting. Smaller AIC values imply a better approximation to the model sought.

In this work we propose a strategy to simultaneously address the four challenges described above that relies on the use of mixed-integer dynamic optimization (MIDO) methods. Our approach adopts a structured mathematical framework to represent the kinetics of the processes that is flexible enough to reproduce a set of plausible network topologies (by implementing slight modifications on a basic model formulation). The power-law [20] and the saturable and cooperative formalisms are examples of such general kinetic representations [9]. Based on this type of general kinetic modeling framework, we develop our systematic parameter estimation method that provides as output a set of potential reaction and regulatory topologies for the target network along with the associated model parameters. We illustrate the capabilities of our approach using the GMA kinetic representation, a canonical model structure that uses the power-law kinetic formalism [21,22].

Results and discussion

As a proof-of-concept, we have tested the capabilities of our approach through its application to a case study taken from Voit and Almeida [23]. The system considered is a four-constituent pathway branched with six velocities and two regulatory signals. X_1 is generated from X_0, and its production is inhibited by X_3 which is produced from X_1 via intermediate X_2. X_1 yields also X_4, which promotes the degradation of X_3 (see Figure 1).

Parameter estimation when the regulatory structure is known

We shall first show that the proposed method is capable of appropriately identifying the model parameters using dynamic data when the regulatory structure is known. This is the classical parameter estimation problem that is solved in many applications. To this end, we first produce dynamic data without error from the reference system using a specific set of parameter values. Then, this *in silico* data is labeled as experimental and we use the proposed method to estimate the model parameters. We define a dynamic optimization model that contains a set of dynamic differential equations describing the system's kinetics. This dynamic model is reformulated into a nonlinear program (NLP) using orthogonal collocation on finite elements. This NLP does not contain binary variables because we assume that the regulatory signals are known. The aforementioned NLP was implemented in GAMS 23.7.3 and calculated with CONOPT 3.15A on a PC/AMD Athlon at 2.99 Ghz using a single core. The NLP features 302 variables and 285 constraints, and was solved in 2.3 CPU seconds. As expected, we obtain estimated parameters values that are very close to the original ones (see Table 1), and a least square error of 1.45×10^{-6}.

Non-linear kinetic models, like the GMA representation, have a certain degree of plasticity that allows different parameter sets to fit the same data. Clear parameter

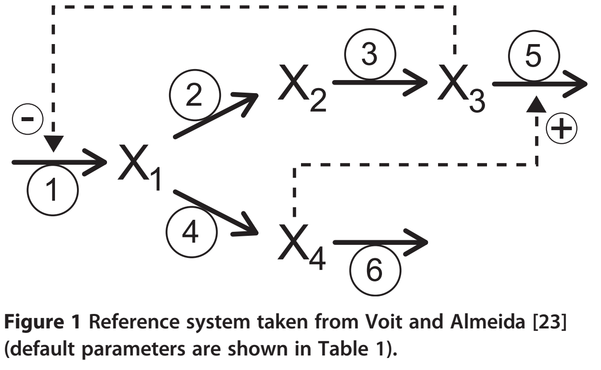

Figure 1 Reference system taken from Voit and Almeida [23] (default parameters are shown in Table 1).

Table 1 Original and predicted parameters values

Parameter	Original parameters	Proposed algorithm
f_{13}	−0.8	−0.7999
f_{21}	0.5	0.4996
f_{32}	0.75	0.7494
f_{41}	0.5	0.5006
f_{53}	0.5	0.4996
f_{54}	0.2	0.1996
f_{64}	0.8	0.8010
γ_1	12	12.000
γ_2	8	8.0031
γ_3	3	3.0034
γ_4	2	1.9965
γ_5	5	5.0014
γ_6	6	5.9967

Data is error free (one experiment with only one observation by time point).

trends are obtained by fixing a given parameter and fitting the remaining ones. As an example, Figure 2 shows the results of fixing f_{32} at different values and fitting the other parameters. All the points in the figure lead to residuals below 5.88×10^{-4}, indicating that it is possible to obtain good fits with different parameter sets. Similar patterns are obtained if we choose to fix any other parameter of the set.

As observed, the model is rather flexible, as there are many combinations of parameters values leading to very low residuals and essentially the same fit to the data. In practical terms, this means that given an experiment and an estimation procedure, we could obtain different parameter sets that closely reproduce the experimental measurements, but that differ from the actual values with which the dynamic profile was generated *in silico*. Thus, estimated parameter values don't help comparing the obtained fit with the reference model. In practice, the residual error and the resulting time profiles should be used to assess the fit.

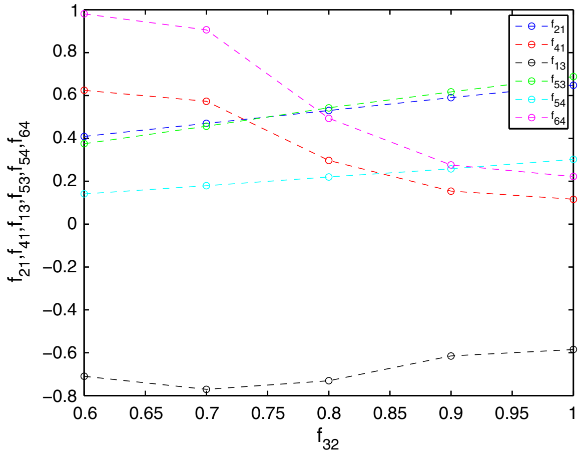

Figure 2 Values of the fitted parameters for different values of f32. Each point was generated by fixing f32 and solving the NLP free of error.

We will now consider the effect of noisy data on fitting the model, as such noise plays a key role in evaluating any proposed method for identifying the regulatory structure of a network. To explore the influence of random experimental uncertainty, we generated 100 dynamic profiles from the reference model by introducing statistical noise. For this, we applied Monte Carlo sampling assuming that every data point follows a normal distribution with standard deviation values of 0.5, 1, 5 and 10% of the actual nominal value. For comparison purposes, we use the same perturbation experiment as in the previous example. Table 2 shows the parameter values and the associated residuals obtained for four of the samples generated, while Figure 3 depicts the profiles associated with a standard deviation of 10%. We can appreciate that despite the different parameter values, the various fitted models lead to similar residuals. Note that although the regulatory structure is fixed, we obtain parameter values representing either positive or negative regulatory effects (f_{54}) of X_4 on v_5. This is a consequence of the "experimental error" introduced in the noisy data. That error may force the estimation procedure to an optimum involving a set of parameter values that may be different from the set that generates the noiseless data. In addition, as seen above, different parameters sets can be used to produce similar time courses. This means that there are coupled parameters in the system, which may also contribute for the estimation of regulatory interactions with reversed signals.

In general, even in simple cases as the one considered here, it will be difficult to obtain a consistent estimation from a single time-series. Identifying the parameter set that is more likely to be the correct one requires simultaneous fitting to additional time-series, resulting from more than one set of experiments. By doing so, we will constraint further the admissible parameter sets (see [24]). In Table 3, we show the results of fitting three different experiments with experimental error. Each experiment corresponds to an alternative perturbation on the initial concentration of metabolite X_3 (0.2, 1.2, and 2.2). These perturbations force the system to move across different dynamic regimes, producing additional information that helps in the identification of appropriate parameter values. As observed, the estimated parameters are more consistent over the various experiments. They are also closer to the actual parameter set selected for generating the data. Note, however, that it is still possible to find solutions involving alternative regulatory topologies with good fit to data (f_{54} acting as an inhibitor in Profile 2).

Table 2 Parameters values with noisy data (one experiment)

		10%		
	Profile 1	**Profile 2**	**Profile 3**	**Profile 4**
f_{13}	−0.14	−0.27	−0.84	−0.79
f_{21}	0.26	0.47	0.4	0.29
f_{32}	0.44	1	0.64	0.41
f_{41}	0.04	0	0.9	1
f_{53}	0	0.26	0.42	0.12
f_{54}	−0.06	0.04	0.1	−0.12
f_{64}	0.13	0.07	1	1
Residual	1.88	1.67	1.68	2.29
		5%		
	Profile 5	**Profile 6**	**Profile 7**	**Profile 8**
f_{13}	−0.282	−0.532	−0.631	−0.893
f_{21}	0.56	0.618	0.306	0.6
f_{32}	1	1	0.436	1
f_{41}	0	0.092	0.761	0.742
f_{53}	0.368	0.639	0.273	0.298
f_{54}	0.127	0.244	0.021	0.279
f_{64}	0.064	0.158	1	1
Residual	0.4128	0.4203	0.5706	0.4482
		1%		
	Profile 9	**Profile 10**	**Profile 11**	**Profile 12**
f_{13}	−0.881	−0.427	−0.859	−0.71
f_{21}	0.571	0.523	0.5	0.414
f_{32}	0.885	0.809	0.758	0.608
f_{41}	0.587	0.078	0.661	0.656
f_{53}	0.479	0.467	0.507	0.402
f_{54}	0.2	0.176	0.197	0.136
f_{64}	1	0.162	1	1
Residual	0.0207	0.0163	0.0167	0.0227
		0.5%		
	Profile 13	**Profile 14**	**Profile 15**	**Profile 16**
f_{13}	−0.845	−0.744	−0.843	−0.765
f_{21}	0.535	0.472	0.496	0.453
f_{32}	0.816	0.714	0.749	0.673
f_{41}	0.556	0.492	0.647	0.643
f_{53}	0.492	0.439	0.497	0.443
f_{54}	0.201	0.167	0.196	0.164
f_{64}	0.916	0.816	1	1
Residual	0.0052	0.0041	0.0042	0.0057

We solved a total of 100 problems, each corresponding to a different replication, generated randomly see Additional file 1: Table S1). The table shows the 16 cases for which the residual error is low.

Identifying the regulatory structure

Performance using error free data

After testing the capabilities of the method when the structure is known, we studied its ability to identify the regulatory topology of the model. To this end, we explore the performance of the method using one experiment with low experimental error (i.e., assuming that the data follow normal distributions with a standard deviation of 0.5%). Larger errors result in a wider set of alternative structures and for simplicity's sake we shall not discuss them here.

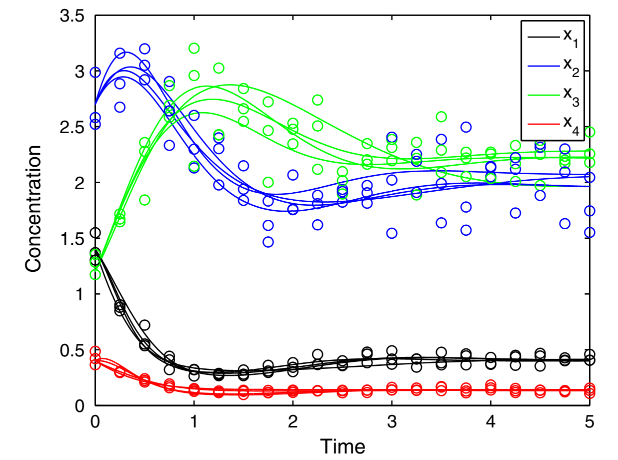

Figure 3 Adjusted profiles for four different noisy data sets (i.e. one experimental condition and four replications) with a standard deviation of 10%.

In order to simplify the search, we fix a maximum of two metabolites (the substrate of the reaction, which is given by the stoichiometric information, and one possible additional modifier, which is not *a priori* characterized) as potential variables affecting each velocity.

We note that it is typical to have some *a priori* knowledge about the biological system one is interested in. The complexity of the regulatory interactions in the identification problem is reduced if such knowledge can be used to constrain further both, the number of potential regulatory signals in the model and their signs (positive, negative). In such cases, we can introduce specific

Table 3 Parameter values obtained from simulated noisy data (with noisy data (three experiments))

	Profile 1	**Profile 2**	**Profile 3**	**Profile 4**
f_{13}	−0.67	−0.64	−0.62	−0.92
f_{21}	0.33	0.9	0.49	0.69
f_{32}	0.42	1	0.73	1
f_{41}	0.64	0	0.38	0.26
f_{53}	0.49	0.66	0.3	0.4
f_{54}	0.05	−0.95	0.22	0.34
f_{64}	1	1	0.53	0.58
Residual	6.96	7.10	5.39	4.89

We solved a total of 100 problems, generated randomly. See Additional file 1: Table S2.

constraints for the relevant parameters to be fitted. For example, in our case kinetic-order corresponding to the substrates of a reaction must be positive.

The MINLP model that simultaneously fits the parameters and infers probable regulatory interactions was implemented in GAMS 23.7.3 and solved with the solver SBB in the same computer as before. The model has 72 binary variables, 391 continuous variables and 414 equations. The solution time was in the order of few minutes for each simulation.

Our algorithm identifies a set of compatible systems, since the model has enough flexibility to play with the regulatory structure as well as the kinetic parameters when minimizing the residuals. The method identifies topologies that are quite close and that show very small residuals, but it is unable to uniquely identify the original topology (see Additional file 1: Table S3 for a list of topologies generated and their associated kinetic parameters and residuals). As an example, in Figure 4, we compare three completely different regulatory structures that produce almost indistinguishable results and similar fitting to the actual dynamics, leading to residual values of 0.00223, 0.00283 and 0.00316 (Figure 5).

As before, one strategy for increasing the possibility of correctly identifying the “true” regulatory structure is to use additional time-series data of the same system under different sets of initial conditions. To this end, we changed the initial concentrations of X_3 (0.2, 1.2, and 2.2). The MINLP model was again implemented in GAMS and solved with SBB in the same computer. In this case, the MINLP features 72 binary variables, 967 continuous variables and 980 equations. The solution time was in the order of few minutes for each simulation.

In Figure 6 we show the dynamic profiles associated with three different topologies identified by the MINLP. A complete list of network topologies and associated kinetic parameters and residuals is provided as (Additional file 1: Table S4). With three time series, the method identifies not only the actual topology, but also several structures that contain the original one (i.e., topologies that account for all the actual regulatory effects plus other signals that were not present originally). Again, we obtained slightly different parameter sets in each case, since the model flexibility is rather large.

Additional remarks

The use of MIDO techniques combined with orthogonal collocation allows posing the parameter estimation task

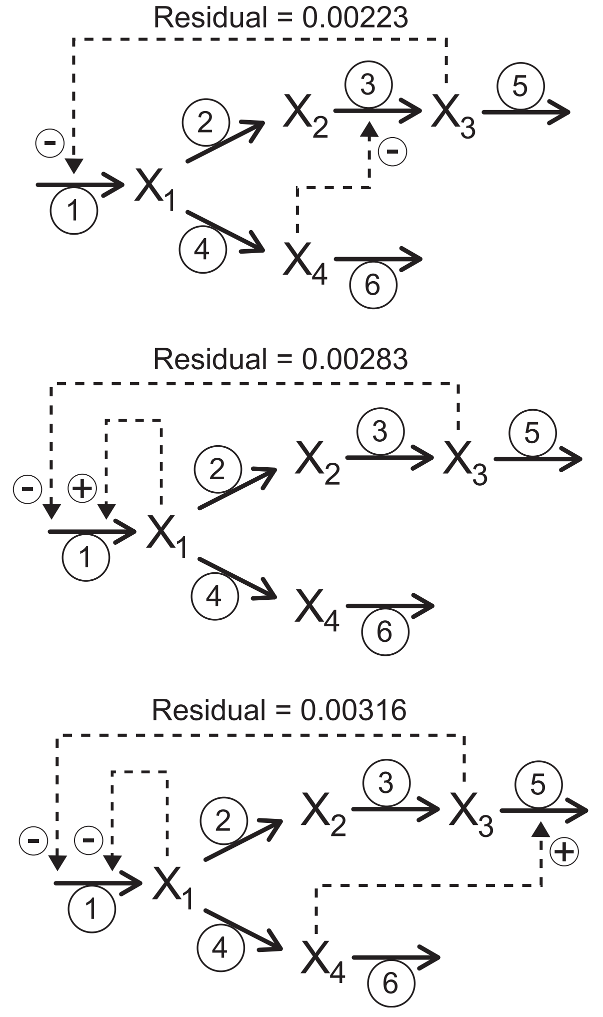

Figure 4 The proposed method identifies different regulatory topologies that essentially produce the same output. We show here the associated profiles corresponding to three regulatory structures with lowest residual values obtained by analyzing data from a single experiment with one replicate (see parameters values and residuals in Additional file 1: Table S3.

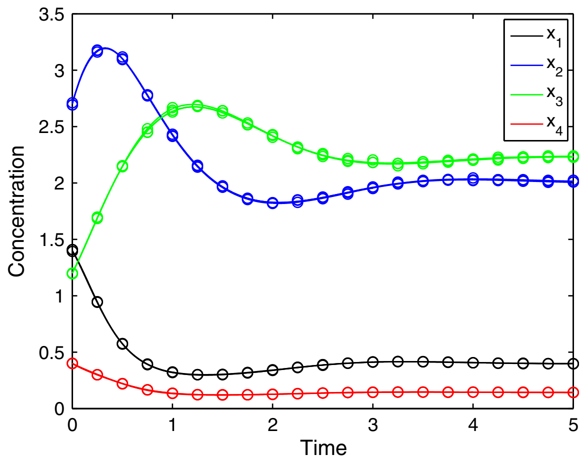

Figure 5 Dynamic responses corresponding to the three different topologies of Figure 4. Parameter values are indicated on Additional file 1: Table S3.

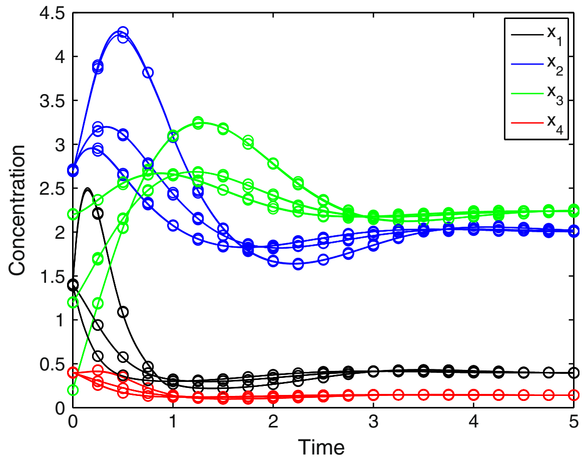

Figure 6 The Profiles generated from three different topologies and three experiments with one replication each. The experiments are generated from the base case by applying different perturbations in the initial concentration of X_3. Details on the topology and associated parameters are provided on Additional file 1: Table S4.

as an algebraic optimization problem that can be efficiently solved using standard MINLP algorithms. Orthogonal collocation shows some appealing properties (see [25]), but has the drawback of increasing the model size because it adds auxiliary variables and equations that increase the problem complexity. Our MIDO approach, however, can be solved by any MIDO algorithm, and it is not restricted to the use of orthogonal collocation and MINLP reformulations.

A key point in our method is the selection of an appropriate starting point to initialize the MINLP algorithm. Standard MINLP algorithms typically solve an initial NLP where the binary variables are relaxed. If this NLP does not converge, the entire algorithm might fail. An initialization strategy that works well in practice is to integrate first the original kinetic model for some parameter values, and then use the dynamic profiles generated *in silico* to provide a starting point for the NLP solver. Another method consists of solving an auxiliary model where we relax some constraints through the addition of slack variables, and then minimize the summation of the slacks in order to obtain an initial feasible point. With this relaxed model, we can identify a feasible (but not necessarily optimal) solution for the initial NLP.

Even if the MINLP model converges, there is still the issue of getting trapped in local optima during the search. To avoid this, we can run the optimization algorithm from different starting points generated randomly. This strategy does not guarantee convergence to the global optimum, but tends to produce high quality solutions in short CPU times. In contrast, deterministic global optimization methods provide a rigorous interval within which the optimum should fall, but tend to lead to large CPU times (see [26,27]).

In our case, we initialize the NLPs by solving a set of relaxed problems from different starting points and then pass these results to the first NLP solver. This approach provides feasible points from which the model converges to solutions with low residuals.

In general, due to the nonconvex nature of the reformulated MINLP, the nonlinear branch and bound implemented in SBB outperforms the outer-approximation used by DICOPT. This is because the supporting hyperplanes defined in the master MILP solved by DICOPT may chop-off feasible solutions due to the noconvex nature of some nonlinear inequalities.

We note that nonlinear models are hard to handle, and even more so when they contain binary variables. Standard NLP solvers can solve problems containing up to hundreds of thousands of variables and constraints. On the other hand, the computational burden of MIDO (and MINLP) models is rather sensitive to the number of binary variables. For the type of problems we are dealing with, it is difficult to provide a bound on the number of binaries above which the algorithm might fail. In practice, however, we found that this approach efficiently for less than one hundred binaries (around 30 parameters).

From a practical viewpoint, we face the challenging problem of discriminating between compatible regulatory structures for a given data set. On a worst case scenario, our method provides a ranked set of alternative regulatory topologies that can be tested and validated experimentally. If appropriate additional time-series data are available, the set of admissible solutions for testing can be further constrained and reduced. Our method finds a set of alternatives that are consistent with the dynamic data available and that can be further refined using additional information and expert knowledge on the system. (i.e., complementary biological information). For instance, kinetic-orders that correspond to substrates of a reaction may be safely restricted to be positive. Similarly, if we are fairly sure that a given metabolite does not participate in a reaction, its kinetic-order should be fixed to zero.

Our method can also be used to explore hypotheses about the regulatory structure of a system. For instance, we can force some parameters to take negative values, thereby representing inhibition effects, and then perform the optimization so as to determine if the fitting is good enough. Furthermore, we can follow the same procedure in order to identify regulatory effects that are consistent with this hypothesis.

In addition, we note that our approach can be easily adapted in order to work with other model selection criteria besides AIC. We remark, however, that the assessment of different selection criteria would deserve a comprehensive study that is beyond the scope of this work.

The simple examples presented in this paper show that estimating parameters in dynamic kinetic models is far from being an easy job and that models based on the power-law formalism facilitate the estimation task. Although this formalism is suitable for a wide variety of problems, one may argue that it may present some limitations. As an alternative, we can use extensions of this framework such as the Saturable and Cooperative formalism [9], which takes into account saturation effects. In both cases, a key point is the possibility of using a canonical mathematical formalism that facilitates the automatic search of alternative regulatory patterns. The method described here would be applicable to such models via recasting of the Saturating and Cooperative formalism into a power law [28].

Conclusions

In this work we have proposed a rigorous approach based on mathematical programming for the simultaneous identification of the regulatory signals and estimation of the kinetic parameters of models of biochemical networks. Our approach is based on the use of mixed-integer dynamic optimization (MIDO) models that minimize the Akaike criterion, and that can be solved by standard optimization algorithms. Particularly, we solve this MIDO by reformulating it as a mixed-integer nonlinear program (MINLP) using orthogonal collocation on finite elements, which makes it possible to apply standard MINLP solution algorithms in an iterative fashion in order to identify a set of plausible network topologies and associated kinetic parameters.

It is noteworthy that the difficult task of parameter estimation in nonlinear models becomes really complicated as the size of the models increases. Therefore, such estimation typically requires customized solution procedures. One key point is to use the appropriate initial conditions to ensure convergence of the calculations.

The proposed method can contribute to fill the lack of information on the regulatory signals that are in play in a given metabolic scenario. Although we cannot deal with genome-wide models, we have shown that dynamic profiles can be processed to provide clear hypothesis on the underlying regulatory structure. This is an important step towards completing essential information on different metabolic processes that are poorly understood.

Methods

The problem we address here is to infer the regulatory structure of a metabolic system, given a known structure for the reaction network (stoichiometry) and experimental time series for the dynamic behavior of that system. To address this question, and to explore the practical problems associated, we consider the following general representation of a biochemical network:

$$\dot{X}_i = \sum_{r=1}^{p} \mu_{i,r} v_r \quad i = 1, \ldots, n \tag{1}$$

where X_i denotes the concentration of metabolite i, $\mu_{i,r}$ is the stoichiometric coefficient of metabolite i in process r, which indicates the number of molecules of type i produced or destroyed by process r, and v_r is the rate function of this process. In general, v_r is represented as:

$$v_r(X_1, \ldots, \ X_{n+m}, \ \theta) \tag{2}$$

There are two critical issues in defining this model. One is the selection of an appropriate mathematical representation for v_r, which may be a function of an arbitrary number of variables (substrates, products, and modifiers). In most cases the mechanism for each process are unknown and choosing a specific mechanistic rate law, such as a Michaelis-Menten rate law, becomes an act of faith. The other issue is the problem of identifying the regulatory structure of the system.

The most straightforward and theoretically well supported solution to both issues is the use of an approximate formalism based on a standard mathematical representation [10]. By adopting such a kinetic representation, identifying the regulatory structure of the system becomes synonymous to determining the set of values θ for the model parameters that better fit the available data. Hence, without losing generality, and as a first step towards a more complex framework, we will consider the case where the rates are modeled using a power-law formalism. Note, however, that our approach could be easily extended in order to accommodate any other structured kinetic formalism.

Power-law models

Using the power-law representation, the rate v_r is expressed as follows:

$$v_r = \ \gamma_r \prod_{j/1}^{n+m} X_j^{f_{r,j}} \ \ r \ \ = \ \ 1, \ldots, \ \ p \tag{3}$$

where γ_r is an apparent rate constant for reaction r, and $f_{r,j}$ is the kinetic order of metabolite j in that process. Note that this equation accounts for the effect of $n + m$ metabolites (n dependent and m independent) on each reaction.

The advantage of this representation is that the same functional form represents all the rates. The reaction structure of the system will constrain the range of admissible values for some of the parameters. For example, all γ and f parameters for the substrates and catalysts of the reactions are by definition larger than zero. In addition, the values of the *f* parameters for all metabolites that are not directly involved in a given process are zero in the rate that describes the process.

By adopting such a kinetic representation, we can pose the problem of identifying the regulatory signals in a very compact mathematical form. If X_j is a modifier of v_r, then the corresponding kinetic order $f_{r,j}$ will be different from zero (positive if it is an activator, and negative if it is an inhibitor). By substituting (3) into equation (1), we get what is known as a Generalized Mass-Action (GMA) model.

$$\dot{X}_i = \sum_{r=1}^{p}\left(\mu_{i,r}\ \gamma_r \prod_{j=1}^{n+m} X_j^{f_{r,j}}\right)\ \ i = 1, \dots,\ \ n \tag{4}$$

Note that the power-law formalism accounts for both the stoichiometry of the system (*the network structure*), and the reaction and regulatory structures (*kinetic orders*) using a single systematic nonlinear representation. This property is very important for defining a systematic way of exploring alternative regulatory signals. We will make use of this general and compact formalism in the derivation of the equations for the parameter estimation model.

Parameter estimation in a GMA model

Given a set of experimental observations (i.e., time courses for the metabolites), our goal is to find the values of the apparent constants and kinetic orders that minimize the sum of least squared errors between the experimental data and the predicted dynamic profiles. This problem can be expressed in compact form as follows:

$$\begin{aligned}
\min_{\gamma_r, f_{r,j}} \quad & \sum_{u=1}^{k}\sum_{i=1}^{n}\left(X_{i,u}^{\text{exp}} - X_{i,u}^{\text{mod}}\right)^2 \\
s.t. \quad & \dot{X}_j = \sum_{r-1}^{p} \mu_{i,r} v_r \quad i = 1, \dots, n \\
& v_r = \gamma_r \prod_{j=1}^{n+m} X_j^{f_{r,j}} \ \ r = 1, \dots, p \\
& \dot{X}_i(t_0) = X_{0i} \qquad i = 1, \dots, n; t\ \in \left[t_0, t_f\right] \\
& X_{i,u}^{\text{mod}} = X_i(t_u) \qquad i = 1, \dots, n; u\ \ = 1, \dots, k
\end{aligned} \tag{5}$$

where X_i represents the state variables (i.e., metabolite concentrations), X_{0i} their initial conditions, $X_{i,u}^{exp}$ denotes the experimental observations, and $X_{i,u}^{mod}$ are the values calculated by the dynamic model (i.e., model predictions). i is the index for the set of state variables whose derivatives explicitly appear in the model, γ_r and $f_{r,j}$ are the parameters to be estimated, and t_u, is the time associated with experimental point u belonging to the set U of observations. k is the total number of experimental data points and n is the number of time dependent variables.

Conventional parameter estimation approaches seek parameter values that minimize the approximation error assuming a given regulatory scheme (i.e., fixing some $f_{r,j}$ to zero beforehand according to the aprioristic biochemical knowledge of the system). While this assumption simplifies the calculations, it can lead to poor approximations and hamper at the same time the discovery of new regulatory loops. In this work we introduce a rigorous and systematic parameter estimation and network identification method that makes no assumption regarding the regulatory network topology.

To model the existence of a regulatory interaction, we make use of the following disjunction:

$$\begin{gathered}
\begin{bmatrix} Y_{r,j}^{-} \\ f_{r,j}\ \le\ -\varepsilon \end{bmatrix} \ \underline{\vee} \begin{bmatrix} Y_{r,j} \\ -\varepsilon \le f_{r,j}\ \le\ \varepsilon \end{bmatrix} \underline{\vee} \begin{bmatrix} Y_{r,j}^{+} \\ \varepsilon \le f_{r,j} \end{bmatrix}\ j = 1, \dots, n, \\
r = 1, \dots, p \\
Y_{r,j}^{-}, Y_{r,j}, Y_{r,j}^{+}\ \in \{True,\ \ False\}
\end{gathered} \tag{6}$$

In which $Y_{r,j}^{-}$,$Y_{r,j}$ and $Y_{r,j}^{+}$ are Boolean variables that are true if parameter $f_{r,j}$ is negative, zero or positive, respectively, and false otherwise. ε is a very small parameter. Note that only one term of the disjunction can be active (i.e., exclusive disjunction), while the others must be false. Hence, if $Y_{r,j}$ is true, metabolite i takes no part in velocity r. Conversely, if this metabolite has an influence on r, then $Y_{r,j}$ is false and either $Y_{r,j}^{-}$ or $Y_{r,j}^{+}$ will be active. This disjunction can be translated into standard algebraic equations using either the big-M or convex-hull reformulations [29]. By applying the former, we get:

$$\begin{aligned}
& f_{r,j}\ \le\ -\varepsilon + M\left(1 - y_{r,j}^{-}\ \right) && j = 1, \dots,\ \ n, r = 1, \dots, p \\
& -\varepsilon - M\left(1 - y_{r,j}\right)\ \le f_{r,j}\ \le \varepsilon + M\left(1 - y_{r,j}\right) && j = 1, \dots,\ \ n, r = 1, \dots, p \\
& f_{r,j}\ \le\ \varepsilon + M\left(1 - y_{r,j}^{+}\ \right) && j = 1, \dots,\ \ n, r = 1, \dots, p \\
& y_{r,j}^{-}\ \ + y_{r,j}\ \ + y_{r,j}^{+}\ \ = 1 && j = 1, \dots,\ \ n, r = 1, \dots, p \\
& y_{r,j}^{-}\ \ + y_{r,j}\ \ + y_{r,j}^{+}\ \ \in\ \{0, 1\}
\end{aligned} \tag{7}$$

where Boolean variables Y have been replaced by auxiliary binary variables y. In these equations, M is a sufficiently large parameter whose value must be carefully set according to the bounds defined for the kinetic parameters.

A key issue in our approach is how to avoid overfitting. To this end, we make use of the Akaike criterion, which captures the trade-off between the number of kinetic parameters contained in the model and its ability to accurately reproduce the experimental data. If we assume that the error of the observations follows a normal distribution, the Akaike criterion takes the following mathematical form [17]:

$$AIC = \ k \log\left(\frac{\sum_{u=1}^{k}\sum_{i=1}^{n}\left(X_{i,u}^{\exp} - \ X_{i,u}^{\text{mod}}\right)^2}{k}\right) + \ 2\sum_{j=1}^{n}\sum_{r=1}^{p}\left(y_{r,j}^{-} + \ y_{r,j}^{+}\right) + \ C \tag{8}$$

Where *AIC* denotes the value of the Akaike criterion and *C* is a constant value that does not affect the optimization. The parameter estimation problem can be finally posed in mathematical terms using the following MIDO (mixed-integer dynamic optimization) formulation:

$$\begin{aligned}
&(\text{M}) \quad \min_{\gamma_r, f_{r,j}, y_{r,j}^{-}, y_{r,j}, y_{r,j}^{+}} \quad k\log\left(\frac{\sum_{u=1}^{k}\sum_{i=1}^{n}\left(\hat{X}_{i,u} - \ \bar{X}_{i,u}\right)^2}{k}\right) + 2\sum_{j=1}^{n}\sum_{r=1}^{p}\left(y_{r,j}^{-} + y_{r,j}^{+}\right) \\
&s.t. \qquad \dot{X}_j = \sum_{r-1}^{p}\mu_{i,r}v_r \quad i = 1,\ldots,n \\
&\qquad\qquad v_r = \gamma_r\prod_{j=1}^{n+m}X_j^{f_{r,j}} \quad r = 1,\ldots,p \\
&\qquad\qquad \dot{X}_i(t_0) = X_{0i} \qquad i = 1,\ldots,n;t \ \in\left[t_0, t_f\right] \\
&\qquad\qquad \bar{X}_{i,u} = X_i(t_u) \qquad i = 1,\ldots,n;u \ \ = 1,\ldots,k \\
&f_{r,j} \ \le \ -\varepsilon + M\left(1 - y_{r,j}^{-}\ \right) \qquad j = 1,\ldots, \ \ n, r = 1,\ldots,p \\
&-\varepsilon - M\left(1 - y_{r,y}\right) \le f_{r,j} \ \le \varepsilon + M\left(1 - y_{r,j}\right) \quad j = 1,\ldots, \ \ n, r = 1,\ldots,p \\
&f_{r,j} \ \le \ \varepsilon + M\left(1 - y_{r,j}^{+}\ \right) \qquad j = 1,\ldots, \ \ n, r = 1,\ldots,p \\
&y_{r,j}^{-} \ + y_{r,y} + y_{r,j}^{+} \ \ = 1 \qquad j = 1,\ldots, \ \ n, r = 1,\ldots,p \\
&y_{r,j}^{-} \ + y_{r,j} \ \ + y_{r,j}^{+} \ \ \in \ \{0, 1\}
\end{aligned} \tag{9}$$

There are different solution methods to solve this MIDO (see [25]). Without loss of generality, we propose here to reformulate this problem into an equivalent algebraic MINLP (mixed-integer nonlinear program) using orthogonal collocation on finite elements. This allows exploiting the rich optimization theory and software applications available for MINLP in the solution of the MIDO. Note that the reformulated MINLP might be nonconvex. This will give rise to multimodality (i.e., existence of multiple local optima), preventing standard gradient-based solvers from identifying the global optimum. Deterministic global optimization methods could be applied to solve the MINLP, but they might lead to large CPU times given the size and complexity of a standard dynamic problem of this type. Details on the application of deterministic global optimization methods to parameter estimation problems of small/medium size can be found elsewhere [30,31]. For the reasons given above, in this work we will solve the reformulated MINLP using local optimizers.

One important feature of our approach is that rather than calculating a single optimal solution, it identifies a set of plausible regulatory topologies by solving the model iteratively. That is, the model is first solved to identify a potential regulatory configuration represented by a binary solution (i.e., set of values of the binary variables). The model is then calculated again but this time adding the following integer cut, which excludes solutions identified so far in previous iterations from the search space:

$$\begin{aligned}
&\sum_{(r,j)\in ONE_{it}^{-}} {y_{r,j}^{-}}^{it} + \sum_{(r,j)\in ONE_{it}} {y_{r,j}}^{it} + \sum_{(r,j)\in ONE_{it}^{+}} {y_{r,j}^{+}}^{it} \\
&\quad - \sum_{(r,j)\in ONE_{it}^{-}} {y_{r,j}^{-}}^{it} - \sum_{(r,j)\in ONE_{it}} {y_{r,j}}^{it} - \sum_{(r,j)\in ONE_{it}^{+}} {y_{r,t}^{+}}^{it} \\
&\quad \le \ \left| \ ONE_{it}^{-} + ONE_{it} + ONE_{it}^{+} \ \right| - \ 1 \\
&ONE_{it}^{-} \ \ = \{(r,j)|{y_{r,t}^{-}}^{it} = 1 \text{ in the solution obtained in the iteration } it \ \} \\
&ONE_{it} \ \ = \{(r,j)|{y_{r,j}}^{it} = 1 \text{ in the solution obtained in the iteration } it \ \} \\
&ONE_{it}^{+} \ \ = \{(r,j)|{y_{r,j}^{+}}^{it} = 1 \text{ in the solution obtained in the iteration } it \ \} \\
&ZERO_{it}^{-} \ \ = \{(r,j)|{y_{r,j}^{-}}^{it} = 0 \text{ in the solution obtained in the iteration } it \ \} \\
&ZERO_{it} \ \ = \{(r,j)|{y_{r,j}}^{it} = 0 \text{ in the solution obtained in the iteration } it \ \} \\
&ZERO_{it}^{+} \ \ = \{(r,j)|{y_{r,j}^{+}}^{it} = 0 \text{ in the solution obtained in the iteration } it \ \}
\end{aligned} \tag{10}$$

Where ONE_{it} and $ZERO_{it}$ represent the sets of binary variables that take a value of one and zero, respectively, in iteration it of the algorithm. After adding the integer cut, the model is solved again to produce a new regulatory topology, and this procedure is repeated iteratively until a desired number of configurations is generated. Hence, the algorithm produces as output a set of potential network configurations (encoded in the values of the binary solutions) rather than a single topology. Note that these regulatory topologies show a descendant value of the Akaike performance criterion.

Additional file

Additional file 1: Table S1. Parameters values obtained from simulated experiments with noisy data and known regulatory structure. We generate 100 different datasets by adding random noise using a normal distribution with a standard deviation of 10%. **Table S2.** Parameter values for three experiments with noisy data and known regulatory structure (we considered three experiments and solved a total of 100 problems, replications, generated randomly with a normal distribution with a standard deviation of 10%. **Table S3.** Kinetic parameters, Akaike values and residuals corresponding to the regulatory topologies obtained by fitting an *'in silico'* experiment generated from the reference model with added noise (normal distribution with a standard deviation of 0.5% of the actual concentration value). We show the ten best cases sorted by residual value. In yellow we indicate kinetic orders that must be greater than zero as they represent effects of the substrate of the considered reaction. In green, we indicate the regulatory effects that were included in the reference model. In light red, we indicate regulatory effects that are not present in the reference model. **Table S4.** Kinetic parameters, Akaike values and residuals corresponding to the regulatory topologies obtained by fitting three *'in silico'* experiment generated from the reference model with added noise (normal distribution with a standard deviation of 0.5% of the actual concentration value). The experiments are generated from the base case by applying different perturbations in the initial concentration of X_3. We show the ten best cases sorted by residual value. See color meaning in **Table S3**.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GGG suggested the need for the approach and provided the biological problems. AM, GGG, RA, AS and LJ developed the optimization algorithms and performed the numerical analysis. All authors evaluated the results, wrote the paper and contributed to its final form. All authors read and approved the final manuscript.

Acknowledgements

Funding: The authors acknowledges the financial support of the following institutions: Spanish Ministry of Education and Science (CTQ2009-14420-C02, CTQ2012-37039-C02, DPI2012-37154-C02-02, BFU2008-00196/BMC, BFU2010-17704, SGR2009-0809 and ENE 2011-28269-CO3-03), Spanish Ministry of External Affairs (projects PHB 2008-0090-PC), and European Commission (Marie Curie Actions - IAPP program - FP7/251298).

Author details

[1]Departament d'Enginyeria Química, Universitat Rovira i Virgili, Av.Països Catalans 26, 43007 Tarragona, Spain. [2]Departament de Ciències Mèdiques Bàsiques, Institut de Recerca Biomèdica de Lleida (IRBLLEIDA), Universitat de Lleida, Avinguda Alcalde Rovira Roure 80, 25198 Lleida, Spain.

Received: 26 March 2013 Accepted: 14 October 2013
Published: 31 October 2013

References

1. Chou IC, Voit EO: **Recent developments in parameter estimation and structure identification of biochemical and genomic systems.** *Math Biosci* 2009, **219:**57–83.
2. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J: **Structural and practical identifiability analysis of partially observed dynamical mod-elsby exploiting the profile likelihood.** *Bioinformatics* 2009, **25:**1923–1929.
3. Srinath S, Gunawan R: **Parameter identifiability of power-law biochemical system models.** *J Biotechnol* 2010, **149:**132–140.
4. Voit EO: **Characterizability of metabolic pathway systems from time series data.** *Math Biosci* 2013. doi:10.1016/j.mbs.2013.01.008.
5. Maki Y, Tominaga D, Okamoto M, Watanabe S, Eguchi Y: **Development of a system for the inference of large scale genetic networks.** *Pac Symp Biocomput* 2001, **2001:**446–458.
6. Vance W, Arkin A, Ross J: **Determination of causal connectivities of species in reaction networks.** *Proc Natl Acad Sci U S A* 2002, **99**(9):5816–5821.
7. Sriyudthsak K, Shiraishi F, Hirai MY: **Identification of a Metabolic Reaction Network from Time-Series Data of Metabolite Concentrations.** *PLoS One* 2013, **8**(1):e51212.
8. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BØ: **A comprehensive genome-scale reconstruction of Escherichia coli metabolism.** *Mol Syst Biol* 2011, **7:**535. doi:10.1038/msb.2011.65.
9. Sorribas A, Hernandez-Bermejo B, Vilaprinyo E, Alves R: **Cooperativity and saturation in biochemical networks: a saturable formalism using Taylor series approximations.** *Biotechnol Bioeng* 2007, **97:**1259–1277.
10. Alves R, Vilaprinyo E, Hernandez-Bermejo B, Sorribas A: **Mathematical formalisms based on approximated kinetic representations for modeling genetic and metabolic pathways.** *Biotechnol Genet Eng Rev* 2008, **25:**1–40.
11. Moles CG, Mendes P, Banga JR: **Parameter estimation in biochemical pathways: a comparison of global optimization methods.** *Genome Res* 2003, **13**(11):2467–2474.
12. Chis OT, Banga JR, Balsa-Canto E: **Structural identifiability of systems biology models: a critical comparison of methods.** *PLoS One* 2011, **6**(11):e27755. doi:10.1371/journal.pone.0027755.
13. Sorribas A, Samitier S, Canela EI, Cascante M: **Metabolic pathway characterization from transient response data obtained in situ: parameter estimation in S-system models.** *J Theor Biol* 1993, **162:**81–102.
14. Sorribas A, Cascante M: **Structure identifiability in metabolic pathways: parameter estimation in models based on the power-law formalism.** *Biochem J* 1994, **298:**303–311.
15. Vilela M, Chou IC, Vinga S, Vasconcelos AT, Voit EO, Almeida JS: **Parameter optimization in S-system models.** *BMC Syst Biol* 2008, **2:**35.
16. Markon KE, Krueger RF: **An Empirical Comparison of Information-Theoretic Selection Criteria for Multivariate Behavior Genetic Models.** *Behav Genet* 2004, **34:**593–610.
17. Akaike H: **New look at statistical-model identification.** *IEEE Trans Automat* 1974, **19:**716–723. Contr. AC19 716.
18. Schwarz G: **Estimating the dimension of a model.** *Ann Stat* 1978, **6:**461–464.
19. Burnham KP, Anderson DR: *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* 2nd edition. New York: Springer-Verlag; 2002. ISBN 0-387-95364-7.
20. Voit E: *Computational analysis of biochemical systems. A practical guide forbiochemists and molecular biologists.* Cambridge: Cambridge University Press; 2000.
21. Savageau M, Biochemical systems analysis. i: **Some mathematical properties of the rate law for the component enzymatic reactions.** *J Theor Biol* 1969, **25:**365–369.
22. Savageau M: **Biochemical systems analysis. ii. The steady-state solutions for an n-pool system using a power-law approximation.** *J Theor Biol* 1969, **25:**370–379.
23. Voit EO, Almeida J: **Decoupling dynamical systems for pathway identification from metabolic profiles.** *Bioinformatics* 2004, **20:**1670–1681.
24. Wang Y, Joshi T, Zhang XS, Xu D, Chen L: **Inferring gene regulatory networks from multiple microarray datasets.** *Bioinformatics* 2006, **22**(19):2413–2420.
25. Biegler L, Grossmann IE: **Retrospective on optimization.** *Comput Chem Eng* 2004, **28**(8):1169–1192.
26. Miró A, Pozo C, Guillén-Gosélbez G, Egea JA, Jiménez L: **Deterministic global optimization algorithm based on outer approximation for the parameter estimation of nonlinear dynamic biological systems.** *BMC Bioinformatics* 2012, **13**(1):90.
27. Grossmann IE, Biegler L: **Part II. Future perspective on optimization.** *Comput Chem Eng* 2004, **28**(8):1193–1218.
28. Pozo C, Marin-Sanguino A, Guillén-Gosalbez G, Jimenez L, Alves R, Sorribas A: **Steady-state global optimization of non-linear dynamic models through recasting into power-law canonical models.** *BMC Syst Biol* 2011, **5:**137.
29. Vecchietti A, Sangbum L, Grossmann I: **Modeling of dicrete/continuous optimization problems: Characterization and formulation of**

disjunctions and their re-laxations. *Comput Chem Eng* 2003, **27**:433–448.

30. Esposito W, Floudas C: **Global optimization for the parameter estimation of differential-algebraic systems.** *Ind Eng Chem Res* 2000, **39**(5):1291–1310.
31. Rodriguez-Fernandez M, Egea J, Banga J: **Novelmetaheuristic for parameter estimation in nonlinear dynamic biological systems.** *BMC Bioinf* 2006, **7**:483.

doi:10.1186/1752-0509-7-113
Cite this article as: Guillén-Gosálbez *et al.*: Identification of regulatory structure and kinetic parameters of biochemical networks via mixed-integer dynamic optimization. *BMC Systems Biology* 2013 **7**:113.