UNIVERSITAT ROVIRA I VIRGILI

DYNAMIC MATHEMATICAL TOOLS FOR THE IDENTIFICATION OF REGULATORY STRUCTURES AND KINETIC PARAMETERS IN SYSTEMS BIOLOGY.

Antoni Miró Roig

Dipòsit Legal: T 1768-2014

DOCTORAL THESIS

Antoni Miró Roig

DYNAMIC MATHEMATICAL TOOLS FOR THE IDENTIFICATION OF REGULATORY STRUCTURES AND KINETIC PARAMETERS IN SYSTEMS BIOLOGY

ROVIRA I VIRGILI UNIVERSITY

Department of Chemical Engineering

UNIVERSITAT ROVIRA I VIRGILI

Tarragona
2014

Antoni Miró Roig

DYNAMIC MATHEMATICAL TOOLS FOR THE IDENTIFICATION OF REGULATORY STRUCTURES AND KINETIC PARAMETERS IN SYSTEMS BIOLOGY

DOCTORAL THESIS

Supervised by:
Prof. **Gonzalo Guillén Gosálbez**
Prof. **Laureano Jiménez Esteller**

Rovira i Virgili University
Department of Chemical Engineering
SUSCAPE research group

Tarragona
2014

UNIVERSITAT
ROVIRA I VIRGILI

Departament d'Enginyeria Química
Campus Sescelades,
Avda. Països Catalans, 26
43007 Tarragona
Tel: 977 55 86 75
Fax: 977 55 96 67

Gonzalo Guillén Gosálbez and Laureano Jiménez Esteller , associate professors in the Department of Chemical Engineering,

CERTIFY that the present study, entitled "Dynamic mathematical tools for the identification of regulatory structures and kinetic parameters in systems biology", presented by Antoni Miró for the award of the degree of Doctor, has been carried out under our supervision at the Department of Chemical Engineering of the University Rovira i Virgili and meets the eligibility requirements for European distinction.

Tarragona, 8th September, 2014

Dr. Gonzalo Guillén Gosálbez

Dr. Laureano Jiménez Esteller

Agradecimientos

Ante todo quiero expresar mi más sincero agradecimiento a mis directores de tesis Dr. Gonzalo Guillén y Dr. Laureano Jiménez. Quiero darles las gracias por la oportunidad que me dieron para ingresar en el programa de doctorado. Quiero agradecerles su apoyo, consejos, motivaciones y extensos conocimientos que me han aportado, pero especialmente la gran paciencia que han tenido hacia mí. Sin su ayuda, esta tesis no habría sido posible. Quiero dar las gracias al Dr. Sebastian Sager por acogerme en su grupo de investigación durante 3 meses, por su apoyo y por los conocimientos que me aportó. Agradezco también a los miembros del tribunal al participar en la evaluación de esta tesis. Quiero mostrar mi gratitud al Ministerio de Ciencia e Innovación por adjudicarme la beca FPI. Quiero agradecer también la colaboración del grupo de investigación de Bioestadística y Biomatemáticas de la Universidad de Lleida (IRBLleida). Así como también a todos los grupos de investigación que han colaborado en la realización de los artículos en los que se basa esta tesis.

Quiero dar mi agradecimiento también a todo el personal de secretaria del DEQ por la gestión de todos los trámites administrativos durante todo el transcurso del doctorado. Quiero agradecer también a todo el personal de mantenimiento y limpieza porque gracias a su trabajo hemos podido hacer el nuestro en óptimas condiciones. Quiero dar las gracias también la Universidad, a sus directivos y a todos los profesores que he tenido durante toda la formación que he recibido. No estoy menos agradecido a mis compañeros de SUSCAPE: ha sido un placer trabajar con vosotros.

Quiero dar las gracias a toda mi familia en especial a mis padres por la educación la tolerancia i el amor recibidos. Os quiero mucho. A todos mis amigos por serlo. ¿Y por qué no? A mí mismo.

Quiero dar las gracias y que me perdone también todo aquel que me haya ayudado y no lo mencione en estos agradecimientos. Finalmente no quiero despedirme sin antes agradecerte también a ti querido lector que estas leyendo estas palabras: ¡GRACIAS!

Solo sé que no se nada y,
al saber que no sé nada, algo sé;
porque sé que no sé nada.

Sócrates (470-399 A.C.)

Summary

In this thesis we developed mathematical programming tools to facilitate the modelling of biological systems. Modeling biological systems is a noteworthy task in systems biology. In this thesis we focused on two challenging tasks: parameter estimation and regulatory structure identification.

The parameter estimation task can be posed as an optimization problem in which the sum of squares between experimental and simulated data is minimized. Regulatory structure identification can be addressed in a similar way. In this case, the signals appear in a model as parameters accounting for the influence that metabolites others than the substrates of a reaction have on its velocity. Parameter estimation of biological systems is particularly challenging due to their dynamic and nonlinear nature.

The main goal of this thesis is to develop mathematical programming tools to surmount these difficulties. Particularly, we make use of global (i.e. outer approximation) and dynamic (i.e. orthogonal collocation on finite elements) optimization tools to cope with the main challenges arising in this area. This PhD dissertation is presented using three articles (two already published and one ready to be submitted to an international peer reviewed journal).

On all publications we deal with parameter estimation problems with differential equations embedded. In order to solve this type of problems, we use dynamic optimization techniques. Among all available dynamic optimization methods, we selected a direct simultaneous approach: the orthogonal collocation on finite elements method. Simultaneous methods allow performing automatic differentiation with respect to the control and state variables, avoiding the need to calculate the derivatives numerically. Unfortunately, the discretization step can lead to large scale NLPs that are difficult to solve. In the orthogonal collocation method, both the control and state profiles are approximated using polynomials and discretized in time by means of finite elements. Particularly we used Lagrange polynomials whose collocation coefficients are distributed at the shifted (between 0 and 1) roots of the orthogonal Legendre polynomials (see *Orthogonal collocation approach* section in [1] for further information).

Deterministic global optimization strategies are the only ones that can ensure convergence to the global optimum of a non-convex problem within a desired tolerance in a

finite number of iterations. On the other hand, stochastic methods rely on metaheuristics in order to guide the search for "*good*" solutions from a series of pseudorandom generated points. These methods, which are often based on physical and biological analogies, tend to provide near optimal solutions in low CPU times, yet they offer no guarantee of global optimality.

In the first publication [1] we presented a deterministic global optimization algorithm for the parameter estimation of nonlinear biological systems. This approach is based on an outer approximation algorithm which offers a theoretical guarantee of convergence to the global optimum. In addition to the solution itself, this method provides a rigorous interval within the global optimal solution must fall. Unfortunately, this method requires a considerable amount of CPU time to ensure global optimality.

In this work we reformulate the set of ordinary differential equations describing the dynamics of the biological system into an equivalent set of algebraic equations through the use of orthogonal collocation methods, giving rise to a nonconvex nonlinear programming (NLP) problem. This nonconvex NLP is decomposed into two hierarchical levels: a master mixed-integer linear programming problem (MILP) that provides a rigorous lower bound on the optimal solution, and a reduced-space slave NLP that yields an upper bound. The master problem is a relaxation of the original NLP (*i.e.*, it overestimates its feasible region) and hence provides a rigorous lower bound on its global optimum. A valid upper bound on the global optimum is obtained by optimizing the original NLP locally. This NLP is initialized with the solution provided by the MILP. The algorithm iterates between these two levels until a termination criterion is satisfied.

A rigorous relaxation of the original model is constructed by replacing the nonconvex terms in the reformulated model by convex estimators. The solution of the convex relaxation provides a valid lower bound on the global optimum. Specifically, in this work the bilinear terms are replaced by piecewise McCormick relaxations (see *Piecewise McCormick-based relaxation section* in [1]).

We illustrated the performance of the proposed algorithm through its application to two challenging benchmark parameter estimation problems: the isomerisation of α-pinene and the inhibition of HIV proteinase. The final goal in these problems is to obtain the set of values of the model parameters with which the model response is as close as possible to the experimental data. For comparison purposes, we used the global optimization package BARON (Branch And Reduce Optimization Navigator). BARON is a commercial software for solving nonconvex optimization problems to global optimality that implements the latest theoretical and algorithmic developments in global optimization.

In the first case study, which focuses on the α-pinene, BARON was able to find the global optimum but failed at reducing the optimality gap below the specified tolerance after 12h of CPU time. In contrast, our algorithm closed the gap in less than 3h. The differences between this algorithm and BARON were more remarkable in the HIV proteinase case study, where BARON failed to identify any feasible solution after 12h of CPU time. Our algorithm closed the gap in approximately 4000 CPU s, and outperformed BARON and other parameter estimation methods, improving the best known solution, and providing a rigorous lower bound on the minimum error that can be attained.

The overwhelming majority of parameter estimation methods assumes a given structure and considers a fix regulatory scheme. This simplification is motivated by the difficulty in identifying regulatory effects. In this PhD thesis, we propose a strategy to simultaneously address the challenging tasks of estimating the parameters and regulatory topology of biochemical networks from time-series data.

In the second publication [2] we proposed a rigorous and systematic parameter estimation and network identification method that makes no assumption regarding the regulatory network topology. The capabilities of this methodology were illustrated through its application to a case study taken from Voit and Almeida [3]. In this case study, a canonical model structure in the context of the power-law kinetic formalism called GMA (Generalized Mass-Action) was used for representing the systems' kinetics. Our approach is not restricted to this particular representation, as it can accommodate any general representation flexible enough to account for the regulatory interactions in a biological network.

To model the existence of a regulatory interaction, we apply disjunctive programming techniques that allow transforming the problem into a mixed-integer dynamic optimization (MIDO) problem through the use of the big-M reformulation (See *section*

4.2). Particularly, we solved this MIDO by reformulating it into a mixed-integer nonlinear programming (MINLP) problem using orthogonal collocation on finite elements, which makes it possible to apply standard MINLP solution algorithms iteratively in order to identify a set of plausible network topologies and associated kinetic parameters. Our MIDO approach can be solved using any standard MIDO solution algorithm, and it is not restricted to the use of orthogonal collocation and MINLP reformulations.

One important feature of the approach followed 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 an integer cut, which excludes solutions identified so far in previous iterations from the search space.

We tested first the extent to which the method can identify the model parameters when the regulatory structure is known and assuming no error. As expected we obtained values very close to the reported ones. We observed that parameter trends can be obtained by fixing a given parameter and fitting the remaining ones. 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*.

Next, we considered the effect of noisy data on fitting the model. We performed the calculations first for one experiment, that is, a single configuration for the initial conditions for state variables. Later we repeated the calculations for three different experiments. The perturbations force the system to move across different dynamic regimes, producing additional information that helps in the identification of appropriate parameter values. In both cases, considering one and three experiments, we noticed that despite the different parameter values, the various fitted models lead to similar residuals.

Lastly, we explored the ability of the method to identify the regulatory structure using one and three experiments with low experimental error. For one experiment, the method identifies topologies that are quite close and that show very small residuals, but it is unable to uniquely identify the original topology. Considering three experiments, 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).

In the last publication [4] we presented an alternative approach based on bi-objective optimization to simultaneously identify the regulatory interactions along with the kinetic parameters (assuming a kinetic representation) from time series data. The performance of this strategy is tested through its application to the same case study taken from Voit and Almeida [3] using the ε–constraint method (see section 2.3.1). We illustrated the ability of this novel strategy to identify the regulatory structure using one and three experiments in four different uncertainty scenarios. *In silico* data were generated via Monte Carlo sampling from normal distributions that assuming standard deviations of 5, 10, 15 and 30%. For comparison purposes, we solved the same problem using the single-optimization approach previously developed [2].

We then assessed the quality of the predictions made by the models produced by each approach under untested conditions (using a validation set).

After performing the calculations for one and three experiments, the results showed that for the majority of instances, the minimization of the Akaike information criterion (AIC) as single objective produces models with better predictive accuracy, that is, models with lower residuals in the validation set (as well as lower AIC values). In addition, we found that the AIC values of the models identified by both approaches are rather similar. Therefore, the alternative models cannot be directly discarded, since their AIC values are obtained using a limited number of points, and because of this the model with minimum AIC value might not be the best possible model.

In summary, this thesis introduces a set of advanced mathematical tools for identifying the regulatory structure and kinetic parameters of dynamic biochemical systems. The tools used are based on dynamic optimization (DO), mixed integer dynamic optimization (MIDO), multi-objective optimization (MOO) and global optimization (GO).

The thesis is organized as follows: in the first section we introduce the topic of the thesis, the section 2 provides the general background in mathematical programming and other tools used in this thesis, in the section 3 the fundamentals of model selection are introduced, in the section 4 the type of problems addressed are introduced. In the section 5, we illustrate the capabilities and the numerical results of these approaches applied to systems biology problems. In the last part, the conclusions and future work sections are outlined. The publications derived from the work can be found in the *Articles* section.

Resumen

En esta tesis desarrollamos herramientas de programación matemática para abordar la modelización de sistemas biológicos. El modelado de sistemas biológicos es uno de los cometidos más importantes en biología de sistemas. En esta tesis nos centramos en dos tareas complejas: la estimación de parámetros y la identificación de la estructura regulatoria.

La estimación de parámetros puede formularse como un problema de optimización en el que se minimiza la suma de cuadrados entre los datos experimentales y simulados. La identificación de la estructura regulatoria puede abordarse de la misma manera, pero en este caso las señales aparecen en el modelo como parámetros que representan la influencia de los metabolitos en las velocidades de reacción. Determinar los parámetros de sistemas biológicos es particularmente complejo debido a su naturaleza dinámica y no lineal.

El objetivo principal de esta tesis es desarrollar herramientas de programación matemática para superar estas dificultades. En particular, hacemos uso de herramientas de optimización global (*outer approximation*) y dinámica (colocación ortogonal en elementos finitos). Esta tesis doctoral se presenta utilizando tres artículos que se han publicado o están listos para ser presentados a revistas internacionales arbitradas.

En todas las publicaciones nos ocupamos de problemas de estimación de parámetros con ecuaciones diferenciales incorporadas. Para revolver este tipo de problemas hay que recurrir a técnicas de optimización dinámica. Entre todos los métodos de optimización dinámicos disponibles se optó por un enfoque directo simultáneo: el método de la colocación ortogonal en elementos finitos. Los métodos simultáneos permiten la diferenciación automática con respecto a las variables de control y de estado, evitando la necesidad de calcular las derivadas numéricamente. Desafortunadamente, la discretización puede conducir a problemas no lineales (NLP) de elevada complejidad que son

difíciles de resolver. En el método de colocación ortogonal, tanto los perfiles de control como los de estado, se aproximan mediante polinomios y se discretizan en el tiempo por medio de elementos finitos. Concretamente se usaron los polinomios de Lagrange cuyos coeficientes de colocación se distribuyeron en las raíces de los polinomios ortogonales de Legendre normalizados (entre 0 y 1) (ver la sección *Orthogonal collocation approach* en [1] para más información).

Las estrategias de optimización global deterministas son las únicas que pueden garantizar la convergencia al óptimo global de un problema no convexo con una tolerancia deseada en un número finito de iteraciones. Por otro lado, los métodos estocásticos se basan en meta-heurísticas para guiar la búsqueda de soluciones "*buenas*" partiendo de una serie de puntos generados pseudo-aleatoriamente. Estos métodos están inspirados a menudo en analogías físicas y biológicas y son capaces de obtener soluciones casi óptimas en tiempos de CPU reducidos. Sin embargo, no ofrecen ninguna garantía de optimalidad global.

En la primera publicación [1] presentamos un algoritmo de optimización global determinista para la estimación de parámetros en sistemas biológicos no lineales. Este enfoque se basa en el algoritmo *outer approximation* el cual ofrece una garantía teórica de convergencia hacia el óptimo global. Además de la propia solución, este método proporciona un intervalo riguroso que contiene la solución óptima global. Por desgracia, este método requiere una cantidad considerable de tiempo de CPU para garantizar la optimalidad global.

Nuestro trabajo se basa en reformular el sistema de ecuaciones diferenciales ordinarias en un conjunto equivalente de ecuaciones algebraicas mediante el uso de métodos de colocación ortogonal, dando lugar a un problema de programación no lineal (NLP) no convexo. Este NLP no convexo se descompone en dos niveles jerárquicos: un problema *master* de programación lineal entera mixta (MILP) que proporciona una cota inferior rigurosa de la solución óptima global, y un problema NLP *esclavo* en el espacio reducido que ofrece una cota superior. El problema *master* es una relajación del problema NLP original (es decir, se sobreestima su región factible) y por lo tanto proporciona un límite inferior riguroso en su óptimo global. La cota superior sobre el óptimo global se obtiene optimizando el problema NLP original localmente. Este problema NLP se inicializa utilizando la solución aportada por el problema *master* MILP como punto inicial. El algoritmo itera entre estos dos niveles hasta que uno de los criterios de finalización se satisface.

La relajación rigurosa del modelo original se construye sustituyendo los términos no convexos en el modelo reformulado utilizando estimadores convexos. La solución de la relajación convexa proporciona una cota inferior válida en el óptimo global. Concretamente, en este trabajo los términos bilineales se reemplazan por relajaciones *piecewise* de McCormick (véase la sección *Piecewise McCormick-based relaxation* en [1]).

Ilustramos el rendimiento del algoritmo propuesto a través de su aplicación a dos problemas de referencia en la estimación de parámetros: la isomerización del α-pineno y la inhibición de la proteinasa del HIV. El objetivo en estos problemas es obtener el conjunto de valores de los parámetros del modelo de tal manera que su respuesta sea lo más cercana posible a los datos experimentales. Para comparar se utilizó el paquete de optimización global BARON (*Branch And Reduce Optimization Navigator*). BARON es un software comercial para la resolución de problemas de optimización no convexos que identifica el óptimo global del problema considerando una tolerancia deseada.

En el caso del α-pineno, BARON fue capaz de encontrar el óptimo global, pero no logró reducir el intervalo de optimalidad por debajo de la tolerancia especificada después de 12h de tiempo de CPU. Por el contrario, el algoritmo desarrollado logró alcanzar dicho intervalo en menos de 3 horas. Las diferencias entre nuestro algoritmo y BARON fueron más notables en el caso de estudio de la proteinasa del HIV. BARON no identificó ninguna solución factible después de 12h de tiempo de CPU. Nuestro algoritmo cerró el intervalo en aproximadamente 4000 segundos de CPU, superando claramente a BARON así como a otros métodos de estimación de parámetros, mejorando la mejor solución conocida, y proporcionando un límite inferior riguroso en el mínimo error que se puede alcanzar.

La inmensa mayoría de los métodos de estimación de parámetros asume una estructura dada y considera un esquema regulatorio fijo. Esta simplificación está motivada por la dificultad en la identificación de los efectos regulatorios. Esta tesis doctoral propuso una estrategia para abordar simultáneamente las difíciles tareas de estimación de parámetros y de identificación de la topología de regulación de las redes bioquímicas a partir de datos de series temporales.

En la segunda publicación [2] proponemos un método riguroso y sistemático de estimación de parámetros y de identificación de la red que no hace ninguna suposición con respecto a la topología de la red de regulación. Las capacidades de nuestra metodología se ilustran a través de su aplicación a un caso de estudio tomado de Voit y Almeida [3].

En este ejemplo, se utilizó como representación cinética un modelo de estructura canónica en el contexto del formalismo cinético de la ley de potencia llamada GMA (*Generalizad Mass-Action*). Sin embargo, nuestro enfoque no está restringido a esta representación en particular, ya que cualquier representación general lo suficientemente flexible como para tener en cuenta las interacciones de regulación en una red biológica puede ser utilizada para el mismo propósito.

Para modelar la existencia de una interacción regulatoria, hacemos uso de programación disyuntiva para plantear un problema MIDO (*mixed-integer dynamic optimization*) mediante el uso de la reformulación *big-M (*Véase la sección 4.2). En particular, se resuelve el problema MIDO reformulándolo como un problema de programación no lineal entera mixta (MINLP) utilizando la colocación ortogonal en elementos finitos, lo que hace posible la aplicación de algoritmos estándar para MINLP de forma iterativa con el fin de identificar un conjunto de topologías de red plausibles con sus parámetros cinéticos correspondientes. Este problema MIDO, sin embargo, se puede resolver con cualquier algoritmo MIDO, y no necesariamente usando colocación ortogonal y reformulaciones MINLP.

Una característica importante del enfoque desarrollado es que en lugar de calcular una única solución óptima, identifica un conjunto de topologías reguladoras plausibles resolviendo el modelo de forma iterativa. Es decir, el modelo se resuelve primero para identificar una potencial configuración regulatoria representada por una solución binaria (un conjunto de valores de las variables binarias). Dicho modelo se calcula entonces de nuevo, pero esta vez añadiendo un corte entero, que excluye a las soluciones identificadas hasta ahora en anteriores iteraciones del espacio de búsqueda.

Primero se probó la capacidad de nuestro método para identificar los parámetros del modelo cuando la estructura regulatoria es conocida y suponiendo que no hay error. Como era de esperar se obtuvieron los valores estimados de los parámetros muy cercanos a los originales. Hemos observado que se pueden obtener tendencias de los parámetros fijando el valor de un parámetro dado y ajustando los restantes. En términos prácticos, esto significa que dado un experimento y método de estimación, se podrían obtener diferentes conjuntos de parámetros que reproducen fielmente las medidas experimentales, pero que difieren de los valores con los que el perfil dinámico se generó *in silico*.

A continuación, hemos considerado el efecto de los datos con ruido en el ajuste del modelo. Hemos realizado los cálculos primero para un solo experimento, es decir, una única configuración para las condiciones iniciales de las variables de estado. Luego repetimos los cálculos para tres experimentos diferentes. Las perturbaciones fuerzan al sistema a moverse a través de diferentes regímenes dinámicos, produciendo información adicional que es de ayuda en la identificación de los valores apropiados de los parámetros. En ambos casos, tanto considerando un único experimento así como tres experimentos, constatamos que a pesar de los diferentes valores de los parámetros, los distintos modelos ajustados conducen a residuales similares.

Por último, hemos explorado el rendimiento del método utilizando uno y tres experimentos con un error experimental pequeño. Para un experimento, el método identifica topologías que se asemejan bastante a la original y que muestran residuos muy pequeños, pero no es capaz de identificar de forma unívoca la topología original. Considerando tres experimentos, el método identifica no sólo la topología real, sino también varias estructuras que contienen la original (es decir, las topologías que tienen en cuenta todos los efectos reguladores reales más las otras señales que no estaban presentes originalmente).

En la última publicación [4] hemos presentado un enfoque alternativo basado en la optimización bi-objetivo para identificar simultáneamente las interacciones reguladoras junto con los parámetros cinéticos (suponiendo una representación cinética) a partir de datos de series de tiempo. Esta estrategia fue evaluada a través de su aplicación al mismo caso de estudio tomado de Voit y Almeida [3]. Aquí ilustramos la capacidad de esta nueva estrategia para identificar la estructura regulatoria utilizando uno y tres experimentos en cuatro escenarios diferentes de incertidumbre. Los datos *in silico* fueron generados mediante muestreo de Monte Carlo usando una sola muestra y suponiendo desviaciones del 5, 10, 15 y 30% en las distribuciones normales. Para fines comparativos, resolvimos el mismo problema con el enfoque de optimización de un objetivo único desarrollado anteriormente [2].

Posteriormente se evaluó qué método es capaz de generar mejores modelos, entendiendo por mejores aquéllos que predicen mejor el comportamiento del sistema bajo condiciones no testadas (utilizando el set de validación). Después de realizar los cálculos para uno y tres experimentos, los resultados mostraron que para la mayoría de los casos la minimización del criterio de información de Akaike (AIC) como único objetivo produce modelos con una precisión mejor de predicción, es decir, modelos con residuales más pequeños en el conjunto de datos de validación (así como valores más bajos del AIC). Sin embargo, estas soluciones no pueden ser directamente descartadas, ya que cuando los modelos tienen valores similares del AIC, el modelo con el menor valor del AIC puede no ser la mejor opción.

Idealmente el mínimo valor del AIC debe pertenecer a la frontera de Pareto obtenida en el enfoque bi-criterio. Sin embargo, en la práctica esta condición no se cumple en todos los casos debido a dos razones. Por un lado, la expresión modificada que tenemos usar en este caso para el cálculo de la AIC (es decir, el AIC corregido, también conocido como AICc) no cumple algunas de las propiedades que se requieren para establecer la analogía entre el AIC y el bi-criterio entre calidad y complejidad del ajuste. Por otro lado, no podemos garantizar el óptimo global de las soluciones de los MINLPs resueltos con un optimizador local (véase la sección *Methods* en [4] para más información).

La tesis se organiza de la siguiente manera: En la primera sección introducimos el tema de la tesis, la sección 2 proporciona el marco general de la programación matemática y otras herramientas utilizadas en esta tesis. En la sección 3 se introducen los fundamentos de la selección de modelos, mientras que en la sección 4 se introducen el tipo de problemas abordados. En la sección 5 se ilustran las capacidades y los resultados numéricos de estos enfoques aplicados a problemas de biología de sistemas. En la última parte se describen las conclusiones y el trabajo futuro. Las publicaciones derivadas de esta tesis se encuentran en la sección *Articles*.

Table of contents

1. Introduction

Since the beginning of the past decade, modern high-throughput techniques have led to an explosion in the rate of data generated for characterizing the dynamics of genomic, proteomic and metabolic responses in biological systems. Consequently, one of the most fundamental challenges in systems biology is to glean biochemical significance from mounds of data. The behavior of biological systems is in many cases too complex to allow intuitive predictions and require the support of powerful theoretical tools from mathematics and physical sciences. Particularly, mathematical models of biochemical systems are becoming essential in systems biology to complement and extract information from time series.

The task of biomathematical modeling involves the conversion of the observed biological system phenomena into a simplified mathematical analogue that mimics its behavior and that is easier to study, predict, manipulate and optimize than the biological system itself. The typical procedure to end up with a reliable model comprises five 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); (iv) inferring the system's regulatory structure; and (v) checking the predictive performance of the model (model validation).

The first challenge requires compiling information available about the system in order to generate its corresponding stoichiometric matrix. The next challenge entails the selection of the appropriate mathematical model among the different representations available. This step depends on the previous knowledge about the system. If enough information is known, mechanistic formulations based on physical sciences (*e.g.*, law of mass action, Michaelis–Menten rate law…) are a good choice. Unfortunately, optimizing these systems is not a straightforward task as it usually leads to complex mathematical formulations [5]. If the degree of knowledge is lower, it is often more convenient to use a generic formulation capable of capturing the nonlinear dynamics while keeping the model relatively simple. Canonical models are particularly useful for this purpose. In addition, canonical models render possible parameter estimation and topology identification tasks simultaneously [6].

Among them, models using the so called power-law formalism show a good compromise between accuracy and simplicity [7]. This group includes the S-System and the General Mass Action (GMA) models, which seem a promising alternative [8,9]. The main advantage of these models is that they can capture the nonlinearities required to describe the regulatory processes of the networks. Additionally, these models constitute a very general framework since any kind of metabolic network can be represented through their formulations [10]. GMA models only differ from S-System models in the way in which the branching points are handled [11]. In S-System models, all the input flows in the branching point are collected and modeled together as if they were a single flow. The same procedure is followed for the outputs so that, finally, the concentration of the metabolite being balanced is the result of just two contributions. On the other hand, in GMA models each process is approximated separately so that there are as many contributions as actual flows in the real system [12]. If the metabolic network only contains nodes that result from the contribution of one input flow and one output, the S-System and GMA representations coincide.

In systems biology there is a strong tendency to build very complex models. In this situation, when a model has too many parameters it is said to be overfitted. Overfitted models should be avoided since the task of biomathematical modeling is not trying to model the data perfectly, instead it has more to do with recovering information from the time series data. In other words, since data contain both information and noise the goal is to extract the information that applies to the process in general rather than that contained in the particular data set. Conversely underfitted models are highly biased from reality and therefore their predictions might be unreliable.

Clearly, a trade-off between under- and overfitting is needed but we cannot rely on intuition to assess such a trade-off, instead a criterion based on deep information theory is demanded. Information-theoretic criteria such as the Akaike information criterion (AIC) [13] or the Bayesian information criterion (BIC) [14], are now perceived as important measures to assess quality of the fitting. AIC is often preferred over BIC because it has a more immediate connection to the theory of information [15]. AIC captures the trade-off between the complexity (measured by the number of parameters), and accuracy of the fitting. AIC selects the fitted approximating model that is estimated, on average, to be closest to the unknown full reality [15]. Smaller AIC values imply a better approximation to the model sought (See section 3).

The third challenge we deal with is to determinate the appropriate numerical parameter values. The aim here is to obtain the set of parameter values that make the model response consistent with the data observed. Particularly, the parameter estimation task can be formulated as an optimization problem in which the sum of squared residuals between the measured and simulated data is minimized.

The fourth challenge can be addressed in a similar way as the third one, since parameters accounting for the influence that metabolites others than the substrates of a reaction have on its velocity can be easily incorporated into the parameter estimation model.

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*. Recovering the network topology and associated kinetic parameter values from time-series data is a very challenging task of paramount importance in systems biology. Usually, characterization of the regulatory topology is more difficult than parameter estimation.

The type of optimization problem being faced and the technical challenges to be solved depend upon the biological model of choice, upon the experimental data available, upon computational issues, and upon the specific mathematical formalism used. Studying dynamic responses of biological systems is particularly appealing in systems biology. Dynamic biological systems are described through nonlinear ordinary differential equations (ODEs) that provide the concentration profiles of certain genes, proteins and metabolites over time. The biomathematical modeling of these systems gives rise to dynamic optimization problems which are hard to solve.

Existing approaches to optimize dynamic models can be roughly classified as direct or indirect (also known as variational) [16]. Direct methods make use of gradient-based nonlinear programming (NLP) and can in turn be divided into sequential and simultaneous. In sequential approaches, the optimization of the control variables, which are discretized, are performed by a NLP solver, whereas the ODE is calculated externally, that is, both steps are executed in a sequential manner. In contrast, in simultaneous strategies, both the control and state profiles are approximated using polynomials (*e.g.*, Lagrange polynomials) and discretized in time by means of finite elements [17, 18]. In the latter strategy, the ODE system is replaced by a system of algebraic equations that is optimized with a standard gradient-based NLP solver. Simultaneous approaches allow performing automatic differentiation with respect to the control and state variables, avoiding the need to calculate the derivatives numerically as is the case in the sequential approach. Unfortunately, the discretization step can lead to large scale NLPs that are difficult to solve. Multiple shooting methods serve as a bridge between sequential and simultaneous approaches.

Optimization problems involving biological systems are usually of nonconvex nature, which gives rise to multiple local solutions (*i.e.*, multimodality). Because of this, traditional gradient-based methods used in the sequential and simultaneous approaches may fall in local optima. In the context of parameter estimation, these local solutions should be avoided, since they may lead to inaccurate models that are unable to predict the system's performance precisely.

Global optimization (GO) algorithms are a special type of techniques that attempt to identify the global optimum in nonconvex problems. These methods can be classified as stochastic and deterministic. Stochastic GO methods are based on probabilistic algorithms that provide near optimal solutions in short CPU times. Despite having shown great potential with large-scale problems like parameter estimation [19], these methods have as major limitation that they are unable to guarantee convergence to the global optimum in a finite number of iterations. In other words, they provide solutions whose optimality (*i.e.*, quality) is unknown, and may or may not be globally optimal. In contrast, deterministic global optimization methods ensure global optimality within a desired tolerance, but lead to larger computational burdens. Hence, in addition to the solution itself, these methods provide as output a rigorous interval within which the best possible solution (*i.e.*, global optimum) must fall.

Two main deterministic GO methods exist: spatial branch-and-bound (sBB) [17, 20-22], and outer approximation [23]. Both algorithms rely on computing valid lower and upper bounds on the global optimum. These bounds tend to approach as iterations proceed, thus offering a theoretical guarantee of convergence to the global optimum.

A rigorous lower bound on the global optimum of the original nonconvex problem is obtained by solving a valid relaxation that contains its feasible space. To construct this relaxed problem, the nonconvex terms in the original formulation are replaced by convex envelopes that overestimate its feasible region. There are different types of convex envelopes that provide good relaxations for a wide variety of nonconvexities (See section 2.1.1). These relaxations are the main ingredient of deterministic GO methods and play a key role in their performance. In general, tighter relaxations provide better bounds (*i.e.*, closer to the global optimum), thereby expediting the overall solution procedure. A valid upper bound on the global optimum is obtained by optimizing the original NLP locally. This NLP is initialized using the solution provided by the MILP as starting point. The solution of this NLP is used to tighten the MILP, so the lower and upper bounds tend to converge as iterations proceed.

Once the model is well defined, the task of the last challenge is to test the validity of the model, that is, the model whose parameters are predicted should be able to predict the systemic responses under yet untested experimental conditions. To this end, we carry out a model validation procedure.

This thesis is devoted to overcoming the five challenges described above in the process of building a mathematical model in systems biology. Specifically, a deterministic outer approximation-based algorithm was developed in [1] for the global optimization of dynamic problems in the context of the parameter estimation of models of biological systems. This approach is based on the reformulation of a differential-algebraic system into an equivalent set of algebraic equations through the use of the orthogonal collocation on finite elements method evaluated at the shifted roots of Legendre polynomials (See *Orthogonal collocation approach* section in [1]). The resulting NLP is then decomposed into two hierarchical levels: a master MILP that provides a rigorous lower bound on the global optimal solution, and a reduced-space slave NLP problem that provides an upped bound. Two case studies consisting of the isomerisation of α-pinene and the inhibition of HIV proteinase were solved. The results obtained were compared with those produced by the state-of-art commercial global optimization package BARON (Branch And Reduce Optimization Navigator).

In [2], we presented an approach for simultaneously estimating the parameters and regulatory topology of biochemical networks from dynamic time-series data. Following this approach, we reformulated a mixed-integer dynamic optimization (MIDO) problem into a mixed-integer nonlinear programming (MINLP) problem through the use of orthogonal collocation methods. The goal was to identify the solution that minimizes the Akaike information criterion (AIC). We tested the capabilities of our approach through its application to a case study taken from Voit and Almeida (2004).

The last contribution of this Thesis [4] consists of an alternative method to determine the regulatory topology of biochemical networks from dynamic time-series data. In this publication, the inference task is posed as a bi-objective MIDO problem in which the complexity and the deviation from reality (*i.e.*, the squared residual of the fitting of time series data) are simultaneously minimized. This problem was solved by applying the ε-constraint method (see section 2.3.1), which identifies a set of candidate models with an increasing number of regulatory interactions. The MIDO problems are further reformulated into non-convex MINLP models after complete discretization based on orthogonal collocation on finite elements. This method was applied to the same case study taken from Voit and Almeida (2004).

For comparison purposes, we solved the same problem using the single-optimization approach previously developed. We thereby assessed the performance of both methods using a Cross-validation (CV) strategy and computing the AIC value for each model.

2. Mathematical programming and optimization

Although optimization started as a methodology of academic interest, it has become a useful technology with significant impact in almost all areas of engineering and science [24]. In mathematical programming, optimization problems are generally posed as minimizations (by reversing the sign of the objective function, we can easily pose maximization problems as well):

$$\begin{array}{lll} SOO & \min & f_1 \\ & s.t. & h\left(x,y\right)=0 \\ & & g\left(x,y\right)\leq 0 \\ & & x\in R, y\in Z \end{array} \tag{1}$$

Single Objective Optimization problems (SOO) are composed of different parts. On the one hand, the objective function f_1 can be understood as the performance index of a given solution. Feasible alternatives (the set of which is sometimes referred to as search space or solution space) are defined by the constraints in the problem. In particular, *h(x,y)* represents equality constraints whereas *g(x,y)* refers to inequality constraints. Regarding the decision variables, these can either be continuous (denoted by *x*) or integer (represented by *y*). Note that widely-used binary variables are a particular case of integer ones.

The nature of an optimization problem is given by the particular combination of variables and equations it embeds. As a result, one may face linear programming problems (LP, continuous variables and linear equations), non-linear programming problems (NLP, continuous variables and one or more non-linear equations), mixed-integer linear programming problems (MILP, continuous and integer variables, and linear equations) and mixed-integer non-linear programming problems (MINLP, continuous and integer variables, and at least one non-linear equation) among others. Special distinction needs to be made regarding whether the NLP is convex or not, as this second case may give rise to multiple local optimal solutions (*i.e.*, multimodality). The existence of multiple local sub-optimal solutions is a handicap when addressing these problems as standard algorithms may get trapped in them during the search, reporting a solution far away from the global one.

An optimization problem is said to be convex when its objective function and its feasible space are both convex. A feasible space is convex if and only if the inequality constraints are convex and the equality constraints are affine (*i.e.*, linear). In a convex search space, any linear combination of two points of the feasible space leads to a point belonging to the same space, whereas in a non-convex one, it does not (Figure 1). Note that according to this definition any problem involving integer variables is non-convex, since its solution space is defined by disjoint regions. In practice, however, MINLP formulations are in general referred to as non-convex only when the NLP resulting from fixing the values of their integer variables is non-convex. Similarly, MILPs are non-convex because of the presence of binary variables.

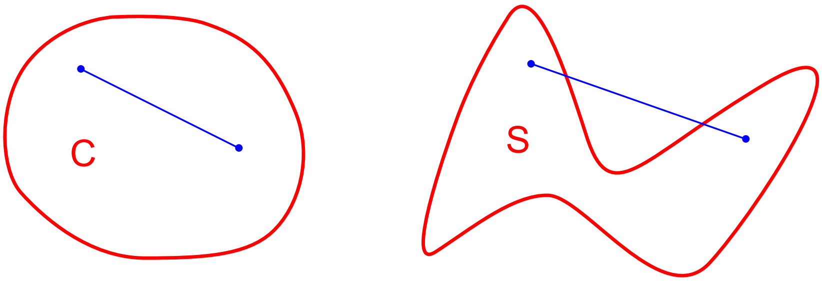

Figure 1 Example of a convex space C and a non-convex space S.

2.1. Global optimization

Deterministic global optimization strategies are the only ones that can ensure convergence to the global optimum of a non-convex problem within a desired tolerance in a finite number of iterations. Some of these methods have been implemented in software applications (for instance, a spatial branch and reduce algorithm is implemented in BARON, the state-of-art global optimization solver).

Here, we should distinguish between stochastic and deterministic approaches. Stochastic methods rely on meta-heuristics in order to guide the search for "*good*" solutions from a series of pseudorandom generated points. These methods are often based on physical and biological analogies and are capable of obtaining near optimal solutions in low CPU times, yet they offer no guarantee of global optimality in a finite number of iterations. On the other hand, as already mentioned, deterministic methods are rigorous and, thus, can guarantee global optimality within a desired optimality gap. These methods are based on calculating valid lower and upper bounds on the global optimum of the problem that are gradually tightened until a desired optimality criterion is satisfied. The main drawback of such strategies is that they require a large number of iterations to converge, and sometimes, even after large CPU times, they cannot close the optimality gap (defined as the absolute value of the relative difference between the upper and the lower bounds) bellow certain limits [25]. The search for the global optimum can be expedited by exploiting the mathematical properties of the specific problem. Hence, there is still room for improvement in this area by devising customized algorithms for specific applications.

In this thesis, we have developed efficient deterministic global optimization techniques for non-convex NLPs arising in parameter estimation studies. From now on, we will refer to deterministic global optimization simply as global optimization.

2.1.1. Relaxations in global optimization

One key feature of any global optimization algorithm is its capability of predicting valid lower bounds on the global optimum. This is usually accomplished by solving a so called convex relaxation of the original nonconvex problem. A relaxation is an auxiliary problem obtained with an objective function that underestimates the original one and a search space that contains that of the original problem.

To program a valid MILP relaxation, we apply the following approach. We first reformulate the NLP using the symbolic reformulation method proposed by Smith and Pantelides [26]. This technique reformulates any system of nonlinear equations into an equivalent canonical form with the only nonlinearities being bilinear product, linear fractional, simple exponentiation and univariate function terms with the following standard form:

$$
\begin{aligned}
&\min_{w} \qquad w_{obj} \\
&s.t. \qquad Aw = b \\
&w^{l} \le w \le w^{u} \\
&y \in \left[y^{l}, \ldots, y^{u} \right] \\
&w_{k} \equiv w_{i} w_{j} \quad \forall \left(i, j, k \right) \in \mathscr{T}_{\text{bt}} \\
&w_{k} \equiv \frac{w_{i}}{w_{j}} \quad \forall \left(i, j, k \right) \in \mathscr{T}_{\text{lft}} \\
&w_{k} \equiv w_{i}^{n} \quad \forall \left(i, k, n \right) \in \mathscr{T}_{\text{et}} \\
&w_{k} \equiv \text{fn}\left(w_{i} \right) \quad \forall \left(i, k, n \right) \in \mathscr{T}_{\text{uft}}
\end{aligned}
\tag{2}
$$

where vector *w* comprises continuous variables *x* as well as integers *y*, while the sets $\mathscr{T}_{\text{bt}}$, $\mathscr{T}_{\text{lft}}$, $\mathscr{T}_{\text{et}}$ and $\mathscr{T}_{\text{uft}}$ are the bilinear product, linear fractional, simple exponentiation and univariate function terms, respectively. A rigorous relaxation of the original model is constructed by replacing the nonconvex terms in the reformulated model by convex estimators. The solution of the convex relaxation provides a valid lower bound on the global optimum.

The objective of a global optimization algorithm is to approach the lower and upper bounds produced to the globally optimal solution. In the case of the lower bound, this can only be accomplished by means of tight relaxations. Hence, in this thesis we studied how to obtain tight relaxations for the problems of interest.

2.2. Dynamic optimization

Dynamic optimization, sometimes called optimal control, aims to determine a set of time profiles for a dynamic system that optimize a given performance index subject to specified constraints. We consider the general formulation of a dynamic optimization problem described by a set of ordinary differential equations (ODE):

$$
\begin{aligned}
\min_{\theta,u(t)} \quad & J\Big[x(t)\Big] \\
s.t. \quad & \dot{x}(t) = f\Big[t, x(t), u(t), \theta\Big], \qquad x(0) = x_0 \\
& h\Big[t, x(t), u(t), \theta\Big] = 0 \\
& g\Big[t, x(t), u(t), \theta\Big] \leq 0 \\
& x(t)^L \leq x(t) \leq x(t)^U \\
& u(t)^L \leq u(t) \leq u(t)^U
\end{aligned}
\tag{3}
$$

where *f* is the vector of differential equations, *h* is the vector of equality constraints, *g* is the vector of inequality constraints, *x(t)*, is the vector of state variables, x_0 its initial conditions and *u(t)* is the vector of control variables.

Among all the methods available to solve dynamic optimization problems, we focus on simultaneous approaches in the context of direct dynamic optimization methods. Simultaneous approaches are particularly appealing because they allow performing automatic differentiation with respect to the control and state variables, avoiding the need to calculate the derivatives numerically. The downside is that the discretization leads to NLP problems of considerable size.

Particularly, our approach entails the complete discretization of control and state variables using orthogonal collocation on finite elements [27, 28]. Notice that the dynamic optimization problems we solve here are not strictly speaking optimal control problems, but rather problems in the context of parameter estimation. Because of this, we do not have to deal with control profiles (see *Orthogonal collocation approach* section in [1] for further details).

2.3. Multi-objective optimization

Sometimes it might be interesting to evaluate alternatives considering more than one criterion. This can be accomplished by appending additional objectives to the problem formulation and by solving the resulting multi-objective optimization (MOO) problem of the following form:

$$\begin{aligned} MOO \quad & \min \quad F = \left\{ f_1, ..., f_B \right\} \\ & s.t. \quad h\left(x, y\right) = 0 \\ & \qquad\;\; g\left(x, y\right) \leq 0 \\ & \qquad\;\; x \in \Re,\, y \in \left\{0,1\right\} \end{aligned} \tag{4}$$

Recall that the difference between problem MOO and problem SOO relies on the objective function. In particular, in problem SOO, f_1 can be regarded as a single objective function whereas in problem MOO, F is a vector containing a set of B objectives ranging from f_1 to f_B.

The vector containing the individual minimum of all the objectives is referred to as the utopia point. This point is in general unattainable due to the trade-off existing between the different objectives. As a result, the solution to this kind of problems is usually composed by a set of points instead of a single one. These points are known as Pareto optimal solutions and form the so-called Pareto frontier. A solution is said to be Pareto optimal when it is not possible to improve one of the objectives without worsening any of the others. For this reason, points in a given Pareto set are all considered to be equally optimal (see [29] for further information). The most popular MOO methods are the weighted sum and the epsilon constraint methods. The former suffers from a well-reported inability to obtain non-convex parts of the Pareto frontier. For this reason, it was not used in this thesis (see [29] for a description of this method). The epsilon constraint does not show this limitation and has been thus adopted in this work due to its simplicity and ease of implementation.

2.3.1. Epsilon constraint

In this method, one objective (main objective) is regarded as main objective, while the rest (secondary objectives) are transferred to auxiliary constraints that impose bounds ε on them:

$$
\begin{array}{lll}
EC & \min & f_1 \\
 & s.t. & f_b \leq \varepsilon_b^n \qquad b = 2,...,B \\
 & & h\left(x,y\right) = 0 \\
 & & g\left(x,y\right) \leq 0 \\
 & & x \in \Re,\, y \in \left\{0,1\right\}
\end{array}
\tag{5}
$$

The values of the epsilon parameters are obtained by first optimizing each objective individually and then splitting the interval defined by the best $\underline{\left(f_b\right)}$ and worst $\overline{\left(f_b\right)}$ values obtained for each objective in this optimization, into a set of subintervals.

One important feature of the epsilon constraint method is that it transforms a MOO problem into a set of single-objective problems, which can be solved by means of any single-objective optimization method.

3. Model Selection: the Akaike information criterion (AIC)

Full reality cannot be extracted from the analysis of a finite amount of data [30]. In the model selection problem, the critical issue is the selection of the best model to use among a wide range of possibilities.

In the model selection literature, it is often assumed that the set of candidate models contains the true model [15]. However, since models are merely approximations to full reality, there is no such thing as the "true model", except in the case of Monte Carlo simulations where a model is used to generate data that mimics reality. George Box made the famous statement: "*All models are wrong but some are useful*". Multimodel inference tries to rank models relative to each other.

When we build a model, we are trying to minimize the loss of information. The Kullback-Leibler information, represent the information lost when approximating reality. Inference from multiple models using methods based on the Kullback–Leibler (K-L) information is preferred among other statistical methods [15]. In particular, null hypothesis testing approaches, which provide arbitrary dichotomies (*e.g.*, significant *vs.* nonsignificant), are particularly limited in model selection [31] and often perform poorly [32]. On the other hand, the use of subjective data inference often leads to overfitted models. Overfitted models contain too many parameters and should be avoided since they include noise as a structural part of the model. Conversely, underfitted models would ignore some important effects that are actually supported by the data.

The concept of parsimony is a fundamental philosophical issue in science and can be seen as the tradeoff between underfitting and overfitting. The principle of parsimony is closely related to Occam's razor. Occam's razor advocates to "*shave away all that is not needed*", Parsimony plays an important role in scientific thinking in general and in modeling in particular [33].

In the early 1970s Hirotugu Akaike presented the Akaike information criterion (AIC), a new paradigm for model selection in the analysis of empirical data. AIC, which is derived from information theory, is a relatively simple and easy to use scheme for selecting a parsimonious model for the analysis of empirical data. AIC establishes a fundamental relationship between Boltzmann's entropy and K-L information (dominant paradigms in information and coding theory), and maximum likelihood (the dominant paradigm in statistics) [34].

Information-theoretic criteria such as AIC based on K-L information and Bayesian Information Criterion (BIC) based on Bayes factors are now perceived as important measures to assess the quality of the fitting. AIC is often preferred over BIC because it has a more immediate connection to the theory of information [15]. AIC captures the trade-off between under and overfitting considering the principle of parsimony.

The "Akaike information criterion" or AIC is *an estimate of the expected, relative distance between the fitted model and the unknown true mechanism (perhaps of infinite dimension) that actually generated the observed data* [15].

Mathematically the Akaike information criterion is described as follows:

$$AIC = 2k + n\left(\log(2\pi) + \log\left(\sum_{u=1}^{k}\sum_{i=1}^{n}\left(\hat{X}_{i,u} - \overline{X}_{i,u}\right)^{2}\right) - \log \mathrm{n} + 1\right) \tag{6}$$

Where AIC denotes the Akaike information criterion, *k* is the number of estimated parameters plus one (the standard deviation σ^2) and *n* is the number of experimental data points.

When the ratio between the number of parameters to be estimated and the number of experimental data points is low (*i.e.*, n/k < ~40 [15]), it is recommended to use a corrective term, giving rise to the following corrected AICc expression:

$$AIC_C = AIC + \frac{2k(k+1)}{n-k-1} \tag{7}$$

Burnham and Anderson [15] strongly recommend using AICc in these instances because using the AIC increases the probability of selecting models that have too many parameters, (*i.e.*, overfitting).

4. <u>Identification of kinetic parameters and regulatory structures in nonlinear dynamic biological systems</u>

4.1 <u>Parameter estimation</u>

Parameter estimation has been viewed as an optimization problem for at least nine decades. The type of optimization problem being faced and the technical challenges to be solved depend upon the biological model of choice, upon the experimental data available, upon computational issues, and upon the specific mathematical formalism used.

The study of dynamic responses of biological systems is particularly appealing in systems biology. Dynamic biological systems are described through nonlinear ordinary differential equations (ODEs) that provide the concentration profiles of certain genes, proteins and metabolites over time. The biomathematical modeling of these systems gives rise to dynamic optimization problems which are hard to solve.

We consider dynamic parameter estimation optimization problems of the following form:

$$\begin{aligned} &\min_{\theta,\hat{z}_u} \sum_{j\in JM}\sum_{u\in U}\left(\hat{z}_{u,j}-\overline{z}_{u,j}\right)^2 \\ &s.t. \quad \dot{z}_j = g\left(z,\theta,t\right) \quad \forall j\in J \\ &\qquad z_j\left(t_0\right) = z_0 \quad \forall j \in J \\ &\qquad t\in\left[t_0,t_f\right] \\ &\qquad \hat{z}_{u,j} = z_j\left(t_u\right) \quad \forall u\in U; \quad \forall j \in JM \end{aligned} \tag{8}$$

Where $\dot{z}_j$ represents the state variables (*i.e.*, metabolite concentrations), z_0 their initial conditions, $\hat{z}_{u,j}$ represents the experimental data variables, $\overline{z}_{u,j}$ are the experimental observations, *J* is the set of state variables whose derivatives explicitly appear in the model, θ are the parameters to be estimated and t_u, is the time associated with the *u*th experimental data point in the set U. Our solution strategy relies on reformulating the nonlinear dynamic optimization problem as a finite dimensional NLP by applying a complete discretization using orthogonal collocation on finite elements.

The method devised for globally optimizing the NLP that arises from the reformulation of the parameter estimation problem (Eq. 8) is based on the outer approximation algorithm [23]. This approach relies on decomposing the original NLP into two subproblems at different hierarchical levels: a lower level based on a master MILP problem, and an upper level slave NLP problem. The master problem is a relaxation of the original NLP (*i.e.*, it overestimates its feasible region) and hence provides a rigorous lower bound on its global optimum. The slave NLP yields a valid upper bound when it is solved locally. The algorithm iterates between these two levels until the optimality gap (*i.e.*, the relative difference between the upper and lower bounds) is reduced below a given tolerance (see *Results and discussi*on section in [1] for further information). The capabilities of our algorithm are tested through its application to two case studies: the isomerisation of α-Pinene and the inhibition of HIV proteinase. The results obtained are compared with those produced by the state-of-art commercial global optimization package BARON. Our algorithm is proved from these numerical examples to produce near optimal solutions in a fraction of the CPU time required by BARON. (see table 2 and 3 in *Results and Discussion* in [1])

One of the key steps in this work is the selection of the appropriate mathematical representation. Kinetic models based on the so-called power-law formalism show a good compromise between accuracy and simplicity. Among them, the S-System and GMA representations are promising alternatives for the kinetic modeling of biological systems. The main advantage of these models is that they can capture the non-linearities required to describe the regulatory processes of the networks.

In particular, the GMA mathematical representation of a metabolic network containing *n* internal metabolites whose concentration varies due to the action of *p* flows can be expressed as follows:

$$\dot{X}_i = \sum_{r=1}^{p} \left(\mu_{i,r} \gamma_r \prod_{j=1}^{m+n} X_j^{f_{r,j}} \right) \quad i = 1,\ldots,n \tag{9}$$

where μ_{ir} is the stoichiometric coefficient of metabolite *i* in process *r*, γ_r is the basal-state enzyme activity, X_j corresponds to the concentration of metabolite *j* and f_{rj} is the kinetic order of metabolite X_j in process *r*, which quantifies its effect on the considered rate. Note that contributions of the *m* (independent) external metabolites are also accounted for in this representation.

4.2. Parameter estimation with GMA models

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:

$$\begin{aligned}
\min_{\gamma_r, f_{r,j}} \quad & \sum_{u=1}^{k} \sum_{i=1}^{n} \left(X_{i,u}^{\mathrm{exp}} - X_{i,u}^{\mathrm{mod}} \right)^2 \\
s.t. \quad & \dot{X}_j = \sum_{r=1}^{p} \mu_{i,r} v_r \quad i = 1,\ldots,n \\
& v_r = \gamma_r \prod_{j=1}^{n+m} X_j^{f_{r,j}} \;\; r = 1,\, \ldots,\, p \\
& \dot{X}_i \left(t_0 \right) = X_{0i} \quad i = 1,\ldots,n;\, t \in \left[t_0, t_f \right] \\
& X_{i,u}^{\mathrm{mod}} = X_i \left(t_u \right) \quad i = 1,\ldots,n; u = 1,\ldots,k
\end{aligned} \tag{10}$$

where $\dot{X}_i$ represents the state variables (*i.e.*, metabolite concentrations), X_{0i} their initial conditions, $X_{i,u}^{\exp}$ denotes the experimental observations, and $X_{i,u}^{\mathrm{mod}}$ 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.

4.3. Structure identification with GMA models

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 use the following disjunction:

$$\begin{bmatrix} Y_{r,j}^{-} \\ f_{r,j} \le -\varepsilon \end{bmatrix} \underline{\vee} \begin{bmatrix} Y_{r,j} \\ -\varepsilon \le f_{r,j} \le \varepsilon \end{bmatrix} \underline{\vee} \begin{bmatrix} Y_{r,j}^{+} \\ \varepsilon \le f_{r,j} \end{bmatrix} \quad j = 1,\dots,n; r = 1,\dots,p$$
$$Y_{r,j}^{-}, Y_{r,j}, Y_{r,j}^{+} \in \{True, False\} \tag{11}$$

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 convexhull reformulations [35]. By applying the former, we get:

$$\begin{array}{ll}
f_{r,j} \le -\varepsilon + M\left(1 - y_{r,j}^{-}\right) & j = 1,\ldots,n; r = 1,\ldots,p \\
-\varepsilon - M\left(1 - y_{r,j}\right) \le f_{r,j} \le \varepsilon + M\left(1 - y_{r,j}\right) & j = 1,\ldots,n; r = 1,\ldots,p \\
\varepsilon \le f_{r,j} + M\left(1 - y_{r,j}^{+}\right) & j = 1,\ldots,n; r = 1,\ldots,p \\
y_{r,j}^{-} + y_{r,j} + y_{r,j}^{+} = 1 & j = 1,\ldots,n; r = 1,\ldots,p \\
y_{r,j}^{-}, y_{r,j}, y_{r,j}^{+} \in \left\{0,1\right\} &
\end{array} \tag{12}$$

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.

We tested the performance of our algorithm for simultaneously estimating the parameters and regulatory topology of biochemical networks from dynamic time-series data through its application to a case study taken from Voit and Almeida [3]. In this approach a mixed-integer dynamic optimization (MIDO) problem was reformulated into an MINLP through the use of orthogonal collocation methods. The objective function of this MINLP was the minimization of the Akaike information criterion (AIC).

We explored an alternative approach to elucidate the network topology of a given biological system that consists of solving a bi-criteria optimization problem in which the complexity and the deviation from reality (i.e., the squared residual of the fitting of time series data) are simultaneously minimized. Mathematically, the parameter estimation task and the identification of the regulatory interactions are both posed in mathematical terms as a multi-objective mixed-integer dynamic optimization (moMIDO) problem. We solved this problem using the ε-constraint method which takes the following form:

$$
\begin{array}{ll}
\min\limits_{\gamma_r, f_{r,j}, y^-_{r,j}, y_{r,j}, y^+_{r,j}} & \sum\limits_{u=1}^{k}\sum\limits_{i=1}^{n}\left(\hat{X}_{i,u} - \bar{X}_{i,u}\right)^2 \\
s.t. & \sum\limits_{j=1}^{n}\sum\limits_{r=1}^{p}\left(y^-_{r,j} + y^+_{r,j}\right) \leq \varepsilon^e \quad e = 1,\ldots,E \\
 & \dot{X}_i = \sum\limits_{r=1}^{p}\mu_{i,r}v_r \quad i = 1,\ldots,n \\
 & v_r = \gamma_r \prod\limits_{j=1}^{n+m} X_j^{f_{r,j}} \; r = 1,\ldots,p \\
 & \dot{X}_i\left(t_0\right) = X_{0i} \quad i = 1,\ldots,n; t \in \left[t_0, t_f\right] \\
 & \bar{X}_{i,u} = X_i\left(t_u\right) \quad i = 1,\ldots,n; u = 1,\ldots,k \\
 & f_{r,j} \leq -\varepsilon + M\left(1 - y^-_{r,j}\right) \qquad j = 1,\ldots,n; r = 1,\ldots,p \\
 & -\varepsilon - M\left(1 - y_{r,j}\right) \leq f_{r,j} \leq \varepsilon + M\left(1 - y_{r,j}\right) \qquad j = 1,\ldots,n; r = 1,\ldots,p \\
 & \varepsilon \leq f_{r,j} + M\left(1 - y^+_{r,j}\right) \quad j = 1,\ldots,n; r = 1,\ldots,p \\
 & y^-_{r,j} + y_{r,j} + y^+_{r,j} = 1 \qquad j = 1,\ldots,n; r = 1,\ldots,p \\
 & y^-_{r,j}, y_{r,j}, y^+_{r,j} \in \left\{0,1\right\}
\end{array}
\tag{13}
$$

This MIDO model is reformulated into an equivalent multi-objective mixed-integer nonlinear programming MINLP (moMINLP) problem using orthogonal collocation on finite elements (See *Orthogonal collocation approach* section in [1] for further details).

5. Results

In this section we provide a brief description about the most relevant results obtained. Further details can be found in the original publications attached to this document. See *Case studies* section in [1] and *Results and Discussion* in [2] and [4].

5.1. Deterministic global optimization for parameter estimation

This work presented a deterministic global optimization method for the parameter estimation of biological systems. This approach allows globally optimizing medium-sized biological problems and provides a rigorous interval within which the global solution must fall. We tested its performance through its application to two case studies: the isomerisation of α-pinene and the inhibition of HIV.

5.1.1. The isomerisation of α-Pinene

This process was originally studied by Fuguitt and Hawkins [36], who carried out one single experiment for reporting the experimental concentrations (mass fraction) of the reactant and the four products measured at eight time intervals. In this homogeneous chemical reaction, α-pinene (γ_1 in Figure 2) is thermally isomerised to dipentene (γ_2) and allo-ocimene (γ_3), which in turn yields α- and β-pyronene (γ_4) and a dimer (γ_5) (Figure 2).

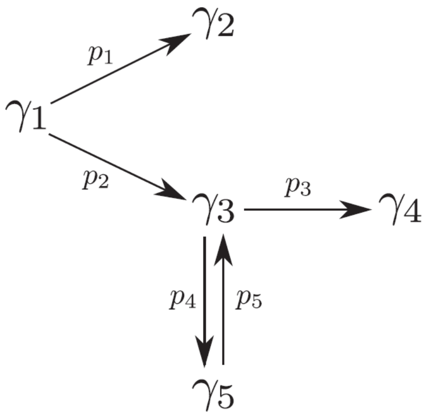

Figure 2 Proposed mechanism describing the thermal isomerization of α-Pinene

In this particular case study, the bounds on the collocation coefficients were tightened following a bound contraction procedure in order to reduce the space search. To solve this problem, an optimality gap of 5% was set as main termination criterion. For comparison purposes, we solved the same problem with the standard global optimization package BARON using its default settings. BARON was able to find the global optimum, but failed at reducing the optimality gap below the specified tolerance after 12h of CPU time. In contrast, our algorithm closed the gap in less than 3h (see Table 2 in *Results and Discussion* in [1]). The algorithm developed was able to find the same solution reported by the literature [19] providing in addition rigorous bounds on the global optimum.

5.1.2. Inhibition of HIV proteinase

This case study was originally examined by Kuzmic [37]. The enzyme HIV proteinase (E), which is only active in a dimer form, was added to a solution of an irreversible inhibitor (I) and a fluorogenic substrate (S). The product (P) is a competitive inhibitor for the substrate (Figure 3).

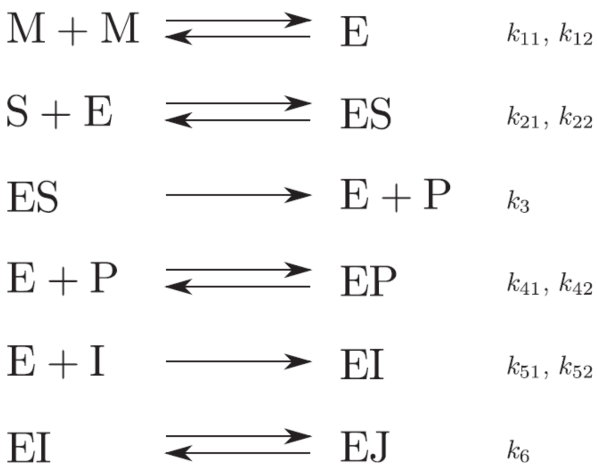
M + M ⇄ E k_{11}, k_{12}
S + E ⇄ ES k_{21}, k_{22}
ES → E + P k_3
E + P ⇄ EP k_{41}, k_{42}
E + I → EI k_{51}, k_{52}
EI ⇄ EJ k_6

Figure 3 Proposed mechanism describing the irreversible inhibition of HIV proteinase.

Mendes and Kell [38] [solved this problem using simulated annealing and reported its first known solution. Rodriguez-Fernandez et al. [19] improved that solution by means of a scatter search metaheuristic, which required a fraction of the time employed by Mendes' simulated annealing.

In this case study, the master problem was further tightened by adding a special type of strengthening cuts. These cuts were generated by temporally decomposing the original full space MILP into a series of MILPs in each of which we fitted only a subset of points of the original dataset, and remove the continuity equations corresponding to the extreme elements included in the sub-problem. The cuts were expressed as inequalities (which were added to the master problem) that impose lower bounds on the error of a subset of elements for which the sub-MILPs were solved. These bounds were hence obtained from the solution of a set of MILP sub-problems.

BARON failed to identify any feasible solution after 12h of CPU time. In contrast, our algorithm was able to obtain the global optimum (See Table 3 in *Results and Discussion* in [1]) with a gap of 18.64% in approximately 4,000 CPU s (See Table 4 in *Results and Discussion* in [1]). Remarkably, the solution found by our algorithm improves the best known solution reported by Rodriguez-Fernandez *et al.* [19]. Hence, our algorithm clearly outperformed other parameter estimation methods, improving the best known solution [19, 38], and providing a rigorous lower bound on the minimum error that can be attained.

As observed, solving medium-sized parameter estimation problems in systems biology to global optimality using deterministic methods is very challenging. Improving both the time required to solve them and the quality of the bounds attained are still open issues.

5.2. Identification of regulatory structure and kinetic parameters

The problem of identifying the regulatory structure and kinetic parameters of a biological system can be formally stated as follows: given a known structure of a reaction network (stoichiometry), and experimental time series data for the dynamic biological system, the goal is to determine the potential reaction and regulatory topologies for the target network along with the associated model parameters.

5.2.1. Branched pathway taken from Voit and Almeida (2004)

We have tested the capabilities of our approach through its application to a case study taken from Voit and Almeida [3]. 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 4).

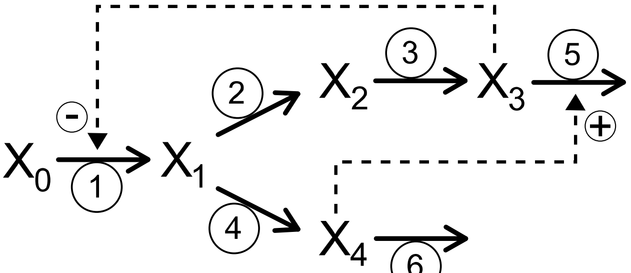

Figure 4 Reference system taken from Voit and Almeida [3] [Voit EO 2004].

We addressed in first place the traditional parameter estimation problem, assuming that the regulatory structure is known and error-free. As expected, we obtained estimated parameter values that are very close to the original ones (see Table 1 in *Results and Discussion* in [2]). We observed that GMA models have a certain degree of plasticity that allows different parameter sets to fit the same data. For instance, figure 2 in *Results and Discussion* in [2] shows the results of fixing f_{32} at different values and fitting the other parameters.

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, respectively. Despite obtaining different parameter values, the various fitted models lead to similar residuals. Although the regulatory structure is fixed, we obtained parameter values representing either positive or negative regulatory effects.

We later considered an alternative perturbation on the initial concentration of metabolite X_3 (0.2, 1.2, and 2.2) for the same values of Monte Carlo sampling. These perturbations force the system to move across different dynamic regimes, producing additional information that helps in the identification of appropriate parameter values. As expected, the estimated parameters are more consistent over the different experiments. They are also closer to the actual parameter set selected to generate the data. However, it is still possible to find solutions involving alternative regulatory topologies with good fit to data.

We later studied the ability of our method to identify the regulatory topology of the model. To this end, we explored 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%). 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 characterized *a priori*) as potential variables affecting each velocity. In addition, we introduced kinetic-order constraints corresponding to those substrates of a reaction which must be positive. The method identifies topologies that are quite close and which show very small residuals, but it is unable to uniquely identify the original topology. see Table S3 in Additional file in [2]: for a list of topologies generated and their associated kinetic parameters and residuals.

As before, we considered three different initial conditions for X_3 (0.2, 1.2, and 2.2). With these 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). See Table S4 in Additional file in [2]:

The simple example presented in this paper show that estimating parameters in dynamic kinetic models is very challenging. In this context, models based on the power-law formalism can greatly facilitate the estimation task.

5.3. Bi-criteria approach for the identification of regulatory structure and kinetic parameters

In this last publication we propose an alternative strategy to characterize biological pathways from time series data. This new approach addresses the problem from a different perspective, instead of minimizing the Akaike information criterion this method explores a wider range of alternatives which are identified by simultaneously minimizing the complexity and the deviation from reality (*i.e.*, the squared residual of the fitting of time series data) using a multi-objective approach.

5.3.1. Branched pathway taken from Voit and Almeida (2004)

We use as benchmark problem the same case study as in the previous publication, an artificial branched pathway taken from Voit and Almeida [3] [Voit EO 2004] (See section 5.2.1).

We first explore the performance of our method for one single experiment *(i.e.*, one single configuration for the initial conditions) and assuming that the regulatory structure is unknown. We consider noisy data that is generated from the *in silico* model assuming that the “true” dynamic profile (*i.e.*, the one generated from the *in silico* model free of error) follows a normal distribution with standard deviation values of 5, 10, 15 and 30% (with respect to the actual nominal values).

We used the ε-constraint method to generate 10 Pareto solutions with an increasing number of regulatory interactions for each level of uncertainty (5, 10, 15 and 30%). Each of these solutions are obtained by solving a single-objective problem in which the sum of squared deviations between experimental and simulated data is kept as main objective, while the number of regulatory interactions is transferred to an auxiliary constraint.

For comparison purposes, we solved the same problem using the single-optimization approach previously developed in the previous paper [2]. In this method, the AIC is minimized as unique objective. This algorithm is executed iteratively in order to produce a set of potential regulating configurations along with the corresponding parameter values. We generated 10 candidate models that are sorted according to their AIC values. We assessed their performance comparing both methods using a CV strategy and computing the AIC value for each model. We also computed the AIC and of the real metabolic network (i.e. the one we used to generate experimental data, without the noise) and depicted these points as blue triangles. Note that such last analysis is only possible when dealing with an academic problem like this, since in a real situation these points are unknowable because they represent reality itself. However in this case we want to exploit this possibility in order to assess how far the generated models are from reality.

We repeated the procedure taking into account three experiments. Thus, we changed the initial concentrations of X_3 (0.2, 1.2, and 2.2). These perturbations force the system to move across different dynamic regimes, producing additional information that constrains further the feasible network configurations.

In most cases (in both instances entailing one and three experiments) the minimization of the AIC as single objective produces models with better predictive accuracy, that is, models with lower residuals in the validation set (as well as lower AIC values). However, the Pareto points generated using the bi-objective approach show AIC values very close to the minimum AIC value computed by the single-objective optimization approach. Regarding to the real model, it has worse AIC and better CV values than other models obtained by the aforementioned methods, which is the final prove that the best AIC is not always the best model. Hence, the models corresponding to the Pareto solutions of the bi-objective optimization formulation cannot be directly discarded, since when models have similar AIC values, the model with the lowest AIC value may not be the best. This is because the AIC value is calculated with a limited number of points. If this calculation was carried out in the space of infinite samples, then the solution with minimum AIC value would be indeed the best one. Another reason is that there might be additional biological considerations to take into account when choosing a final model.

6. Conclusions

The numerical results obtained and the algorithms developed have provided us the knowledge to draw the following conclusions:

In the first work we have proposed a novel strategy for globally optimizing parameter estimation problems with embedded nonlinear dynamic systems. The method presented was tested through two challenging benchmark problems: the isomerisation of α-pinene and the inhibition of HIV proteinase. The proposed algorithm was able to identify the best known solution, which was originally reported by Rodriguez-Fernandez *et al.* [19], in the case of the α-pinene, and improved the best known one in the HIV proteinase case study. In both cases, rigorous lower bounds were provided on the global optimum, making it possible to determine the optimality gap of the solutions found. The method proposed produced promising results, surpassing the capabilities of BARON.

In the second 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. 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). The proposed method can contribute to fill the lack of information on the regulatory signals that are in play in a given metabolic scenario. The example presented in this paper show that models based on the power-law formalism are particularly appealing for the simultaneous parameter estimation and regulatory structure identification tasks.

In the third paper we have presented an alternative approach based on multi-objective optimization to simultaneously identify the regulatory interactions along with the kinetic parameters (assuming a kinetic representation) from time series data. Our method is a generalization of a previous method that focused on minimizing the AIC as unique criterion. The new alternative method solves a bi-objective model that seeks to minimize simultaneously the problem complexity and the residual. This bi-objective model is solved by the epsilon constraint method, providing as output a set of Pareto optimal models. Ideally, the model yielding the minimum AIC should be identified by the bi-criteria approach. In practice, this might not always hold because of two reasons.

On the one hand, the modified expression we need to use in this case for computing the AIC (i.e. the corrected AIC, also known as AICc) does not hold some of the properties that are required to establish the analogy between the AIC and the bi-criteria quality-complexity of the fit. On the other hand, we cannot guarantee the global optimality of the solutions calculated by the reformulated MINLPs, which are solved by a local optimizer (see *Methods section* in [4] for more information).

Both approaches, the single-objective and bi-objective one show different performance depending on the case study, and there is no clear winner. Sometimes the single-objective identifies the solution with minimum AIC value, but others the bi-objective does it instead.

In the academic example addressed in this manuscript we have had the opportunity to compare the models obtained with the two methods exposed with the real network (i.e., with reality itself). Results of this comparison highlight the fact that both methods are quite similar in terms of the reliability of the models generated and that the bottleneck for obtaining better models is the ability to manage bigger experimental samples that will lead to bigger and more complex problems to solve. Further, this analysis evidenced that, for small samples (in the training set), the model yielding the minimum AIC might not be the one showing the best performance under new experimental data since the real network obtained worse AIC values than other models obtained by the aforementioned methods.

Because of this, the minimum AIC model should be carefully revised and compared with other models with similar AIC values so as to select a final model to be used in practice. In this context, we recommend to calculate a set of models with low AICs and residuals, and further assess them taking into account their accuracy, complexity and additional biological knowledge of the system.

7. Future Work

Our future work will focus on making the global optimization approach more efficient through the use of tailored cutting planes and decomposition strategies. We can further tighten the search space of the original problem by adding additional constraints.

These constraints could be obtained by decomposing the master problem into a set of MILP sub-problems that would optimize the error of only a subset of elements.

The global optimization algorithm can also be improved through the use of hybridization of deterministic methods with stochastic approaches. Reducing the time required to generate tight bounds is still a major issue in deterministic global optimization. We can speed up our algorithm by applying stochastic approaches for attaining rigorous bounds on the solution of the master problem and hence on the global solution of the original problem.

In this thesis, we used the piecewise McCormick envelopes to relax the bilinear terms. Our methodology should be extended to other nonlinearities.

Our goal would be to develop a software package (*e.g.*, a toolbox in Matlab…) to automate the calculations, so our approach can be easily used by a wider community. This is a challenging task, since nonlinear models are hard to handle and typically require customized solution procedures.

Many biological systems exhibit saturable and cooperative interactions. The GMA formalism, however, cannot account for these phenomena. The next step will be to extend our method to more accurate and complex representations like the Saturable and Cooperative (SC) formalism.

Ideally, the bi-objective approach should identify the minimum AIC solution. In practice, however, this does not happen due to the presence of nonconvexities that lead to multiple local optima in which local optimizers might get trapped during the search. Hence, there is a clear need for developing more efficient global optimization tools to assess such studies.

The next step to build a biomathematical model, will involve the selection of the final model among the models that compete for the first position. When we have models with similar AIC values, the model with the lowest AIC value will not be always the best.

From the experience we acquired during the development of this thesis, we conclude that the inference of regulatory signals and the estimation of the associated parameters are both very challenging tasks. There is still much work to be done in this area, but we strongly believe that such an effort is worthy.

8. References

1. 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.

2. Guillén-Gosálbez G, Miró A, Alves R, Sorribas A, Jiménez L: Identification of regulatory structure and kinetic parameters of biochemical networks via mixed-integer dynamic optimization. *BMC Systems Biology* 2013, 7:113

3. Voit EO, Almeida J: Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics* 2004, 20:1670–1681.

4. Miró A, Pozo C, Guillén-Gosálbez G, Jiménez L: Bi-objective mixed-integer dynamic optimization approach for identifying the regulatory structure and kinetic parameters of biochemical networks. *BMC Systems Biology* 2014 (Submitted)

5. Polisetty, P, Gatzke, E, Voit, E: Yield optimization of regulated metabolic systems using deterministic branch-and-reduce methods. *Biotechnology and Bioengineering* 2008, 99(5), 1154–1169.

6. Chou I-C, Voit E O: Recent developments in parameter estimation and structure identification of biochemical and genomic systems *Math. Biosci*. 2009, 219 57–83

7. Marin-Sanguino A, Voit E, Gonzalez-Alzon C, Torres, N: Optimization of biotechnological systems through geometric programming. *Theoretical Biology & Medical Modelling* 2007, 4–38.

8. Voit E: Optimization in integrated biochemical systems. *Biotechnology and Bioengineering* 1992, 40(5), 572–582.

9. Voit E: Design principles and operating principles: the yin and yang of optimal functioning. *Mathematical Biosciences* 2003, 182, 81–92.

10. Alves R, Vilaprinyo E, Hernndez-Bermejo B, Sorribas A: Mathematical formalisms based on approximated kinetic representations for modeling genetic and metabolic pathways. *Biotechnology & Genetic Engineering Reviews* 2009, 25, 1–40.

11. Curto R, Sorribas A, Cascante M: Comparative characterization of the fermentation pathway of Saccharomyces cerevisiae using biochemical systems theory and metabolic control analysis: Model definition and nomenclature. *Mathematical Biosciences* 1995, 130, 25–50.

12. Voit, E: *Computational analysis of biochemical systems. A practical guide for biochemists and molecular biologists.* Cambridge 2000: Cambridge University Press.

13. Akaike H: New look at statistical-model identification. *IEEE Trans Automat* 1974, 19:716–723. Contr. AC19 716.

14. Schwarz G: Estimating the dimension of a model. *Ann Stat* 1978, 6:461–464.

15. 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.

16. Kameswaran S, Biegler L: Simultaneous dynamic optimization strategies: recent advances and challenges. *Comput & Chem Eng* 2006, 30(10-12):1560–1575.

17. Esposito W, Floudas C: Global optimization for the parameter estimation of differential-algebraic systems. *Ind & Eng ChemRes* 2000, 39(5):1291–1310.

18. Cizniar M, Salhi D, Fikar M, Latifi M: A MATLAB package for orthogonal collocations on finite elements in dynamic optimisation. In Proc *15 Int Conference Process Control*, Volume 5:058f.

19. Rodriguez-Fernandez M, Egea J, Banga J: Novelmetaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinf* 2006, 7:483.

20. Esposito W, Floudas C: Deterministic global optimization in nonlinear optimal control problems. *J Global Optimization* 2000, 17:97–126.

21. Papamichail I, Adjiman C: A rigorous global optimization algorithm for problems with ordinary differential equations. *J Global Optimization* 2002, 24:1–33.

22. Singer A, Barton P: Global solution of optimization problems with parameter-embedded linear dynamic systems. *J Optimization Theory and Appl* 2004, 121(3):613–646.

23. Kesavan P, Allgor R, Gatzke E, Barton P: Outer approximation algorithms for separable nonconvexmixed-integer nonlinear programs. *Math Programming* 2004, 100(3):517–535.

24. Vecchietti A, Sangbum L, Grossmann I: Modeling of discrete/continuous optimization problems: characterization and formulation of disjunctions and their relaxations. *Computers and chemical engineering* 2003, 27, 433-448.

25. Marin-Sanguino A, Voit E O, Gonzalez-Alcon C, Torres N V: Optimization of biotechnological systems through geometric programming. *Theor Biol Med Model* 2007, 4, 38.

26. Smith E, Pantelides C: A symbolic reformulation/spatial branch-and-bound algorithm for the global optimisation of nonconvex MINLPs. *Comput & Chem Eng* 1999, 23(4-5):457–478.

27. Cuthrell J, Biegler L: On the optimization of differential-algebraic process systems. *AIChE J* 1987, 33(8):1257–1270.

28. Tieu D, Cluett W, Penlidis A: A comparison of collocation methods for solving dynamic optimization problems. *Comput & Chem Eng* 1995, 19(4):375–381.

29. Ehrgott, M.: *Multicriteria Optimization*. Berlin: Springer, 1998.

30. deLeeuw, J. (1988). *Model selection in multinomial experiments*. Pages 118–138 in T. K. Dijkstra (ed.) *On model uncertainty and its statistical implications. Lecture Notes in Economics and Mathematical Systems*, Springer-Verlag, New York, NY.

31. Royall R M: Statistical evidence: a likelihood paradigm. *Chapman and Hall* 1997, London.

32. Akaike H: Likelihood of a model and information criteria. *Journal of Econometrics* 1981, 16, 3–14.

33. Forste, M R, Sober E: How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal of the Philosophy of Science* 1994, 45, 399–424.

34. deLeeuw J: Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. Pages 599–609 in S. Kotz, and N.L. Johnson (eds.) *Breakthroughs in statistics*. Vol. 1. 1992, Springer-Verlag, London.

35. 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.

36. Fuguitt R, Hawkins J: Rate of the thermal isomerization of α-Pinene in the liquid phase1. *J AmChemSoc* 1947, 69(2):319–322.

37. Kuzmic P: Program DYNAFIT for the analysis of enzyme kinetic data: application to HIV proteinase. *Anal Biochem* 1996, 237(2):260–273.

38. Mendes P, Kell D: Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 1998, 14(10):869.

Miró *et al. BMC Bioinformatics* 2012, **13**:90
http://www.biomedcentral.com/1471-2105/13/90

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Deterministic global optimization algorithm based on outer approximation for the parameter estimation of nonlinear dynamic biological systems

Anton Miró[1], Carlos Pozo[1], Gonzalo Guillén-Gosálbez[1*], Jose A Egea[2] and Laureano Jiménez[1]

Abstract

Background: The estimation of parameter values for mathematical models of biological systems is an optimization problem that is particularly challenging due to the nonlinearities involved. One major difficulty is the existence of multiple minima in which standard optimization methods may fall during the search. Deterministic global optimization methods overcome this limitation, ensuring convergence to the global optimum within a desired tolerance. Global optimization techniques are usually classified into stochastic and deterministic. The former typically lead to lower CPU times but offer no guarantee of convergence to the global minimum in a finite number of iterations. In contrast, deterministic methods provide solutions of a given quality (i.e., optimality gap), but tend to lead to large computational burdens.

Results: This work presents a deterministic outer approximation-based algorithm for the global optimization of dynamic problems arising in the parameter estimation of models of biological systems. Our approach, which offers a theoretical guarantee of convergence to global minimum, is based on reformulating the set of ordinary differential equations into an equivalent set of algebraic equations through the use of orthogonal collocation methods, giving rise to a nonconvex nonlinear programming (NLP) problem. This nonconvex NLP is decomposed into two hierarchical levels: a master mixed-integer linear programming problem (MILP) that provides a rigorous lower bound on the optimal solution, and a reduced-space slave NLP that yields an upper bound. The algorithm iterates between these two levels until a termination criterion is satisfied.

Conclusion: The capabilities of our approach were tested in two benchmark problems, in which the performance of our algorithm was compared with that of the commercial global optimization package BARON. The proposed strategy produced near optimal solutions (i.e., within a desired tolerance) in a fraction of the CPU time required by BARON.

Background

Elucidation of biological systems has gained wider interest in the last decade. Despite recent advances, fundamental understanding of life processes still requires powerful theoretical tools from mathematics and physical sciences. Particularly, mathematical modelling of biological systems is nowadays becoming an essential partner of experimental work. One of the most challenging tasks in computational modelling of biological systems is the estimation of the model parameters. The aim here is to obtain the set of parameter values that make the model response consistent with the data observed. Parameter estimation can be formulated as an optimization problem in which the sum of squared residuals between the measured and simulated data is minimized. The biological model dictates the type of optimization problem being faced. Many biological systems are described through nonlinear ordinary differential equations (ODEs) that provide the concentration profiles of certain metabolites over time. Recent methodological developments have enabled the generation of some dynamic profiles of gene networks

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

and protein expression data, although the latter are still very rare. In this context, there is a strong motivation for developing systematic techniques for building dynamic biological models from experimental data. The parameter estimation of these models gives rise to dynamic optimization problems which are hard to solve.

Existing approaches to optimize dynamic models can be roughly classified as direct or indirect (also known as variational) [1]. Direct methods make use of gradient-based nonlinear programming (NLP) solvers and can in turn be divided into sequential and simultaneous. In sequential approaches, the optimization of the control variables, which are discretized, is performed by a NLP solver, whereas the ODE is calculated externally, that is, both steps are executed in a sequential manner. In contrast, in simultaneous strategies, both the control and state profiles are approximated using polynomials (e.g., Lagrange polynomials) and discretized in time by means of finite elements [2,3]. In the latter strategy, the ODE system is replaced by a system of algebraic equations that is optimized with a standard gradient-based NLP solver. Simultaneous approaches can handle dynamic systems with unstable modes and with path constraints [1]. Furthermore, they allow performing automatic differentiation with respect to the control and state variables, avoiding the need to calculate the derivatives numerically as is the case in the sequential approach. Unfortunately, the discretization step can lead to large scale NLPs that are difficult to solve.

Models of biological systems are typically highly nonlinear, which gives rise to nonconvex optimization problems with multiple local solutions (i.e., multimodality). Because of this, traditional gradient-based methods used in the sequential and simultaneous approaches may fall in local optima. In the context of parameter estimation, these local solutions should be avoided, since they may lead to inaccurate models that are unable to predict the system's performance precisely.

Global optimization (GO) algorithms are a special class of techniques that attempt to identify the global optimum in nonconvex problems. These methods can be classified as stochastic and deterministic. Stochastic GO methods are based on probabilistic algorithms that provide near optimal solutions in short CPU times. Despite having shown great potential with large-scale problems like parameter estimation [4], these methods have as major limitation that are unable to guarantee convergence to the global optimum in a finite number of iterations. In other words, they provide solutions whose optimality (i.e., quality) is unknown, and may or may not be globally optimal. In contrast, deterministic global optimization methods ensure global optimality within a desired tolerance, but lead to larger computational burdens. Hence, in addition to the solution itself, these methods provide as output a rigorous interval within which the best possible solution (i.e., global optimum) must fall. Despite recent advances in deterministic global optimization methods [5,6], their application to parameter estimation has been quite scarce. Two main deterministic GO methods exist: spatial branch-and-bound (sBB) [2,5-7], and outer approximation [8]. Both algorithms rely on computing valid lower and upper bounds on the global optimum. These bounds tend to approach as iterations proceed, thus offering a theoretical guarantee of convergence to the global optimum.

A rigorous lower bound on the global optimum of the original nonconvex problem is obtained by solving a valid relaxation that contains its feasible space. To construct this relaxed problem, the nonconvex terms in the original formulation are replaced by convex envelopes that overestimate its feasible region. There are different types of convex envelopes that provide relaxations for a wide variety of nonconvexities. These relaxations are the main ingredient of deterministic GO methods and play a key role in their performance. In general, tighter relaxations provide better bounds (i.e., closer to the global optimum), thereby expediting the overall solution procedure.

To the best of our knowledge, Esposito and Floudas were the first to propose a deterministic method for the global solution of dynamic optimization problems with embedded ODEs [2]. Their approach relies on reformulating the problem as a nonconvex NLP using orthogonal collocation on finite elements. This reformulated NLP was then solved by means of a sBB method. To this end, they constructed a convex relaxation of the reformulated problem following the αBB approach previously proposed by the authors [5-7]. Despite being valid for twice continuous differentiable functions, these relaxations may provide weak bounds in some particular cases and therefore lead to large CPU times when used in the context of a spatial branch and bound framework [9].

This work proposes a computational framework for the deterministic global optimization of parameter estimation problems of nonlinear dynamic biological systems. The main contributions of our work are: (1) the application of deterministic global optimization methods to dynamic models of biological systems, and (2) the use of several known techniques employed in dynamic (i.e., orthogonal collocation on finite elements) and global optimization (i.e., symbolic reformulation of NLPs and piecewise McCormick envelopes) in the context of an outer approximation algorithm. The approach presented relies on discretizing the set of nonlinear ODEs using orthogonal collocation on finite elements, thereby transforming the dynamic system into an equivalent nonconvex NLP problem. A customized outer approximation algorithm that relies on a mixed-integer linear programming (MILP) relaxation is used in an iterative scheme along with the aforementioned NLP to solve the nonconvex model to global optimality. The MILP relaxation is tightened using a special type of cutting plane that exploits the problem structure, thereby expediting the overall solution procedure.

The capabilities of our algorithm are tested through its application to two case studies: the isomerisation of α-Pinene (case study 1) and the inhibition of HIV proteinase (case study 2). The results obtained are compared with those produced by the state-of-art commercial global optimization package BARON (Branch And Reduce Optimization Navigator). Our algorithm is proved from these numerical examples to produce near optimal solutions in a fraction of the CPU time required by BARON.

Methods

Problem statement

The problem addressed in this work can be stated as follows: given is a dynamic kinetic model describing the mechanism of a set of biochemical reactions. The goal is to determine the appropriate values of the model coefficients (e.g., rate constants, initial conditions, etc.), so as to minimize the sum-of-squares of the residuals between the simulated data provided by the model and the experimental observations.

Mathematical formulation

We consider dynamic parameter estimation optimization problems of the following form:

$$\min_{\boldsymbol{\theta},\hat{z}_u} \quad \sum_{j\in JM}\sum_{u\in U}(\hat{z}_{u,j}-\bar{z}_{u,j})^2 \tag{1}$$

$$\text{s.t.} \quad \dot{\boldsymbol{z}}_j = \boldsymbol{g}(\boldsymbol{z},\boldsymbol{\theta},t) \qquad \forall j \in J \tag{2}$$

$$z_j(t_0) = \boldsymbol{z}_0 \qquad \forall j \in J \tag{3}$$

$$t \in [\,t_0, t_f] \tag{4}$$

$$\hat{z}_{u,j} = z_j(t_u) \qquad \forall u \in U; \quad \forall j \in JM \tag{5}$$

Where $\dot{\boldsymbol{z}}$ represents the state variables (i.e., metabolite concentrations), $\boldsymbol{z}_0$ their initial conditions, $\hat{\boldsymbol{z}}_{u,j}$ represents the experimental data variables, $\bar{\boldsymbol{z}}_{u,j}$ are the experimental observations, J is the set of state variables whose derivatives explicitly appear in the model, $\boldsymbol{\theta}$ are the parameters to be estimated and t_u, is the time associated with the uth experimental data point in the set U.

Our solution strategy relies on reformulating the nonlinear dynamic optimization problem as a finite-dimensional NLP by applying a complete discretization using orthogonal collocation on finite elements. This NLP is next solved using an outer approximation algorithm (see Figure 1). In the sections that follow, we explain in detail the main steps of our algorithm.

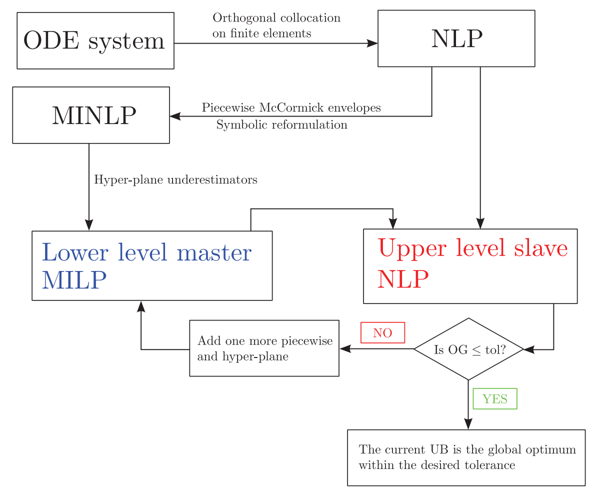
ODE system
Orthogonal collocation on finite elements
NLP
MINLP
Piecewise McCormick envelopes
Symbolic reformulation
Hyper-plane underestimators
Lower level master MILP
Upper level slave NLP
Add one more piecewise and hyper-plane
NO
Is OG ≤ tol?
YES
The current UB is the global optimum within the desired tolerance

Figure 1 Solution Strategy. The system of ODEs is first reformulated into a nonconvex NLP using the orthogonal collocation on finite elements approach. This NLP is decomposed into two levels: a master MILP and a slave NLP. The master MILP, which is constructed using piecewise McCormick envelopes and supporting hyper-planes, provides a rigorous lower bound on the global optimum. The slave NLP corresponds to the original nonconvex NLP that is solved using as starting point the solution of the MILP. The algorithm iterates between these two levels until the optimality gap (i.e., the relative difference between the upper and lower bounds) is reduced below a given tolerance.

Orthogonal collocation approach

There is a considerable number of collocation-based discretizations for the solution of differential-algebraic systems [10]. Without loss of generality, we employ herein the so-called orthogonal collocation on finite elements method [11,12]. Consider the following set of ODE's defined as

$$\dot{\boldsymbol{z}}_j = \boldsymbol{g}(\boldsymbol{z},\boldsymbol{\theta},t) \qquad \forall j \in J \tag{6}$$

The state variables are first approximated using Lagrange polynomials as follows:

$$z_{NK+1}(t) = \sum_{k=0}^{NK}\xi_k\phi_k(t) \qquad \phi_k(t) = \prod_{q=0,\,q\neq k}^{NK}\frac{t-t_q}{t_k-t_q} \tag{7}$$

These polynomials have the property that at the orthogonal collocation points their coefficients, ξ_k, take the value of the state profile at that point. Therefore, the collocation coefficients ξ_k acquire physical meaning which allows to generate bounds for these variables.

Because state variables may present steep variations, the whole solution space is commonly divided into time intervals called finite elements. Hence, the time variable t is divided into NE elements of length $\Delta\eta_e$ and rescaled as $\tau \in [0,1]$. Within each finite element, $NK+1$ orthogonal collocation points $\tau(0)$, $\tau(1)$, $\tau(2)$, $\cdots$, $\tau(NK)$ are distributed at the shifted (between 0 and 1) roots of the orthogonal Legendre polynomial of *NK* degree. Recall that the 0th orthogonal collocation point is located at the beginning of each element (Figure 2).

Following the collocation method [10], the residual equations arising from the combination of Eqs. 6 and 7, are defined for each element *e* in the set *E* and state variable in the set *J*, giving rise to the following constraints:

$$\sum_{k=0}^{NK} \xi_{e,k,j}\dot{\phi}_{e,k,j}(\tau_{k'}) - \Delta\eta_e g_j(\xi_{e,k',j},\, \boldsymbol{\theta},\, t_{e,k'}) = 0$$

$$\forall e \in E \quad k' = 1, \ldots, NK; \quad \forall j \in J \tag{8}$$

The state variables have to be continuous between elements, so we enforce the following continuity constrains:

$$\xi_{e,0,j} - \sum_{k=0}^{NK} \xi_{e-1,k,j}\phi_k(\tau = 1) = 0 \quad e = 2, \ldots, NE \quad \forall j \in J \tag{9}$$

These equations extrapolate the polynomial at element *e*-1, providing an accurate initial condition for the next element *e*.

Moreover, initial conditions are enforced for the beginning of the first element using the following equation:

$$\xi_{1,0,j} - z_{0,j} = 0 \qquad \forall j \in J \tag{10}$$

Recall that collocation points in which time has been discretized will not necessarily match the times at which experimental profiles were registered. Hence, variable $\hat{z}_{u,j}$ is added to determine the value of the model states profiles at times t_u making it possible to fit the model to the experimental points. This is accomplished by adding the following equation:

$$-\hat{z}_{u,j} + \sum_{k=0}^{NK} \xi_{e_u,k,j}\phi_k(\tau_u) = 0 \qquad \forall u \in U; \qquad \forall j \in JM \tag{11}$$

Where τ_u is calculated as follows:

$$\tau_u = \frac{t_u - \eta_{e_u}}{\Delta\eta_{e_u}} \tag{12}$$

Here, the subscript e_u refers to the element where t_u falls, that is, $e_u \equiv \{e : \eta_e \leq t_u < \eta_{e+1}\}$.

NPL formulation

The dynamic optimization problem is finally reformulated into the following NLP:

$$\min_{\boldsymbol{\theta},\xi,\hat{\boldsymbol{z}}_u} \qquad \sum_{j \in JM}\sum_{u \in U}(\hat{z}_{u,j} - \bar{z}_{u,j})^2 \tag{13}$$

$$\text{s.t.} \quad \sum_{k=0}^{NK} \xi_{e,k,j}\dot{\phi}_{e,k,j}(\tau_{k'}) - \Delta\eta_e g_j(\xi_{e,k',j},\, \boldsymbol{\theta},\, t_{e,k'}) = 0$$

$$\forall e \in E \qquad k' = 1, \ldots, NK; \qquad \forall j \in J \tag{14}$$

$$\xi_{e,0,j} - \sum_{k=0}^{NK} \xi_{e-1,k,j}\phi_k(\tau = 1) = 0$$

$$e = 2, \ldots, NE \qquad \forall j \in J \tag{15}$$

$$\xi_{1,0,j} - z_{0,j} = 0 \qquad \forall j \in J \tag{16}$$

$$-\hat{z}_{u,j} + \sum_{k=0}^{NK} \xi_{e_u,k,j}\phi_k(\tau_u) = 0$$

$$\forall u \in U; \qquad \forall j \in JM \tag{17}$$

Results and discussion

Optimization approach

The method devised for globally optimizing the NLP that arises from the reformulation of the parameter estimation problem (Eqs. 13–17) is based on an outer approximation algorithm [8] used by the authors in previous works [13-17]. This approach relies on decomposing the original NLP into two subproblems at different hierarchical levels: a lower level MILP problem and an upper level slave NLP problem. The master problem is a relaxation of the original NLP (i.e., it overestimates its feasible region) and hence provides a rigorous lower bound on its global optimum. The slave NLP yields a valid upper bound when it is solved locally. The algorithm iterates between these two levels until the optimality gap (i.e., the relative difference between the upper and lower bounds) is reduced below a given tolerance (Figure 3). In the following subsections, we provide a detailed description of the algorithm.

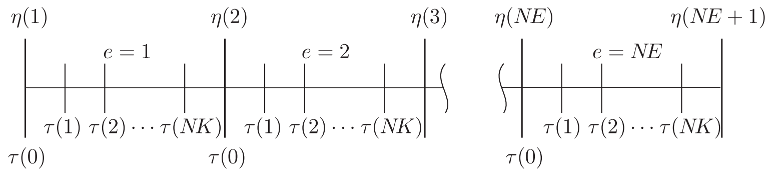
$\eta(1)$
$\eta(2)$
$\eta(3)$
$\eta(NE)$
$\eta(NE+1)$
$e = 1$
$e = 2$
$e = NE$
$\tau(1)$ $\tau(2) \cdots \tau(NK)$
$\tau(1)$ $\tau(2) \cdots \tau(NK)$
$\tau(1)$ $\tau(2) \cdots \tau(NK)$
$\tau(0)$
$\tau(0)$
$\tau(0)$

Figure 2 Orthogonal collocation discretization over finite elements. The time interval is divided into *NE* elements which in turn are divided into *NK* + 1 collocation points evaluated at the shifted orthogonal Legendre polynomials.

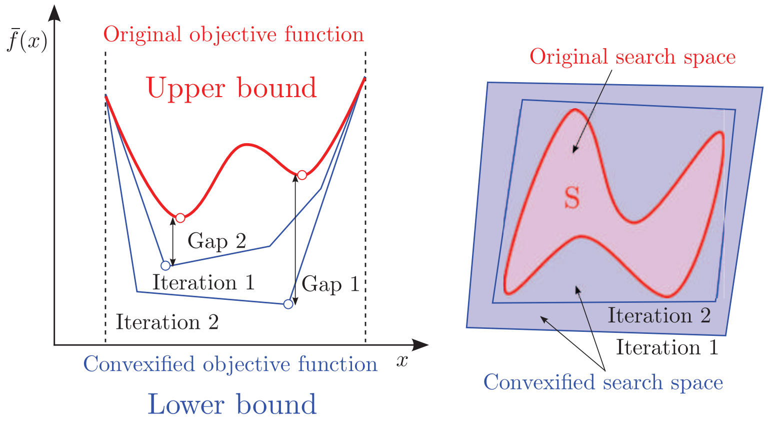

Figure 3 Optimization algorithm based on outer approximation. Our approach decomposes the problem into two subproblems: a master MILP, constructed by relaxing the original model using piecewise McCormick envelopes and hyper-planes, that provides a lower bound, and a slave NLP that yields an upper bound. The algorithm iterates between these two levels until a termination criterion is satisfied.

Lower level master problem

Designing efficient and smart strategies for attaining tight bounds is a mayor challenge in deterministic global optimization. Both the quality of the bounds and the time required to generate them drastically influence the overall performance of a deterministic global optimization algorithm.

Any feasible solution of the original NLP is a valid upper bound and can be obtained by means of a local NLP solver. To obtain lower bounds, we require a rigorous convex (linear or nonlinear) relaxation. This relaxation is obtained by replacing the nonconvex terms by convex overestimators. Since the relaxed problem is convex, it is possible to solve it to global optimality using standard local optimizers. Furthermore, since its feasible region contains that of the original problem and its objective function rigorously underestimates the original one, it is guaranteed to provide a lower bound on the global optimum of the original nonconvex model [18].

Androulakis et al. [19] proposed a convex quadratic relaxation for nonconvex functions named αBB underestimator which can be applied to general twice continuously differentiable functions. This technique, which was used in parameter estimation by Esposito and Floudas [2], might lead in some cases to weak relaxations and therefore poor numerical performance [9].

To construct a valid MILP relaxation, we apply the following approach. We first reformulate the NLP using the symbolic reformulation method proposed by Smith and Pantelides [20]. This technique reformulates any system of nonlinear equations into an equivalent canonical form with the only nonlinearities being bilinear products, linear fractional, simple exponentiation and univariate function terms with the following standard form:

$$\min_{w} \quad w_{\text{obj}} \tag{18}$$

$$\text{s.t.} \quad Aw = b \tag{19}$$

$$w^l \leq w \leq w^u \tag{20}$$

$$y \in [y^l, \ldots, y^u] \tag{21}$$

$$w_k \equiv w_i w_j \qquad \forall (i,j,k) \in \mathcal{T}_{\text{bt}} \tag{22}$$

$$w_k \equiv \frac{w_i}{w_j} \qquad \forall (i,j,k) \in \mathcal{T}_{\text{lft}} \tag{23}$$

$$w_k \equiv w_i^n \qquad \forall (i,k,n) \in \mathcal{T}_{\text{et}} \tag{24}$$

$$w_k \equiv \text{fn}(w_i) \qquad \forall (i,k) \in \mathcal{T}_{\text{uft}} \tag{25}$$

where vector w comprises continuous variables x as well as integers y, while the sets $\mathcal{T}_{\text{bt}}$, $\mathcal{T}_{\text{lft}}$, $\mathcal{T}_{\text{et}}$ and $\mathcal{T}_{\text{uft}}$ are the bilinear product, linear fractional, simple exponentiation and univariate function terms, respectively.

A rigorous relaxation of the original model is constructed by replacing the nonconvex terms in the reformulated model by convex estimators. The solution of the convex relaxation provides a valid lower bound on the global optimum. More precisely, the bilinear terms are replaced by piecewise McCormick relaxations. The fractional terms can be convexified in two different manners. The first is to replace them by tailored convex envelopes that exploit their structure [21]. The second is to express them as bilinear terms by performing a simple algebraic transformation, and then use the McCormick envelopes to relax the associated bilinear term. Univariate functions commonly used in process engineering models (e.g., logarithms, exponentials, and square roots) are purely convex or purely concave, and can be replaced by the exact function-secant pair estimators [22].

The reader is referred to the work by Smith and Pantelides [20] for further details on the symbolic reformulation. We focus next on explaining how the bilinear terms are approximated in the reformulated NLP.

Piecewise McCormick-based relaxation The bilinear terms appearing in the reformulated model are approximated using McCormick's envelopes [23-26]. For bilinear terms, this relaxation is tighter than the αBB-based relaxations [18,27].

Each bilinear term *xy* can be replaced by an auxiliary variable *z* as follows:

$$z = xy \qquad x^L \le x \le x^U \qquad y^L \le y \le y^U \tag{26}$$

The best known relaxation for approximating a bilinear term is given by the McCormick envelopes, obtained by replacing Eq. 26 by the following linear under (Eqs. 27 and 28), and overestimators (Eqs. 29 and 30):

$$z \ge xy^L + x^L y - x^L y^L \tag{27}$$

$$z \ge xy^U + x^U y - x^U y^U \tag{28}$$

$$z \le xy^L + x^U y - x^U y^L \tag{29}$$

$$z \le xy^U + x^L y - x^L y^U \tag{30}$$

In this work we further tighten the McCormick envelopes by adding binary variables [25,28]. Particularly, two additional sets of variables are defined in the piecewise formulation:

- Binary switch: $\lambda \in \{0, 1\}^{N_P}$
- Continuous switch: $\Delta y \in [0, y^U - y^L]^{N_P}$

The binary switch λ is active (i.e., $\lambda(n_P) = 1$) for the segment where x is located $(x^L + a(n_P - 1) \le x \le x^L + an_P)$ and is otherwise inactive. Therefore, the partitioning scheme activates exactly only one $n_P \in \{1, \ldots, N_P\}$ so that the feasible region corresponding to the relaxation of *xy* is reduced from the parallelogram in Figure 4(a) to a significantly smaller one depicted in Figure 4(b).

Eq. 31 enforces that only one binary variable is active:

$$\sum_{n_P=1}^{N_P} \lambda(n_P) = 1 \tag{31}$$

The continuous switch Δy takes on any positive value between 0 and $y^U - y^L$ when the binary switch corresponding to the n_Pth piecewise $\lambda(n_P)$ is active (i.e., $\lambda(n_P) = 1$) and 0 otherwise. Therefore:

$$y = y^L + \sum_{n_P=1}^{N_P} \Delta y(n_P) \tag{32}$$

$$0 \le \Delta y(n_P) \le (y^U - y^L)\lambda(n_P) \qquad n_P = 1, \ldots, N_P \tag{33}$$

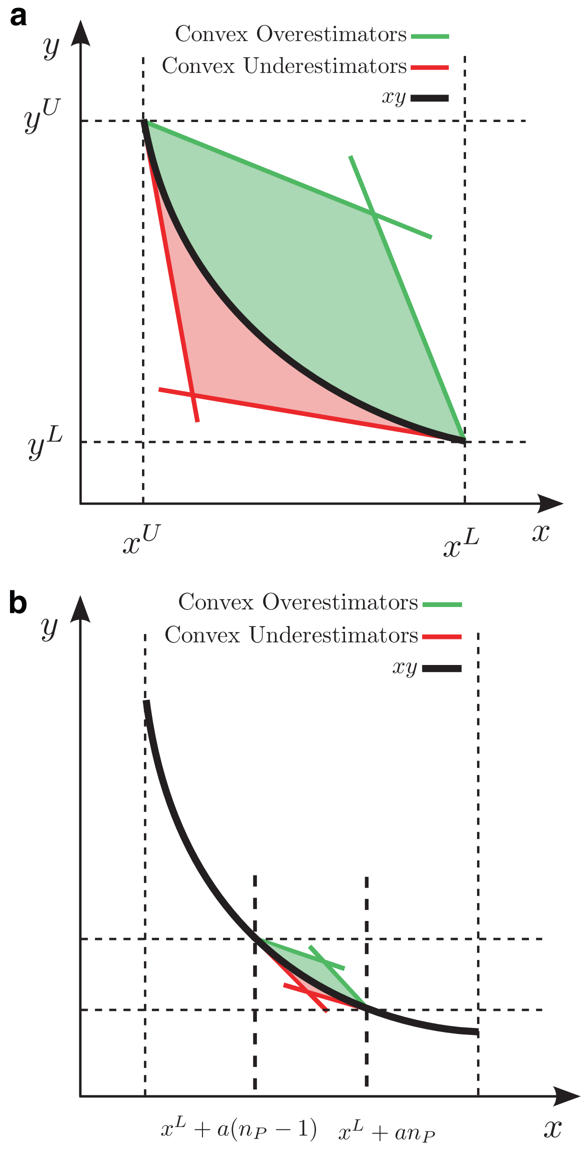

Figure 4 McCormick convex relaxation over the entire feasible region (subfigure (a)) compared to a piecewise McCormick relaxation over a smaller active region (subfigure (b)) where the tightness of the relaxation is improved. We built the master problem by replacing the bilinear terms by piecewise McCormick envelopes. The relaxation can be further improved by adding binary variables.

Finally, the under and overestimators for the active segment are defined in algebraic terms as follows:

$$z \ge xy^L + \sum_{n_P=1}^{N_P} [x^L + a(n_P - 1)]\, \Delta y(n_P) \tag{34}$$

$$z \ge xy^U + \sum_{n_P=1}^{N_P} [x^L + an_P]\, [\Delta y(n_P) - (y^U - y^L)\lambda(n_P)] \tag{35}$$

$$z \leq xy^L + \sum_{n_P=1}^{N_P} [x^L + an_P]\, \Delta y(n_P) \tag{36}$$

$$z \leq xy^U + \sum_{n_P=1}^{N_P} [x^L + a(n_P - 1)]\, [\, \Delta y(n_P) - (y^U - y^L) \times \lambda(n_P)] \tag{37}$$

$$x^L \leq x \leq x^U; \qquad y^L \leq y \leq y^U \tag{38}$$

Note that the discrete relaxation is tighter than the continuous one over the entire feasible region. The introduction of the binary variables required in the piecewise McCormick reformulation gives rise to a mixed-integer nonlinear programming (MINLP) problem, with the only nonlinearities appearing in the objective function. While this MINLP is convex and can be easily solved to global optimality with standard MINLP solvers, it is more convenient to linearize it in order to obtain an MILP formulation, for which more efficient software packages exist. The section that follows explains how this is accomplished.

Hyper-planes underestimation The convex MINLP can be further reformulated into an MILP by replacing the objective function by a set of hyper-planes. For this, we define two new variables as $z'_{u,j} = \hat{z}_{u,j} - \bar{z}_{u,j}$ and $\alpha \geq z'^{2}_{u,j}$. The quadratic terms are then approximated by 1st degree Taylor series. That is, the square terms are replaced by l hyper-planes uniformly distributed between the maximum and minimum desired values of $z'_{u,j}$ (Figure 5) so that the objective function is reduced to a summation of quadratic terms as follows:

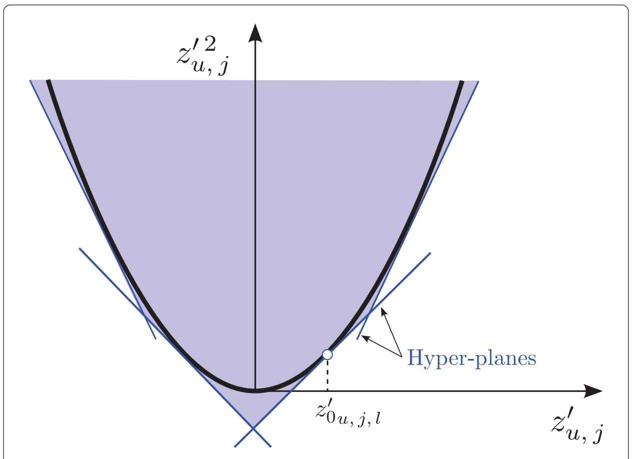

Figure 5 ***x* squared function underestimated by a 1st degree Taylor series.** The objective function is linearized by a first degree Taylor series with *l* hyper-planes.

$$\min_{\theta,\xi,\hat{z}_u} \quad \sum_{j\in JM}\sum_{u\in U} \alpha_{u,j} \tag{39}$$

$$\alpha_{u,j} \geq z'^{2}_{0\;u,j,l} + 2z'_{0u,j,l}(z'_{u,j} - z'_{0u,j,l})$$

$$\forall u \in U \qquad \forall j \in JM \qquad \forall l \in L \tag{40}$$

Upper level slave problem

A valid upper bound on the global optimum is obtained by optimizing the original NLP locally. This NLP is initialized using the solution provided by the MILP as starting point. The solution of this NLP is used to tighten the MILP, so the lower and upper bounds tend to converge as iterations proceed.

Algorithm steps

The proposed algorithm comprises the following steps:

1. Set iteration count it = 0, UB = ∞, LB = $-\infty$ and tolerance error = tol.
2. Set it $=$ it $+$ 1. Solve the master problem MILP.
   (a) If the MILP is infeasible, stop (since the NLP is also infeasible).
   (b) Otherwise, update the current LB making LB $=$ max$_{\text{it}}$(LB$_{\text{it}}$), where LB$_{\text{it}}$ is the value of the objective function of the MILP in the it$^{\text{th}}$ iteration.
3. Solve the slave problem NLP.
   (a) If the NLP is infeasible add one more piecewise term and hyper-plane to the master MILP and go to step 2 of the algorithm.
   (b) Otherwise, update the current UB making UB = min$_{\text{it}}$ (UB$_{\text{it}}$), where UB$_{\text{it}}$ is the value of the objective function of the NLP in the it$^{\text{th}}$ iteration.
4. Calculate the optimality gap OG as OG $= \frac{|\text{UB}-\text{LB}|}{\text{UB}}$.
   (a) If OG $\leq$ tol, then stop. The current UB is regarded as the global optimum within the desired tolerance.
   (b) Otherwise, add one more piecewise section and hyper-plane to the master MILP and go to step 2 of the algorithm.

Remarks:

- There are different methods to update the piecewise bilinear approximation. One possible strategy is to update it by dividing the active piecewise (i.e., the piecewise term in which the solution is located) into two equal-length segments.
- The new hyper-plane term $z'_{0u,j,l}$ is added at the optimal solution of the MILP (solution point $z'_{u,j}$) in the previous iteration.
- The univariate convex and concave terms in the reformulated problem can be either approximated by

the secant or by a piecewise univariate function similarly as done with the McCormick envelopes.

- Our algorithm needs to be tuned prior to its application. This is a common practice in any optimization algorithm. In a previous publication [13], we studied the issue of defining the number of piecewise intervals and supporting hyper-planes in an optimal manner. In practice, however, the optimal number of piecewise terms and hyper-planes is highly dependent on the specific instance being solved, so it is difficult to provide general guidelines on this.
- The approach presented might lead to large computational burdens in large-scale models of complex biological systems. Future work will focus on expediting our algorithm through the addition of cutting planes and the use of customized decomposition strategies.

Case studies

We illustrate the performance of the proposed algorithm through its application to two challenging benchmark parameter estimation problems: the isomerisation of α-Pinene (case study 1) and the inhibition of HIV proteinase (case study 2). The objective in these problems is to obtain the set of values of the model parameters such that the model response is as close as possible to the experimental data. For comparison purposes we used the global optimization package BARON (version 8.1.5). BARON is a commercial software for solving nonconvex optimization problems to global optimality. BARON combines constraint propagation, interval analysis, duality, and enhanced "branch and bound" concepts for efficient range reduction with rigorous relaxations constructed by enlarging the feasible region and/or underestimating the objective function. The interested readers have the possibility to evaluate this software on their own for free in this link: http://www.neos-server.org/neos/solvers/go:BARON/GAMS.html. Our algorithm was implemented in GAMS 23.5.2 using CPLEX 12.2.0.0 for the MILPs and SNOPT 4 for the NLPs subproblems. All the calculations were performed in a PC/AMD Athlon II at 2.99 Ghz using a single core. Data about the size of the models can be found in Table 1.

Case study 1: Isomerisation of α-Pinene

In this first case study, five kinetic parameters describing the thermal isomerisation of α-Pinene are estimated. The proposed reaction scheme for this process is depicted in Figure 6. In this homogeneous chemical reaction, α-Pinene (γ_1) is thermally isomerised to dipentene (γ_2) and allo-ocimene (γ_3), which in turn yields α- and β-Pyronene (γ_4) and a dimer (γ_5). This process was originally studied by Fuguitt and Hawkins [29], which carried out a single experiment reporting the experimental concentrations (mass fraction) of the reactant and the four products measured at eight time intervals.

Table 1 Model size in the last iteration

	Isomerisation of α-Pinene	Inhibition of HIV proteinase
MILP equations	1,836	138,128
MILP continuous variables	1,096	53,321
MILP binary variables	380	3,625
NLP equations	186	16,306
NLP variables	196	16,361

Hunter and McGregor [30] postulated first-order kinetics and proposed the following set of ODE's describing the dynamic process:

$$\frac{d\gamma_1}{dt} = -(p_1 + p_2)\gamma_1 \tag{41}$$

$$\frac{d\gamma_2}{dt} = p_1\gamma_1 \tag{42}$$

$$\frac{d\gamma_3}{dt} = p_2\gamma_1 - (p_3 + p_4)\gamma_3 + p_5\gamma_5 \tag{43}$$

$$\frac{d\gamma_4}{dt} = p_3\gamma_3 \tag{44}$$

$$\frac{d\gamma_5}{dt} = p_4\gamma_3 - p_5\gamma_5 \tag{45}$$

$$\gamma_0 = [\,100,\ 0,\ 0,\ 0,\ 0] \qquad t \in [\,0,\ 36420] \tag{46}$$

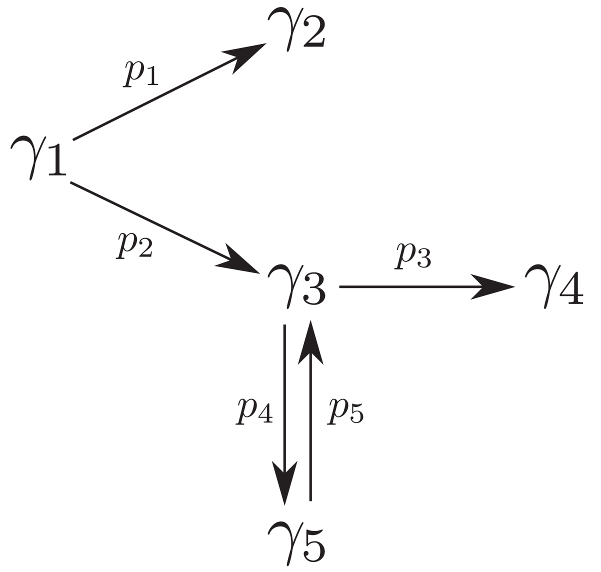

Figure 6 Proposed mechanism describing the thermal isomerization of α-Pinene. In this reaction α-Pinene (γ_1) is thermally isomerized to dipentene (γ_2) and allo-ocimene (γ_3), which in turn yields α- and β-Pyronene (γ_4) and a dimer (γ_5).

Rodriguez-Fernandez et al. [4] addressed this problem by applying a metaheuristic based on the scatter search method. This strategy does not offer any theoretical guarantee of convergence to the global optimum in a finite number of iterations.

Following our approach, the state variables were approximated by Lagrange polynomials using three collocation points evaluated at the shifted roots of orthogonal Legendre polynomials and defining five finite elements of equal length. The nonconvexities in the resulting residual equations are given by the bilinear terms $\theta_i \xi_{e,k,j}$ which were relaxed using piecewise McCormick approximations as described previously. The objective function was underestimated using supporting hyper-planes.

It is well known that the quality of the lower bound predicted by a relaxation strongly depends on the bounds imposed on its variables [31]. Hence bounds on collocation coefficients ($\xi^L_{e,k,j}$ and $\xi^U_{e,k,j}$, originally set to 0 and 100, respectively) were tightened by performing a bound contraction procedure [21,32]. Particularly, tight lower and upper bounds were estimated for each collocation coefficient by maximizing and minimizing its value while satisfying the constraints contained in the master problem. This is a costly process (i.e., if bounds for n variables are to be estimated, $2n$ optimization problems should be solved). For this reason, it was only performed recursively 3 times before the initialization of the algorithm. The MILP was further tightened by adding the following constraint:

$$\sum_{j \in JM} \sum_{u \in U} (\hat{z}_{u,j} - \bar{z}_{u,j})^2 \leq 20 \tag{47}$$

which forces the model to find a solution better than the one obtained at the beginning of the search by locally minimizing the original NLP (i.e., 20 is a rigorous upper bound for the objective function). Furthermore, the parameter θ_i was allowed to take any value within the [0, 1] interval.

The problem was solved with 6 initial hyper-planes. An extra hyper-plane was added in each iteration, but the total number of piecewise terms was kept constant (4 piecewise intervals were considered) in order to keep the MILP in a manageable size. A tolerance of 5% was set as termination criterion.

For comparison purposes, we solved the same problem with the standard global optimization package BARON using its default settings. BARON was able to find the global optimum but failed at reducing the optimality gap below the specified tolerance after 12h of CPU time. In contrast, our algorithm closed the gap in less than 3h (see Table 2). As shown in Table 2, the results obtained agree with those reported in the literature.

Table 2 Global optimization results for the α-Pinene isomerisation problem

	Rodriguez-Fernandez et al.	BARON	Proposed algorithm
Sum of squares	19.87	19.87	19.87
UB	-	19.87	19.87
LB	-	4.112	19.26
Gap (%)	-	79.31	3.056
Iterations	9,518	60,614	2
Time (CPU s)	122	43,200	8,916

Case study 2: Inhibition of HIV proteinase

In this second case study, we considered a much more complex biological dynamic system. Particularly, we studied the reaction mechanism of the irreversible inhibition of HIV proteinase, as originally examined by Kuzmic [33] (Figure 7). Note that this dynamic model has lack of practical identifiability, as reported in Rodriguez-Fernandez et al [4]. Nevertheless, we think that this example is still useful for the purpose of our analysis, since the emphasis here is placed on globally optimizing dynamic models of biological systems rather than analyzing identifiability issues.

The model can be described mathematically through a set of 9 nonlinear ODE's with ten parameters:

$$\frac{d[\mathrm{M}]}{dt} = -2k_{11}[\mathrm{M}][\mathrm{M}] + 2k_{12}[\mathrm{E}] \tag{48}$$

$$\frac{d[\mathrm{P}]}{dt} = k_3[\mathrm{ES}] - 2k_{41}[\mathrm{P}][\mathrm{E}] + 2k_{42}[\mathrm{EP}] \tag{49}$$

$$\frac{d[\mathrm{S}]}{dt} = -k_{21}[\mathrm{S}][\mathrm{E}] + k_{22}[\mathrm{ES}] \tag{50}$$

$$\frac{d[\mathrm{I}]}{dt} = -k_{51}[\mathrm{I}][\mathrm{E}] + k_{52}[\mathrm{EI}] \tag{51}$$

$$\frac{d[\mathrm{ES}]}{dt} = k_{21}[\mathrm{S}][\mathrm{E}] - k_{22}[\mathrm{ES}] - k_3[\mathrm{ES}] \tag{52}$$

$$\frac{d[\mathrm{EP}]}{dt} = k_{41}[\mathrm{P}][\mathrm{E}] - k_{42}[\mathrm{EP}] \tag{53}$$

$$\begin{aligned}\frac{d[\mathrm{E}]}{dt} &= k_{11}[\mathrm{M}][\mathrm{M}] - k_{12}[\mathrm{E}] - k_{21}[\mathrm{S}][\mathrm{E}] + k_{22}[\mathrm{ES}] \\ &\quad + k_3[\mathrm{ES}] - k_{41}[\mathrm{P}][\mathrm{E}] + k_{42}[\mathrm{EP}] - k_{51}[\mathrm{I}][\mathrm{E}] \\ &\quad + k_{52}[\mathrm{EI}]\end{aligned} \tag{54}$$

$$\frac{d[\mathrm{EI}]}{dt} = k_{51}[\mathrm{I}][\mathrm{E}] - k_{52}[\mathrm{EI}] - k_6[\mathrm{EI}] \tag{55}$$

$$\frac{d[\mathrm{EJ}]}{dt} = k_6[\mathrm{EI}] \tag{56}$$

where the following initial conditions and parameters are known:

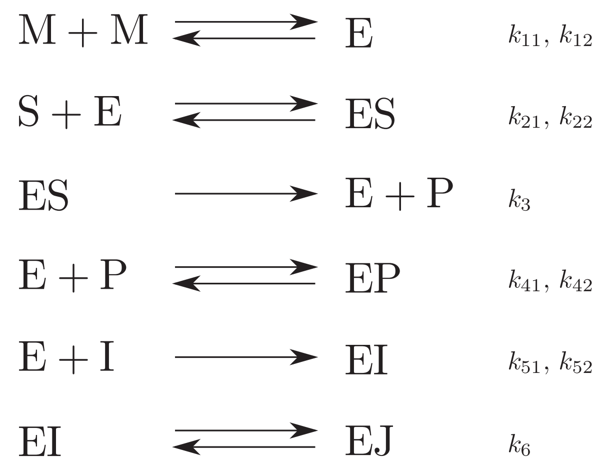
M + M ⇄ E k_{11}, k_{12}
S + E ⇄ ES k_{21}, k_{22}
ES → E + P k_3
E + P ⇄ EP k_{41}, k_{42}
E + I → EI k_{51}, k_{52}
EI ⇄ EJ k_6

Figure 7 Proposed mechanism describing the irreversible inhibition of HIV proteinase. The enzyme HIV proteinase (E), which is only active in a dimer form, was added to a solution of an irreversible inhibitor (I) and a fluorogenic substrate (S). The product (P) is a competitive inhibitor for the substrate.

Table 3 Optimal parameters for the HIV proteinase inhibition problem

Parameter	Rodriguez-Fernandez et al.	Proposed algorithm
Sum of squares	0.01997	0.01961
k_3 (s^{-1})	6.235	5.764
k_{42} (s^{-1})	8,772	968.7
k_{22} (s^{-1})	473	129.9
k_{52} (s^{-1})	0.09726	0.01612
k_6 (s^{-1})	0.01417	0.01337
S_0 exp. 1 (μM)	24.63	24.61
S_0 exp. 2 (μM)	23.32	23.4
S_0 exp. 3 (μM)	26.93	27.05
S_0 exp. 4 (μM)	13.34	13.97
S_0 exp. 5 (μM)	12.5	12.5
E_0 exp. 1 (μM)	0.005516	0.005286
E_0 exp. 2 (μM)	0.005321	0.005168
E_0 exp. 3 (μM)	0.006	0.006
E_0 exp. 4 (μM)	0.004391	0.004428
E_0 exp. 5 (μM)	0.003981	0.004105
offset exp. 1	-0.004339	-0.004234
offset exp. 2	-0.001577	-0.003478
offset exp. 3	-0.01117	-0.0142
offset exp. 4	-0.001661	-0.005177
offset exp. 5	0.007133	0.00486

$$
\begin{gathered}
[\mathrm{M}]_0 = 0 \quad [\mathrm{P}]_0 = 0 \quad [\mathrm{ES}]_0 = 0 \\
[\mathrm{EP}]_0 = 0 \quad [\mathrm{EI}]_0 = 0 \quad [\mathrm{EJ}]_0 = 0 \\
[\mathrm{I}]_0\,(\mathrm{exp1}) = 0 \quad [\mathrm{I}]_0\,(\mathrm{exp2}) = 0.0015 \\
[\mathrm{I}]_0\,(\mathrm{exp3}) = 0.003 \quad [\mathrm{I}]_0\,(\mathrm{exp4}) = 0.004 \\
[\mathrm{I}]_0\,(\mathrm{exp5}) = 0.004
\end{gathered} \tag{57}
$$

$$
k_{11} = 0.1 \quad k_{12} = 0.001 \quad k_{41} = 100
$$

$$
k_{21} = 100 \quad k_{51} = 100 \tag{58}
$$

$$
t \in [0, 3600] \tag{59}
$$

A series of five experiments where the enzyme HIV proteinase (E) (assay concentration 0.004 μM) was added to a solution of an irreversible inhibitor (I) and a fluorogenic substrate (S) (25 μM) were considered. The five experiments were carried out at four different concentrations of the inhibitor (0, 0.0015, 0.003, and 0.004 μM in replicate).

The fluorescence changes were monitored during one hour. The measured signal is a linear function of the product (P) concentration, as expressed in the following equation:

$$
\mathrm{signal} = \varepsilon[\mathrm{P}] + \mathrm{offset} \tag{60}
$$

In this fit, the offset (baseline) of the fluorimeter was considered as a degree of freedom. A certain degree of uncertainty (±50%) was assumed for the value of the initial concentrations of substrate and enzyme (titration errors).

The calibration of a total of 20 adjustable parameters was addressed: five rate constants, five initial concentrations of enzyme and substrate and five offset values. Mendes and Kell [34] solved this problem using simulated annealing and reported its first known solution. Later, Rodriguez-Fernandez et al. [4] improved that solution by means of a scatter search metaheuristic, which required a fraction of the time employed by Mendes' simulated annealing. Recall that, despite producing near optimal solutions in short CPU times, stochastic algorithms provide no information on the quality of the solutions found and are unable to guarantee convergence to the global optimum in a finite number of iterations. On the contrary, the proposed methodology ensures the global optimality of the solution computed within a desired tolerance.

Table 4 Global optimization results for the HIV proteinase inhibition problem

	Rodriguez-Fernandez et al.	BARON	Proposed algorithm
Sum of squares	0.01997	failed	0.01961
UB	-	-	0.01961
LB	-	-	0.01595
Gap (%)	-	-	18.64
Iterations	29,345	263	3
Time (CPU s)	1,294	43,200	4,351

In our study, the state variables were approximated using five orthogonal collocation points and five equal-length finite elements. In this case, the nonconvexities arise from the bilinear terms $\xi_{e,k,j}\xi_{e,k,j}$ and $\theta_i\xi_{e,k,j}$.

The parameter bounds θ_i were set to $\theta_i \in [0, 10^6]$. The lower and upper limits for the collocation coefficients $\xi_{e,k,j,n}$ were fixed to $\xi_{e,k,j,n} \in [0, 37.5]$ except for $\xi_{e,k,\mathrm{E},n} \in [0.002, 0.006]$ and $\xi_{e,k,\mathrm{S},n} \in [12.5, 37.5]$. The bounds for all the offsets were set to $\mathrm{offset}_n \in [-0.5, 0.5]$.

The master problem was further tightened by adding a special type of strengthening cuts. These cuts are generated by temporally decomposing the original full space MILP into a series of MILPs in each of which we fit only a subset of the original dataset, and remove the continuity equations corresponding to the extreme elements included in the sub-problem. The cuts are expressed as inequalities added to the master problem that impose lower bounds on the error of a subset of elements for which the sub-MILPs are solved. These bounds are hence obtained from the solution of a set of MILP sub-problems that optimize the error of only a subset of elements.

This case study was solved with 3 initial piecewise intervals and 6 initial hyper-planes. Two strengthening cuts involving elements 1, 2, 3 and 4, and 2, 3, 4 and 5, respectively, were added as constrains. A tolerance of 20% was used in the calculations. Hyper-planes and piecewise terms were updated at each iteration of the algorithm. In this case, BARON failed to identify any feasible solution after 12h of CPU time.

In contrast, our algorithm was able to obtain the global optimum (Table 3) with a gap of 18.64% in approximately 4,000 CPU s (Table 4). Remarkably, the solution found by our algorithm improves the best known solution reported by Rodriguez-Fernandez et al. [4]. Hence, our algorithm clearly outperformed other parameter estimation methods, improving the best known solution [4,34], and providing a rigorous lower bound on the minimum error that can be attained.

Conclusions

In this work, we have proposed a novel strategy for globally optimizing parameter estimation problems with embedded nonlinear dynamic systems. The method presented was tested through two challenging benchmark problems: the isomerisation of α-Pinene (case study 1) and the inhibition of HIV proteinase (case study 2).

The proposed algorithm identified the best known solution, which was originally reported by Rodriguez-Fernandez et al. [4], in the case of the α-Pinene, and improved the best known one in the HIV proteinase case study. In both cases, rigorous lower bounds were provided on the global optimum, making it possible to determine the optimality gap of the solutions found.

The method proposed produced promising results, surpassing the capabilities of BARON. Our method requires some knowledge on optimization theory as well as skills using modelling systems. Our final goal is to develop a software to automate the calculations, so our approach can be easily used by a wider community. This is a challenging task, since nonlinear models are hard to handle and typically require customized solution procedures. Particularly, nonlinear models must be initialized carefully to ensure convergence even to a local solution. In this regard, the use of an outer approximation scheme that relies on a master MILP formulation is quite appealing, since the outcome of this MILP can be used to initialize the NLP in a robust manner.

Another key point here is how to construct tight relaxations of the nonconvex terms. An efficient algorithm must exploit the problem structure to obtain high quality relaxations and therefore good bounds close to the global optimum. These relaxations can be further tightened through the addition of cutting planes or the use of customized decomposition methods. As observed, there is still much work to be done in this area, but we strongly believe that such an effort is worthy. Furthermore, recent advances in global optimization theory and software applications are paving the way to develop systematic deterministic tools for the global optimization of parameter estimation problems of increasing size. Our future work will focus on making the approach more efficient through the use of tailored cutting planes and decomposition strategies and also through the hybridization of deterministic methods with stochastic approaches.

Competing interest

The authors declare that they have no competing interests.

Authors' contributions

G. G-G suggested the need for the approach and J.A. E provided the biological problems. A. M, C. P, G. G-G and L. J developed the optimization algorithms and performed the numerical analysis. All authors evaluated the results, wrote the paper and contributed to its final form.

Acknowledgements

This work was supported by the Spanish Ministry of Science and Innovation through the doctoral research grant FPI-MICINN reference grant BES-2010-037166 and through the project CTQ2009-14420-C02-01.

Author details

[1]Departament d'Enginyeria Química, Universitat Rovira i Virgili, Tarragona, Spain. [2]Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, Cartagena, Spain.

Received: 7 November 2011 Accepted: 10 May 2012
Published: 10 May 2012

References

1. Kameswaran S, Biegler L: **Simultaneous dynamic optimization strategies: recent advances and challenges.** *Comput & Chem Eng* 2006, **30**(10-12):1560–1575.
2. Esposito W, Floudas C: **Global optimization for the parameter estimation of differential-algebraic systems.** *Ind & Eng Chem Res* 2000, **39**(5):1291–1310.
3. Cizniar M, Salhi D, Fikar M, Latifi M: **A MATLAB package for orthogonal collocations on finite elements in dynamic optimisation.** In *Proc 15 Int Conference Process Control,* **Volume 5**:058f.
4. Rodriguez-Fernandez M, Egea J, Banga J: **Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems.** *BMC Bioinf* 2006, **7:**483.
5. Esposito W, Floudas C: **Deterministic global optimization in nonlinear optimal control problems.** *J Global Optimization* 2000, **17:**97–126.
6. Papamichail I, Adjiman C: **A rigorous global optimization algorithm for problems with ordinary differential equations.** *J Global Optimization* 2002, **24:**1–33.
7. Singer A, Barton P: **Global solution of optimization problems with parameter-embedded linear dynamic systems.** *J Optimization Theory and Appl* 2004, **121**(3):613–646.
8. Kesavan P, Allgor R, Gatzke E, Barton P: **Outer approximation algorithms for separable nonconvex mixed-integer nonlinear programs.** *Math Programming* 2004, **100**(3):517–535.
9. Biegler L, Grossmann I: **Retrospective on optimization.** *Comput & Chem Eng* 2004, **28**(8):1169–1192.
10. Finlayson B: *The method of weighted residuals and variational principles: with application in fluid mechanics, heat and mass transfer, Volume 87*: Academic Pr; 1972.
11. Cuthrell J, Biegler L: **On the optimization of differential-algebraic process systems.** *AIChE J* 1987, **33**(8):1257–1270.
12. Tieu D, Cluett W, Penlidis A: **A comparison of collocation methods for solving dynamic optimization problems.** *Comput & Chem Eng* 1995, **19**(4):375–381.
13. Pozo C, Guillén-Gosálbez G, Sorribas A, Jiménez L: **Outer approximation-based algorithm for biotechnology studies in systems biology.** *Comp & Chem Eng* 2010, **34**(10):1719–1730.
14. Carlos P, Alberto M, Rui A, Gonzalo G, Laureano J, Albert S: **Steady-state global optimization of metabolic non-linear dynamic models through recasting into power-law canonical models.** *BMC Syst Biol*, **5:**137.
15. Pozo C, Guillén-Gosálbez G, Sorribas A, Jiménez L: **A spatial branch-and-bound framework for the global optimization of kinetic models of metabolic networks.** *Ind & Eng Chem Res* 2010.
16. Sorribas A, Pozo C, Vilaprinyo E, Guillén-Gosálbez G, Jiménez L, Alves R: **Optimization and evolution in metabolic pathways: Global optimization techniques in Generalized Mass Action models.** *J Biotechnol* 2010, **149**(3):141–153.
17. Guillén-Gosálbez G, Sorribas A: **Identifying quantitative operation principles in metabolic pathways: a systematic method for searching feasible enzyme activity patterns leading to cellular adaptive responses.** *BMC Bioinf* 2009, **10:**386.
18. Wicaksono D, Karimi I: **Piecewise MILP under-and overestimators for global optimization of bilinear programs.** *AIChE J* 2008, **54**(4):991–1008.
19. Androulakis I, Maranas C, Floudas C: **αBB: A global optimization method for general constrained nonconvex problems.** *J Global Optimization* 1995, **7**(4):337–363.
20. Smith E, Pantelides C: **A symbolic reformulation/spatial branch-and-bound algorithm for the global optimisation of nonconvex MINLPs.** *Comput & Chem Eng* 1999, **23**(4-5):457–478.
21. Quesada I, Grossmann I: **Global optimization algorithm for heat exchanger networks.** *Ind & Eng Chem Res* 1993, **32**(3):487–499.
22. Smith E, Pantelides C: **Global optimisation of general process models.** *NONCONVEX OPTIMIZATION APPL* 1996, **9:**355–384.
23. McCormick G: **Computability of global solutions to factorable nonconvex programs: Part I Convex underestimating problems.** *Math Programming* 1976, **10:**147–175.
24. McCormick G: **Nonlinear programming: Theory, algorithms, and applications.** *JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, USA, 1983, 464*, 1983.
25. Misener R, Thompson J, Floudas C: **APOGEE: Global optimization of standard, generalized, and extended pooling problems via linear and logarithmic partitioning schemes.** *Comput & Chem Eng* 2011, **35:**876–892.
26. Singer A, Barton P: **Bounding the solutions of parameter dependent nonlinear ordinary differential equations.** *SIAM J Sci Comput* 2006, **27**(6):2167–2184.
27. Singer A, Barton P: **Global optimization with nonlinear ordinary differential equations.** *J Global Optimization* 2006, **34**(2):159–190.
28. Karuppiah R, Grossmann I: **Global optimization for the synthesis of integrated water systems in chemical processes.** *Comput & Chem Eng* 2006, **30**(4):650–673.
29. Fuguitt R, Hawkins J: **Rate of the thermal isomerization of α-Pinene in the liquid phase1.** *J Am Chem Soc* 1947, **69**(2):319–322.
30. Hunter W, McGregor J: **The estimation of common parameters from several responses: Some actual examples.** In *Unpublished Report.* The Department of Statistics. University of Winsconsin; 1967.
31. Grossmann I, Biegler L: **Part II. Future perspective on optimization.** *Comput & Chem Eng* 2004, **28**(8):1193–1218.
32. Hansen P, Jaumard B, Lu S: **An analytical approach to global optimization.** *Math Programming* 1991, **52:**227–254.
33. Kuzmic P: **Program DYNAFIT for the analysis of enzyme kinetic data: application to HIV proteinase.** *Anal Biochem* 1996, **237**(2):260–273.
34. Mendes P, Kell D: **Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation.** *Bioinformatics* 1998, **14**(10):869.

doi:10.1186/1471-2105-13-90
Cite this article as: Miró *et al.*: Deterministic global optimization algorithm based on outer approximation for the parameter estimation of nonlinear dynamic biological systems. *BMC Bioinformatics* 2012 **13**:90.

Guillén-Gosálbez *et al. BMC Systems Biology* 2013, **7**:113
http://www.biomedcentral.com/1752-0509/7/113

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

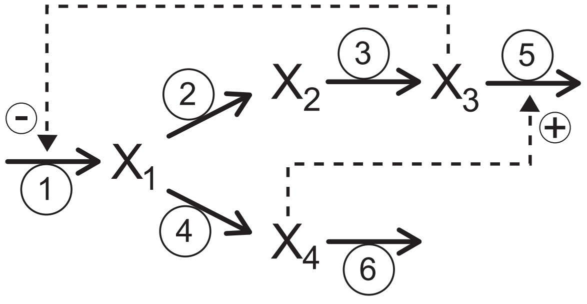

Figure 1 Reference system taken from Voit and Almeida [23] (default parameters are shown in Table 1).

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.*

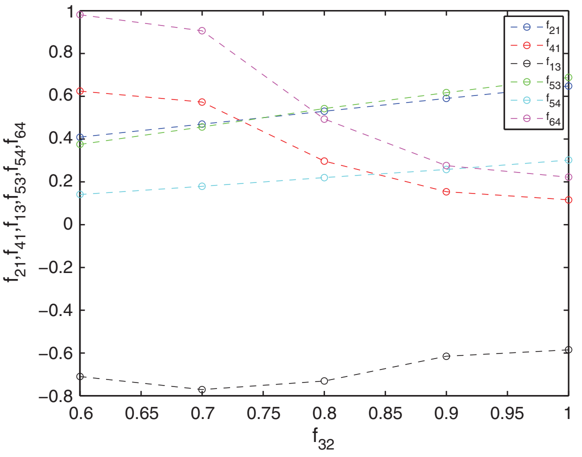

Figure 2 Values of the fitted parameters for different values of f32. Each point was generated by fixing f32 and solving the NLP free of error.

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.

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.

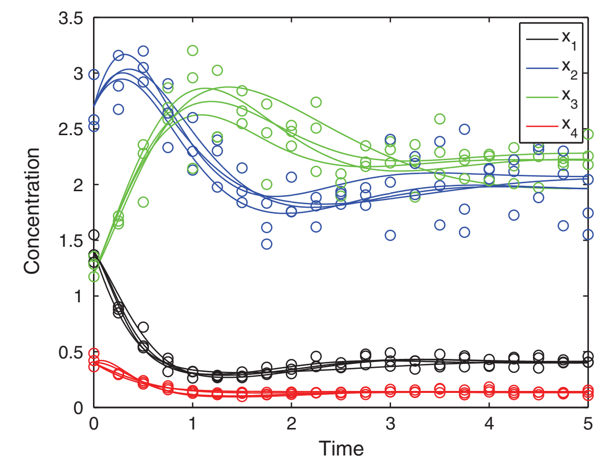

Figure 3 Adjusted profiles for four different noisy data sets (i.e. one experimental condition and four replications) with a standard deviation of 10%.

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

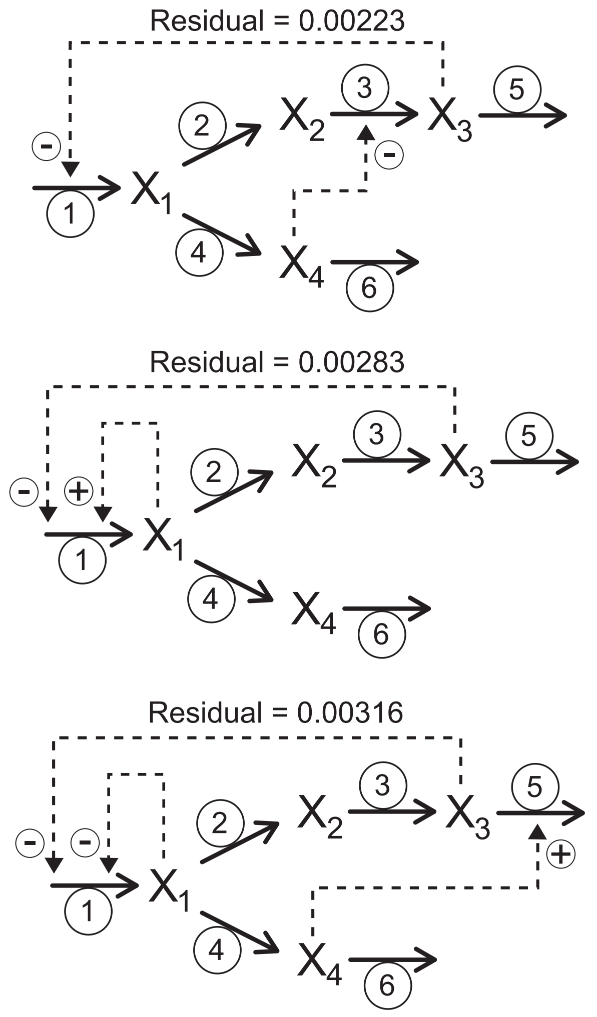

Figure 4 The proposed method identifies different regulatory topologies that essentially produce the same output. We show here the associated profiles corresponding to three regulatory structures with lowest residual values obtained by analyzing data from a single experiment with one replicate (see parameters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

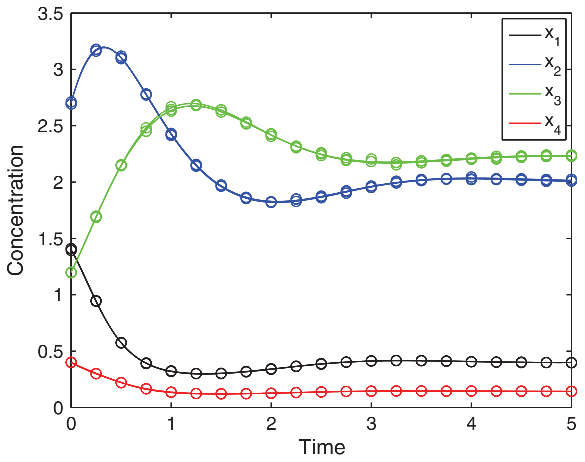

Figure 5 Dynamic responses corresponding to the three different topologies of Figure 4. Parameter values are indicated on Additional file 1: Table S3.

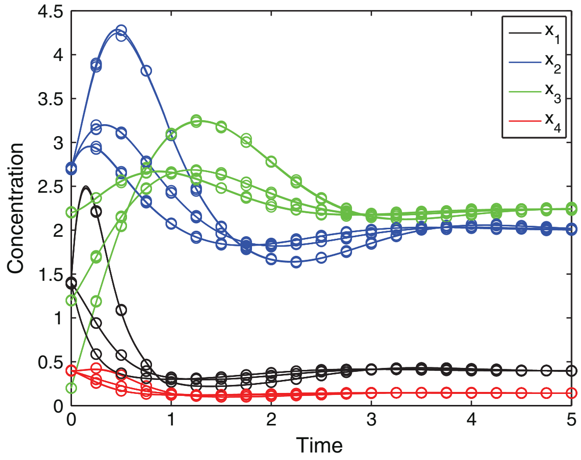

Figure 6 The Profiles generated from three different topologies and three experiments with one replication each. The experiments are generated from the base case by applying different perturbations in the initial concentration of X_3. Details on the topology and associated parameters are provided on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

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

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:

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

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:

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

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:

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

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]:

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

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:

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

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:

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

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.

Supplementary Material for

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]

Corresponding author*

Contact: gonzalo.guillen@urv.cat

Supplementary table 1. 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%. In the estimation task, the non-zero parameters are indicated and all the other parameters are set to zero.

Profile	f_{13}	f_{21}	f_{32}	f_{41}	f_{53}	f_{54}	f_{64}	Residual
1	-0.14	0.26	0.44	0.04	0	-0.06	0.13	1.88
2	-0.27	0.47	1	0	0.26	0.04	0.07	1.67
3	-0.84	0.4	0.64	0.9	0.42	0.1	1	1.68
4	-0.79	0.29	0.41	1	0.12	-0.12	1	2.29
5	-0.77	0.58	1	0.88	0	0.22	1	1.81
6	0.62	0.12	0.2	0	0.41	-1	0.93	2.14
7	-0.65	0.33	0.5	0.27	0.97	0.21	0.5	1.42
8	-0.45	0.4	0.62	0	0.12	0.27	0.28	1.78
9	-0.42	0.66	1	0.19	0	-0.63	0.35	1.93
10	-1	0.79	0.79	0	0.73	0.58	1	1.64
11	-0.64	0.86	1	0	0.49	0.61	1	1.89
12	-0.36	0.33	0.41	0	0.3	0.23	0.38	1.19
13	-0.39	0.26	0.47	0	0	0.07	1	2.01
14	-0.42	0.12	0.18	0.58	0	-0.15	0.58	2.06
15	0.25	0.02	0.53	0	1	-0.16	0.02	5.62
16	-0.54	0.26	0.39	0.41	0.31	0.06	0.54	1.11
17	-0.27	0.11	0.38	0.45	0	-0.2	0.72	1.6
18	0.37	0.07	0.11	0	1	0.12	1	1.53
19	-1	0.71	1	0.13	0.95	0.7	0.31	1.17
20	-0.49	0.26	0.41	0.7	0.22	0.08	1	2.03
21	-0.43	0.84	1	0	0.14	0.11	0.06	2.25
22	0.56	0.14	0.06	0	1	-1	0.87	2.52
23	-0.31	0.57	1	0	0	-0.07	1	1.93
24	-1	0.5	0.89	0.88	0.23	0.13	1	1.27
25	-0.22	0.22	0.28	0.31	0.13	0.04	1	1.28
26	-0.3	0.61	1	0	0.42	0.03	0	2.85
27	-0.05	1	0.67	1	0	-0.12	0.16	37.74

28	-0.76	0.72	1	0	0.36	0.58	1	1.89
29	-0.03	0.91	1	0	0.88	-0.51	0.05	2.33
30	-0.75	0.22	0.53	0.98	0	-0.1	1	2.24
31	-0.11	0.19	0.3	0	0.11	-0.06	0.31	1.71
32	-0.53	0.23	0.37	0.67	0.12	-0.08	1	2.12
33	-0.84	0.72	1	0.1	1	0.77	0.44	1.3
34	-0.77	0.24	0.46	0.95	0.18	-0.05	1	2.12
35	-0.63	0.34	0.55	0.39	0.46	0.26	0.73	2.66
36	-0.55	0.77	1	0	0.63	0.39	0.07	1.38
37	-0.33	0.21	0.41	0	0.12	0.07	1	1.94
38	-0.17	0.27	0.4	0.18	0.02	-0.01	0.26	1.61
39	-0.6	0.21	0.53	1	0	-0.04	1	1.75
40	-0.66	0.32	0.54	0.89	0.16	0.07	1	1.75
41	-0.34	0.18	0.44	0.66	0	-0.03	1	2.07
42	-0.39	0.31	0.31	0.46	0.16	-0.04	1	1.34
43	-0.36	0.23	0.27	0	0.18	-1	1	2.81
44	-1	0.48	0.65	0.87	0.4	0.21	1	1.1
45	-0.9	0.03	0.02	1	1	0.06	1	18.53
46	-0.98	0.95	1	0.28	0.12	0.35	0.38	1.04
47	-0.28	0.49	1	0	0.35	0.16	0.05	2.06
48	-0.38	0.24	0.45	0	1	0.54	1	1.98
49	-0.51	0.29	0.41	0.59	0.29	0.1	1	1.48
50	-0.17	0.2	0.33	0	0.89	0.83	0.49	2.47
51	-0.94	0.54	0.82	0.62	0.58	0.45	1	1.29
52	-0.84	0.3	0.43	1	0.2	-0.24	1	2.35
53	-0.36	1	0.94	0.09	0.38	-0.25	0.03	2.67
54	-0.83	0.61	1	0.71	0.12	0.22	1	1.64
55	-1	0.57	1	0.47	1	0.75	0.78	1.16
56	-0.3	0.36	0.61	0	0.92	1	0.3	1.15
57	-0.43	0.78	1	0	0.53	1	0.94	1.71
58	-0.22	0.43	0.63	0	0.89	1	0.28	2.04
59	-0.37	0.28	0.53	0.04	0.44	0.32	0.35	1.64
60	-0.38	0.52	1	0	0	0.2	0.1	1.78
61	-0.91	0.6	1	0.27	0.02	0.19	0.23	1.79
62	-0.83	0.55	0.62	1	0.13	0	1	1.49
63	-0.52	0.26	0.42	0	0.26	-1	1	2.23
64	-1	0.71	1	0.26	0.16	0.19	0.24	1.85
65	-0.42	0.37	0.48	0.18	0.03	0.12	0.56	1.74
66	-0.36	0.25	0.43	0	0.13	0.04	0.28	1.11
67	-0.87	0.74	1	0.23	0.54	0.25	0.18	1.45
68	0	0.59	1	0	0.09	-0.12	0	9.04
69	-1	0.57	0.9	1	1	0.03	0.23	1.93
70	-1	0.64	1	0.88	0.54	0.32	1	1.62
71	-1	0.73	1	0.7	0.16	0.18	1	1.73
72	-0.38	0.24	0.39	0	0.33	0.16	0.46	1.42
73	-0.24	0.25	0.39	0	0.23	0.12	0.37	1.93
74	-0.36	0.08	0.17	1	0	-0.2	1	2.39
75	-0.35	0.23	0.38	0	0.09	0.09	0.34	1.27
76	-1	0.65	1	0.6	0.7	0.46	1	1.69
77	-1	0.76	1	0.24	0.54	0.08	0.42	1.41
78	-0.65	0.51	1	0.39	0.45	0.19	0.44	2.02
79	0.32	0	0.13	0.01	0.37	-0.01	0.13	4.22

80	-0.54	0.3	0.51	0.48	0.18	0.17	1	1.31
81	-1	0.75	1	0.33	0.23	0.33	0.49	1.67
82	-0.54	0.2	0.39	0.97	0	-0.09	1	1.67
83	-1	0.95	1	0.4	1	1	1	2.12
84	-1	0.72	1	0.32	0.34	0.21	0.3	1.71
85	-0.63	0.32	0.54	0.88	0.12	0.07	1	2.77
86	-1	0.88	1	0.25	0.47	0.3	0.19	1.82
87	-1	0.64	0.87	0	0.17	0.2	1	1.75
88	0	0	0.17	0.01	0.4	-0.12	0	12.4
89	-0.25	0.2	0.3	0	0.16	-0.01	0.86	0.94
90	-1	0.58	0.81	0.01	1	0.37	0.21	2.13
91	-0.67	0.7	1	0.01	1	1	0.44	1.44
92	-0.56	0.6	1	0.17	0.09	0.14	0.17	2.12
93	-0.15	0.22	0.32	0	0.11	-0.01	0.92	1.94
94	-0.27	0.19	0.28	0.89	0	-0.09	1	2.09
95	-0.65	0.26	0.65	1	0.19	0.12	1	1.58
96	0.4	0.21	0.14	0.23	1	0.02	0	5.62
97	-0.42	0.47	0.5	0	0.64	0.82	0.46	1.07
98	-0.75	0.29	0.43	1	0	-0.12	1	1.56
99	-0.33	0.24	0.28	0	0.26	-1	0	3.19
100	-0.15	0.16	0.39	0	0.19	0.12	0.69	1.57

Supplementary table 2. 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%.

Profile	f_{12}	f_{14}	f_{23}	f_{31}	f_{35}	f_{45}	f_{46}	Residual
1	0.33	0.64	0.42	-0.67	0.49	0.05	1	6.96
2	0.9	0	1	-0.64	0.66	-0.95	1	7.1
3	0.49	0.38	0.73	-0.62	0.3	0.22	0.53	5.39
4	0.69	0.26	1	-0.92	0.4	0.34	0.58	4.89
5	0.31	0.71	0.54	-0.49	0.24	0.13	1	5.58
6	0	0.01	0.2	0.18	0.35	-0.03	0.06	22.51
7	0.58	0.27	1	-0.64	0.69	0.26	0.44	6.14
8	0.51	0.12	0.83	-0.72	0.46	0.17	0.17	5.02
9	0.6	0.59	0.8	-0.79	0.64	0.08	0.86	5.96
10	0.56	0	1	-0.57	0.26	-0.46	1	8.3
11	0.31	0.59	0.54	-0.52	0.16	0.12	0.74	7.04
12	0.55	0.5	1	-1	0.69	0.63	1	6.19
13	0.49	0.23	0.6	-0.66	0.6	0.18	0.29	6.58
14	0.61	0.34	1	-0.61	0.42	-0.03	0.48	6.82
15	0.52	0	0.81	-0.52	0.19	0.26	0.52	7.37
16	0.68	0.16	0.98	-0.72	0.87	0.52	0.29	5.65
17	0.56	0.7	1	-0.87	0.41	0.25	1	4.26
18	0.4	0.74	0.69	-0.78	0.3	0.11	1	5.76
19	0.26	0.52	0.48	-0.62	0.33	0	0.85	7.54
20	0.46	0.61	0.77	-0.78	0.38	0.23	1	5.54
21	0.39	0.72	0.62	-0.72	0.22	0.05	1	5.59
22	1	0	1	0.14	1	-1	0	36.61
23	0.17	0.05	0.3	1	0	-1	1	18.13
24	0.6	0.08	0.94	-0.91	0.52	-0.06	0.66	8.37

25	0.55	0	0.84	-0.9	0.57	-0.12	0.65	7.66
26	0.46	0.68	0.67	-0.78	0.48	0.12	1	5.07
27	0.2	0	0.04	1	0.99	-0.05	0.02	43.47
28	0.46	0	0.75	1	0	-1	1	18
29	0.32	0.13	0.43	-0.5	0.29	0.13	0.3	5.73
30	0.5	0.39	0.72	-0.76	0.45	0.33	0.53	6.61
31	0.31	1	0.36	-1	0	-1	0.44	74.29
32	0.31	0.33	0.49	-0.45	0.14	0.1	1	6.32
33	0.64	0.22	1	-0.9	0.7	0.47	0.46	5.23
34	0.72	0.39	1	-0.89	0.82	0.3	0.47	5.6
35	0.72	0.09	1	-0.74	0.64	0.47	0.19	5.71
36	0.44	0.65	0.62	-0.77	0.45	0.19	1	6.93
37	0.75	0.01	0.99	-0.59	0.61	0.2	0.03	7.02
38	0.6	0.12	1	-0.57	0.49	0.25	0.18	5.35
39	0.32	0.63	0.53	-0.61	0.15	-0.02	1	4.77
40	0.6	0.04	0.86	-0.73	0.39	0.36	0.61	8.59
41	0.54	0.67	0.92	-0.95	0.54	0.18	1	6.05
42	0.51	0.6	0.73	-0.87	0.55	0.04	1	4.48
43	0.61	0.1	1	-0.58	0.74	0.35	0.15	4.71
44	0.39	0.63	0.64	-0.78	0.27	-0.02	1	5.15
45	0.7	0	1	-0.59	0.6	0.89	0.55	5.71
46	0.64	0.35	1	-0.96	0.48	0.21	1	7.81
47	0.35	0.33	0.52	-0.59	0.29	0.02	0.71	5.85
48	0.46	0.18	0.69	-0.65	0.36	0.16	0.37	4.94
49	0.66	0.08	1	-0.59	0.72	0.2	0.15	5.39
50	0.62	0	1	-0.96	0.67	0.33	0.45	5.71
51	0	1	0.42	0.38	1	-0.13	1	19.11
52	0.62	0.2	1	-0.83	0.66	0.29	0.32	6.27
53	0	0	1	0.22	1	-1	0	27.25
54	1	0	1	1	1	-1	0.02	21.33
55	0.52	0.09	0.77	-0.64	0.46	0.29	0.39	4.66
56	0.46	0.17	0.67	-0.67	0.4	0.32	0.42	5.94
57	0.3	0.11	0.49	-0.39	0.19	0.13	0.38	5.61
58	0.44	0	0.53	-0.44	0.22	0.23	0.83	5.24
59	0.41	0.63	0.58	-0.77	0.43	0.26	1	7.23
60	0.07	0	0	0.4	1	0.01	0.01	42.55
61	0.17	0.08	0.04	1	0	-0.37	1	33.1
62	0.61	0.59	1	-0.92	0.37	0.17	1	4.71
63	0.63	0	0.97	-0.79	0.58	0	0.02	7
64	0.55	0.3	0.9	-0.68	0.68	0.57	0.53	7.72
65	0.37	0.67	0.52	-0.68	0.38	0.16	1	6.53
66	0.44	0.7	0.71	-0.76	0.38	0.16	1	6.25
67	0.49	0.37	1	-0.7	0.33	0.17	0.59	7.31
68	0.27	0.69	0.4	-0.45	0.14	0.07	1	6.1
69	0.56	0.23	0.82	-0.69	0.34	0.08	0.39	5.05
70	0.63	0	0.94	-0.38	0.49	0.2	0.06	4.82
71	0.57	0.06	0.96	-0.63	0.52	0.19	0.2	5.71
72	0.59	0.2	0.92	-0.71	0.56	0.61	0.48	6.15
73	0.53	0.3	0.95	-0.8	0.49	0.38	0.51	5.75
74	0.63	0.09	0.96	-0.5	0.48	0.04	0.16	5.93
75	0.63	0.49	1	-1	0.7	0.26	0.81	4.41
76	0.42	0.24	0.59	-0.58	0.34	0.19	0.37	5.86

77	0.64	0.32	1	-0.88	0.62	0.23	0.39	7.05
78	0.21	0.02	0.37	-0.01	0.01	-0.01	0	30.56
79	0.56	0.32	0.92	-0.86	0.39	0.37	0.44	5.66
80	0.06	0	0.09	1	0.3	-1	0.87	15.24
81	1	0.01	0.3	0.53	1	0.29	0.01	22.46
82	0.59	0.25	1	-0.69	0.31	0.23	0.53	6.88
83	0.32	0.24	0.59	-0.44	0.28	0.31	0.56	5.19
84	0.48	0	0.82	-0.59	0.29	0.19	0.23	5.8
85	0.47	0.7	0.77	-0.92	0.34	0.03	1	6.19
86	0.68	0	1	-1	0.54	0.28	0.8	5.08
87	0.31	0.41	0.44	-0.51	0.26	0.03	0.71	6.21
88	0	1	0.9	-1	0	-1	1	34.87
89	0.36	0.35	0.68	-0.64	0.33	0.2	0.54	5.18
90	0.56	0.09	1	-0.69	0.46	0.29	0.3	5.72
91	0.14	0.04	0.04	1	0	-0.41	1	35.86
92	0.4	0.68	0.71	-0.62	0.35	0.2	1	5.77
93	0.65	0.08	0.98	-0.86	0.29	0.27	0.54	6.52
94	0.73	0.25	1	-1	0.52	0.12	0.39	6.48
95	0.01	0.51	0.48	1	0.42	-0.15	1	26.25
96	0.46	0.18	0.63	-0.51	0.54	0.16	0.32	6.32
97	0.42	0.31	0.5	-0.54	0.32	0.01	0.5	5.55
98	0	0.02	0.32	-1	0.38	-1	0.32	39.4
99	0.63	0.11	0.98	-0.66	0.85	0.48	0.35	6.14
100	0.55	0	0.71	-0.87	0.5	0.02	0.34	6.31

Supplementary table 3. 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.

Topology		v_1	v_2	v_3	v_4	v_5	v_6	Akaike	Residual
1	X_1	-	0.59	-	0.4	0.27	-	-176.18	0.00223
	X_2	-	-	1	-	-	-		
	X_3	-0.82	-	-	-	0.88	-		
	X_4	-	-	-0.08	-	-	0.62		
2	X_1	0.08	0.55	-	0.31	-	-	-171.12	0.00283
	X_2	-	-	0.83	-	-	-		
	X_3	-0.63	-	-	-	0.62	-		
	X_4	-	-	-	-	0.3	0.46		
3	X_1	-0.01	0.52	-	0.35	-	-	-168.86	0.00316
	X_2	-	-	0.77	-	-	-		
	X_3	-0.76	-	-	-	0.49	-		
	X_4	-	-	-	-	0.21	0.6		
4	X_1	-0.06	0.38	-	0.22	-	-	-168.42	0.00322
	X_2	-	-	0.55	-	0.07	-		
	X_3	-0.58	-	-	-	0.28	-		
	X_4	-	-	-	-	-	0.41		
5	X_1	-	0.48	-	0.36	0.31	-	-167.15	0.00342

	X_2	-	-	0.72	-	-	-		
	X_3	-0.72	-	-	-	0.76	-		
	X_4	-	-	0.02	-	-	0.62		
6	X_1	-	0.33	-	0.27	-	-	-168.45	0.00354
	X_2	-	-	0.48	-	-	-		
	X_3	-0.5	-	-	-	0.28	-0.07		
	X_4	-	-	-	-	-	0.4		
7	X_1	-	0.46	-	0.47	-	-	-166.42	0.00355
	X_2	-	-	0.59	-	-	-		
	X_3	-0.74	-	-0.12	-	0.27	-		
	X_4	-	-	-	-	0.18	0.8		
8	X_1	-	0.41	-	0.28	0.2	-	-166.06	0.0036
	X_2	-	-	0.59	-	-	-		
	X_3	-0.6	-	-0.06	-	0.54	-		
	X_4	-	-	-	-	-	0.52		
9	X_1	-	0.47	-	0.14	-	-	-165.58	0.0037
	X_2	-	-	0.6	-	-	-		
	X_3	-0.56	-	-	-	0.38	-		
	X_4	-	-	0.08	-	0.27	0.27		
10	X_1	-	0.7	-	0.41	0.26	-	-165.57	0.00369
	X_2	-	-	0.89	-	-	-		
	X_3	-0.91	-	-	-	0.87	-		
	X_4	-	-0.15	-	-	-	0.64		

Supplementary table 4. 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. 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.

Topology		v_1	v_2	v_3	v_4	v_5	v_6	Akaike	Residual
	X_1	-	0.52	-	0.3	-	-		
	X_2	-	-	0.77	-0.01	-	-		
	X_3	-0.74	-	-	-	0.54	-		
1	X_4	-	-	-	-	0.2	0.5	-483.49	0.0137
	X_1	-	0.48	-	0.38	-	-		
	X_2	-0.01	-	0.73	-	-	-		
	X_3	-0.75	-	-	-	0.49	-		
2	X_4	-	-	-	-	0.2	0.65	-480.62	0.0143
	X_1	-	0.49	-	0.35	-	-		
	X_2	-	-	0.74	-	-	-		
	X_3	-0.75	-	0.02	-	0.5	-		
3	X_4	-	-	-	-	0.21	0.61	-480.53	0.01431
	X_1	-	0.52	-	0.31	-	-		
	X_2	-	-	0.78	-	-	0.02		
	X_3	-0.76	-	-	-	0.54	-		
4	X_4	-	-	-	-	0.2	0.5	-479.19	0.01463

	X_1	-	0.52	-	0.31	-	-		
	X_2	-	-	0.8	-	-	-		
	X_3	-0.74	-	-	-	0.57	-		
5	X_4	-	-	-	-	0.21	0.54	-484.84	0.0147
	X_1	-	0.52	-	0.3	-	-		
	X_2	0.01	-	0.8	-	-	-		
	X_3	-0.75	-	-	-	0.52	-		
6	X_4	-	-	-	-	0.21	0.54	-478.44	0.0148
	X_1	-	0.53	-0.01	0.29	-	-		
	X_2	-	-	0.79	-	-	-		
	X_3	-0.72	-	-	-	0.57	-		
7	X_4	-	-	-	-	0.22	0.45	-477.18	0.0151
	X_1	-	0.5	-	0.32	-	-		
	X_2	-	-	0.76	-	-	-		
	X_3	-0.72	-	0.04	-	0.55	-		
8	X_4	-	-	-	-	0.21	0.55	-476.31	0.0153
	X_1	-	0.53	-	0.33	-	0.01		
	X_2	-	-	0.77	-	-	-		
	X_3	-0.75	-	-	-	0.53	-		
9	X_4	-	-	-	-	0.19	0.54	-475.8	0.01544
	X_1	-	0.54	-	0.29	-	-		
	X_2	-	-	0.8	-	-	-0.01		
	X_3	-0.73	-	-	-	0.54	-		
10	X_4	-	-	-	-	0.22	0.48	-475.54	0.0155

Bi-objective mixed-integer dynamic optimization approach for identifying the regulatory structure and kinetic parameters of biochemical networks

Anton Miró[a], Carlos Pozo[a], Gonzalo Guillén-Gosálbez[a,b]*, Sebastian Sager[c], Laureano Jiménez[a]

[a] Departament d'Enginyeria Química, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Spain
[b] Centre for Process Integration, School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester M13 9PL, UK
[c] Institute of Mathematical Optimization, Faculty of Mathematics, Otto-von-Guericke-Universität, Universitätsplatz 2, 02-224 39106 Magdeburg, Germany

Abstract

Background: Standard parameter estimation methods seek the parameters values that make the model response consistent with experimental observations assuming a given regulatory structure. Methods that characterize simultaneously both, the regulatory interactions and the associated parameters are few and lack generality. Building on a previous work by the authors, this paper presents an approach to carry out both tasks simultaneously. The parameter estimation problem is posed as a multi-criteria mixed-integer dynamic optimization (MIDO) problem in which the complexity and the quality of the fit are simultaneously minimized. This MIDO problem is solved using the epsilon constraint method, in which one objective is kept in the objective function while the rest are transferred to auxiliary constraints. The MIDO problem is reformulated into a mixed integer nonlinear programming model (MINLP) through orthogonal collocation on finite elements.

Results: A comparison between our method an another one presented in a previous work was carried out using a metabolic network and a series of noisy *in silico*-generated experimental data. Numerical results show that both methods lead to solutions showing similar Akaike information criterion (AIC) and residual values. Our method provides as output a set of plausible models with similar performance. To make a final choice, it is needed to apply advanced biological knowledge on the system. Another important outcome of this paper consists of a theoretical discussion on the use of the AIC in parameter estimation, showing its connections with the method presented herein.

Conclusion: An alternative bi-objective method for parameter estimation and model selection was proposed. The method has theoretical connections with the minimization of the AIC as unique criterion. Numerical results show that both methods provide similar results, with each of them displaying better performance in different case studies, and with no clear winner between the two. When identifying the best model to be implemented in practice we should not focus on a single metric, but rather look at: (i) the value of the AIC; (ii) the residual in the validation set; and (iii) the model complexity; and combine these criteria with biological knowledge of the system. Both of our methods provide as output a set of candidate models rather than a single "best" model, from which biologists should identify the most appropriate ones considering the items mentioned above.

Keywords: Parameter estimation, Structure identification, Akaike criterion, Orthogonal collocation, Dynamic optimization, epsilon-constraint method, multi-objective optimization.

Background

In recent years, the rapid development of high-throughput techniques has produced large amounts of data. In this context, mathematical modeling of biochemical systems has become an essential tool for complementing and extracting information from time series data in systems biology. Parameter values of a given model can be estimated assuming that its structure is known [1-4]. Unfortunately, the network topology is seldom fully characterized and the regulatory details are in many cases not completely understood or entirely missing. Regulatory topology can be inferred from time series data [5-7], but there are few systematic methods that accomplish this task.

In the process of building a reliable model, four main challenges are encountered: (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. Once the model is built, the last task entails testing the quality of the model through model validation. Ideally, the model should be able to predict systemic responses under yet untested experimental conditions.

The first challenge requires compiling information available about the system's mass flow structure in order to generate its corresponding stoichiometric matrix (see [8]). The next challenge entails the selection of the appropriate mathematical model among the different representations available. This step depends on the amount of information available. Mechanistic formulations based on physical sciences (*e.g.*, law of mass action, Michaelis-Menten rate law, etc.) are good choices when detailed information on the system is available. In the remaining cases, it is often preferred to use a generic formulation capable of capturing the nonlinear dynamics while yet keeping the model relatively simple. Canonical models are particularly useful for this purpose, as they facilitate both, parameter estimation and topology identification [9]. The third challenge consists of determining the appropriate numerical parameter values. The aim here is to obtain the set of parameter values that make the model response consistent with the data observed. This parameter estimation task can be formulated as an optimization problem that minimizes the sum of squared residuals between the measured and simulated data. The fourth challenge can be tackled in a similar manner as the third one. Regulatory connections can appear in a model as parameters accounting for the influence that metabolites others than the substrates of a reaction have on its velocity. Hence, determining the structure is equivalent to finding the values of these parameters.

There is a lack of methods that simultaneously identify the regulatory topology of a network along with the associated parameters values. Canonical models facilitate both tasks, as they are based on a general mathematical representation of a system that can model a wide variety of biological effects. The most widely used canonical nonlinear models are the Generalized Mass Action (GMA) and S-system formalisms, which are both based on the Biochemical Systems Theory (BST) [10-14]. Another recently proposed canonical form is the Saturable and Cooperative Formalism [15], which is based on a Taylor approximation that exhibits improved cooperativity and saturation predictions in comparison to other canonical formalisms. This last approach, however, requires a larger number of parameters, thereby increasing the estimation efforts.

In systems biology, there is a strong tendency to build very complex models. In this situation, when a model has too many parameters it is said to be overfitted. Overfitted models must be avoided since in addition to their complexity, they tend to capture noise as a real part of the system's structure [16]. These models display a diminished predictive capacity when they are tested on validation datasets. Conversely, underfitted models must also be avoided because of their inability to capture the system dynamics, which leads to unreliable predictions. Optimization tools provide a sound basis to assess the tradeoff between complexity and accuracy.

Finding the candidate model with the best predictive accuracy is challenging. Different methods have been proposed to carry out this task. One such approach is Cross-validation (CV), which was first proposed to measure the predictive performance of a model, and later expanded in scope to study model selection [17-20]. Simple model validation is a simplified variation of CV that relies on a single split of the data available. Part of the data is used for model fitting (training set), and the remaining part (validation set) is used for assessing the predictive accuracy of the model. This predictive accuracy can be measured, for instance, by the sum of squared residuals between the measured and simulated data. This residual will be, in general, greater than the error in the training set (as the validation set is not used for building the model).

Other approaches to assess the quality of the fitting include those based on information criteria, like Akaike information criterion (AIC) [21] or the Bayesian information criterion (BIC) [22]. AIC is often preferred over BIC because it has a more immediate connection to the theory of information [16]. Asymptotically (*i.e.*, considering a training set of infinite size), the model yielding the minimum AIC will also be the model with the minimum error in the validation set of a CV [23]. This property is valid for any model and makes the AIC method very useful for inference.

Smaller AIC values imply better approximations to reality. Considering this, it is possible to formulate model selection in mathematical terms as a single-objective optimization problem in which the AIC is minimized. With finite data, however, the model with minimum AIC and the one leading to the minimum error in the CV might not be the same. If this is the case, further biological knowledge of the system should be employed to discriminate between them. Furthermore, the model yielding the minimum AIC value for a finite sample is not guaranteed to be the best possible model, since the calculation of the AIC is performed considering only a limited number of samples. Rather than looking for the model with minimum AIC/residual, we should then focus on identifying models with small AICs and residual values, form which the "best" one according to additional biological criteria should be final selected [16]. Therefore, a good strategy for model selection and identification should have the capability of generating a set of plausible models with low AICs/residuals, from which biologists should choose the most appropriate ones taking into account their knowledge on the system.

In a previous work [24], we developed a method based on mixed-integer dynamic optimization (MIDO) that simultaneously identifies the regulatory signals and the kinetic parameters of models of biochemical networks by minimizing the AIC. This algorithm produces as output a set of candidate regulatory topologies (*i.e.*, models) for the target network without previous information about the biological system.

In this paper, we propose an alternative approach for model generation which formulates the parameter estimation task as a bi-criteria optimization problem. The AIC captures the trade-off between model complexity (measured by the number of parameters), and accuracy of the fitting. The way in which the AIC does this is by defining and assigning weights to both terms on the basis of information theory concepts. By solving the bi-criteria problem, we avoid the need to define these weights, thereby generating a set of candidate models representing the optimal trade-off between model accuracy and complexity.

More precisely, the parameter estimation task is posed as a bi-objective mixed-integer dynamic optimization (MIDO) problem in which the complexity and the deviation from reality (*i.e.*, the squared residual of the fitting of time series data) are simultaneously minimized. This problem can be solved by applying the ε-constraint method, which identifies a set of candidate models with an increasing number of regulatory interactions. Each of the sub-problems of the ε-constraint method is solved by reformulating the MIDO into a non-convex mixed-integer nonlinear programming (MINLP) model after complete discretization based on orthogonal collocation on finite elements. The performance of each Pareto optimal model is assessed using a CV strategy and computing also its AIC value. The Pareto set of models showing better performance are finally passed to biologists, which will keep the most promising ones based on their experience.

The overall structure of the manuscript is as follows. The next section presents the numerical results produced by our approach. In the following section the conclusions of the study are drawn. The last section of the paper describes the mathematical formulation and the methods used for its efficient solution.

Results and discussion

Our goal is to determine the regulatory structure of a metabolic network from time series data. An artificial branched pathway taken from Voit and Almeida [25] is used as a test bed to illustrate the capabilities of the method presented. The production of X_1 depends on the source X_0, and at the same time is inhibited by X_3. X_1 produces X_4 and X_2, and X_2leads to X_3. X_4 inhibits the degradation of X_3 (See Figure 1).

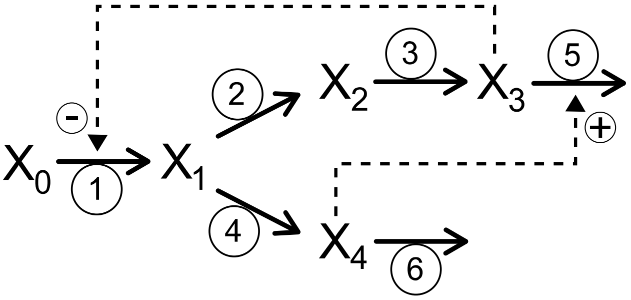

Figure 1 Reference artificial pathway taken from Voit and Almeida [25].

Given is a known structure for the reaction network (stoichiometry) and experimental time series data reflecting the dynamic behavior of a metabolic system, the goal of the analysis is to infer its regulatory structure. A *priori* knowledge about the system is available and can be used to constraint both, the complexity of the regulatory topology (the number of potential regulatory signals) and their signs (positive or negative) in the identification process. In our example, 6 kinetic orders corresponding to the substrates of the reactions must be positive. To reduce the number of possible interactions, a maximum of two metabolites are fixed as potential modifiers for each velocity, and one of them corresponds to the main substrate of the reaction.

The performance of our method is first assessed considering one single experiment *(i.e.*, one single configuration for the initial conditions) and assuming that the regulatory structure is unknown. We consider noisy data that is generated from the *in silico* model assuming that the "true" dynamic profile (*i.e.*, the one generated from the *in silico* model free of error) follows a normal distribution with standard deviation values of 5, 10, 15 and 30% (with respect to the actual nominal values). A single replication is generated from the normal distributions via Monte Carlo sampling. 21 experimental data points uniformly distributed over the entire dynamic simulation time are used to carry out the parameter estimation.

To assess the predictive capabilities of a dynamic model (entailing a specific regulatory structure and kinetic parameters), the *in silico* data was numbered and divided into two partitions. The first partition consisting of 10 experimental points was used for model fitting (training sample), and corresponds to data points numbered with even numbers. The second partition containing 55 experimental points (5 replications for 11 points) simulates untested experimental conditions (validation sample), and includes odd numbered experimental data points. These points were used for model validation (assuming the same level of uncertainty as before, and taking 5 replicates). The validation sample provides insight into the potential risk of overfitting (note that the training sample and the validation sample are fully independent from each other).

The training set is used to estimate the parameters values. In a previous work, a single-objective optimization model that minimizes the AIC criterion [24] was used to find the optimal parameters values of the model from a set of points in the training set. This paper proposes an alternative approach to generate plausible models based on a bi-objective formulation in which rather than minimizing a single objective (i.e., the AIC value), the model simultaneously optimizes the sum of squared deviations (between the "experimental points" in the training set and the ones predicted by the model) and the number of regulatory signals. The ε-constraint method is used to generate 10 Pareto solutions with an increasing number of regulatory interactions for each level of uncertainty (5, 10, 15 and 30%). Each of these solutions was obtained solving a single-objective problem in which the sum of squared deviations between experimental and simulated data is kept as main objective, while the number of regulatory interactions is transferred to an auxiliary constraint. Each of these single-objective MINLP problems was implemented in GAMS 24.2.3 and solved with the solver SBB on a PC/AMD Athlon at 2.99 Ghz using a single core. The model features 84 binary variables, 487 continuous variables and 513 equations. The solution time was in the order of few minutes for each simulation.

For comparison purposes, the same problem was solved using the single-optimization approach previously developed by the authors [24]. In this method, the AIC[1] is minimized as unique objective. This algorithm is executed iteratively in order to produce a set of potential network configurations along with the corresponding parameter values. 10 candidate models were generated and sorted according to their AIC values. The corresponding MINLP problem features 84 binary variables, 487 continuous variables and 512 equations.

Figures 2a, 3a, 4a and 5a show the residuals for each Pareto point (squared error in the training set) *vs* the model complexity (number of regulatory signals), while Figures 2b, 3b, 4b and 5b depict the residuals of the same models but tested in the validation set. The Pareto points are represented by green circles in the figures. The figures show as well the solutions produced using the AIC minimization approach, which are represented by blue squares. Particularly, 10 iterations of the ε-constraint algorithm (each of which requires solving a MIDO model) were executed. The AIC of the real metabolic network (*i.e.*, the one we used to generate experimental data and without noise) is also provided (red triangles). Note that this last analysis is only possible when dealing with an academic problem like the one studied here. The numbers attached to the points in the aforementioned figures denote the value of the AIC for that particular model. In the single-optimization approach, these values correspond to the objective function of the problem, whereas in the multi-objective approach, they are computed after running the MIDO algorithm.

We evaluate first which method is able to generate better models, that is, models that predict better the behavior of the system under untested conditions (using the validation set). In almost of the instances, the minimization of the AIC as single objective produces models with better predictive accuracy, that is, models with lower residuals in the validation set (as well as lower AIC values). However, the Pareto points generated using the bi-objective approach show AIC values very close to the minimum AIC value computed by the single-objective optimization approach. Recall that when models have similar AIC values, the model with the lowest AIC value may not be the best [16]. Furthermore, a finite dataset is used to perform the calculations, so the AIC value is indeed

[1] Although for simplicity we refer to it as AIC, we actually minimize the AICc instead, since the sample is not large enough to allow the appropriate use of the AIC. Refer to Methods section for further details.

an approximation to its "true" value. Additional biological considerations might therefore be applied when selecting the final model to be used.

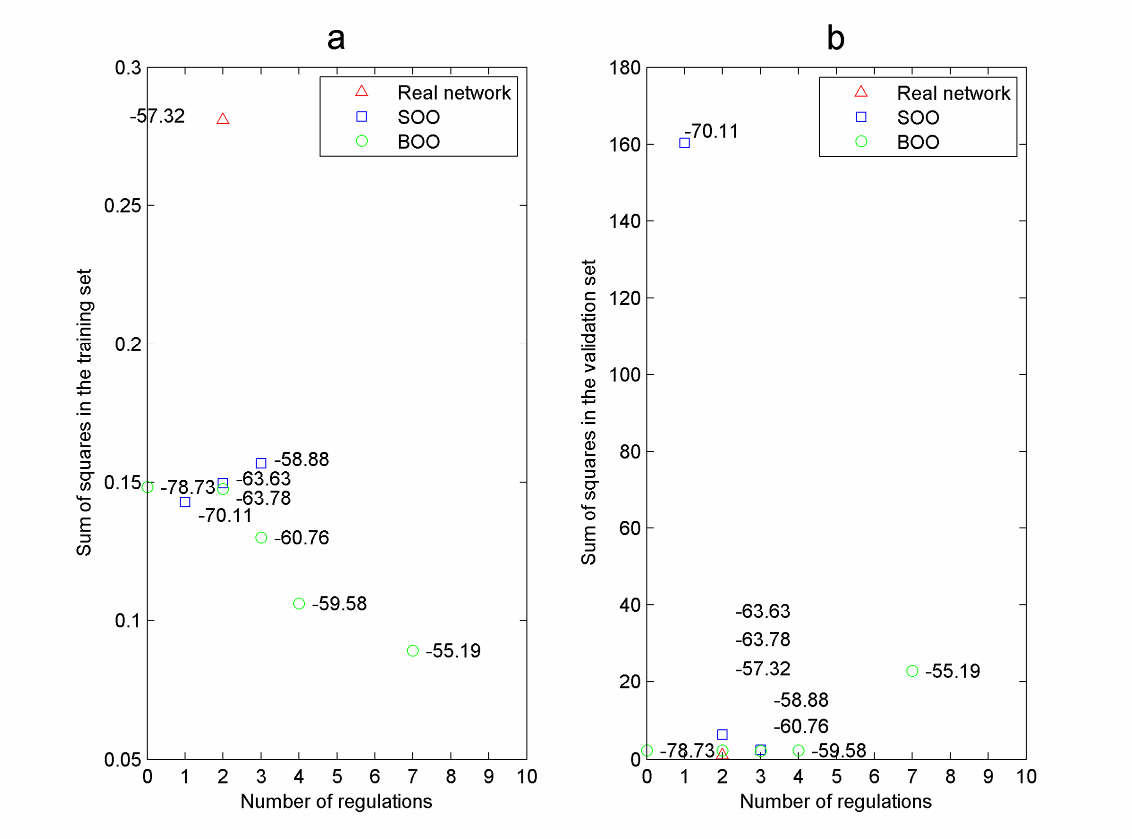

Figure 2 Number of regulations vs sum of squares in the training set (a) and in the validation set (b) for one experiment with 5% error. Blue squares refer to models generated with the single-objective optimization (SOO) approach whereas green circles represent models generated with the bi-objective optimization (BOO) scheme. Red triangles depict the real metabolic network (i.e., the one we used to generate experimental data and without noise). Each point is tagged with its corresponding AIC value.

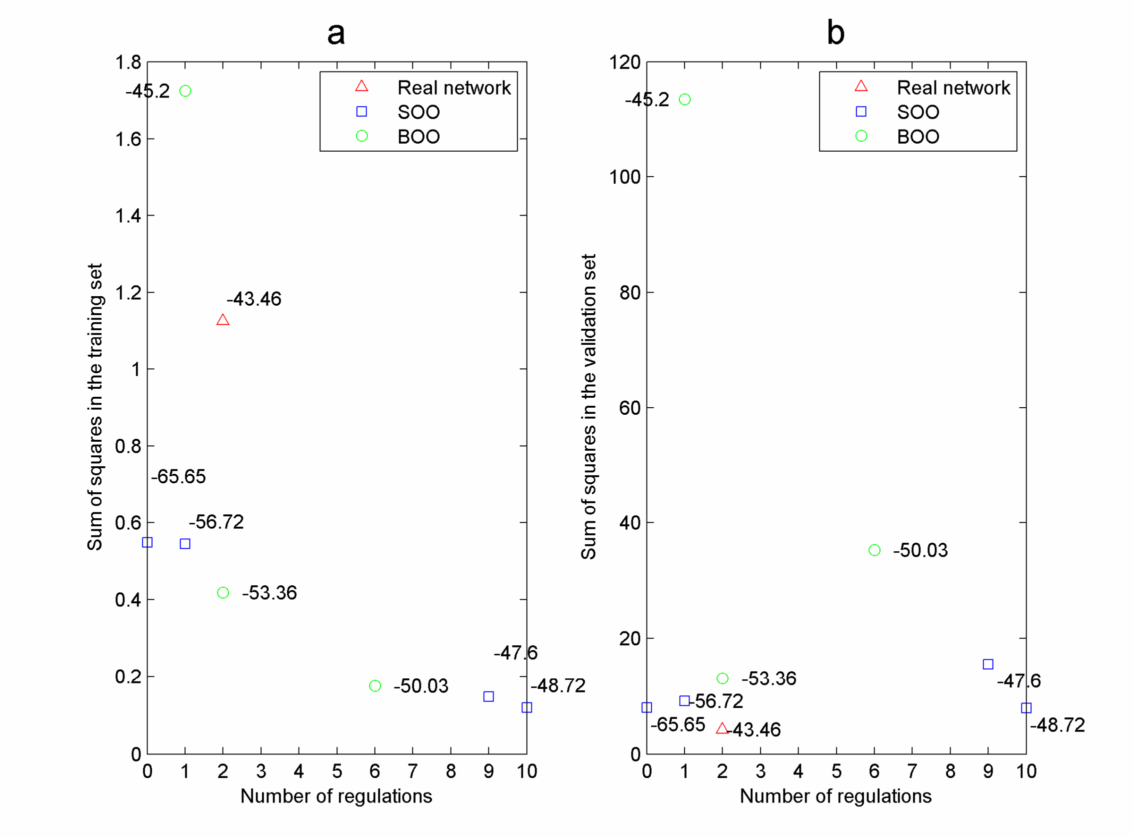

Figure 3 Number of regulations vs sum of squares in the training set (a) and in the validation set (b) for one experiment with 10% error. Blue squares refer to models generated with the single-objective optimization (SOO) approach whereas green circles represent models generated with the bi-objective optimization (BOO) scheme. Red triangles depict the real metabolic network (i.e., the one we used to generate experimental data and without noise). Each point is tagged with its corresponding AIC value.

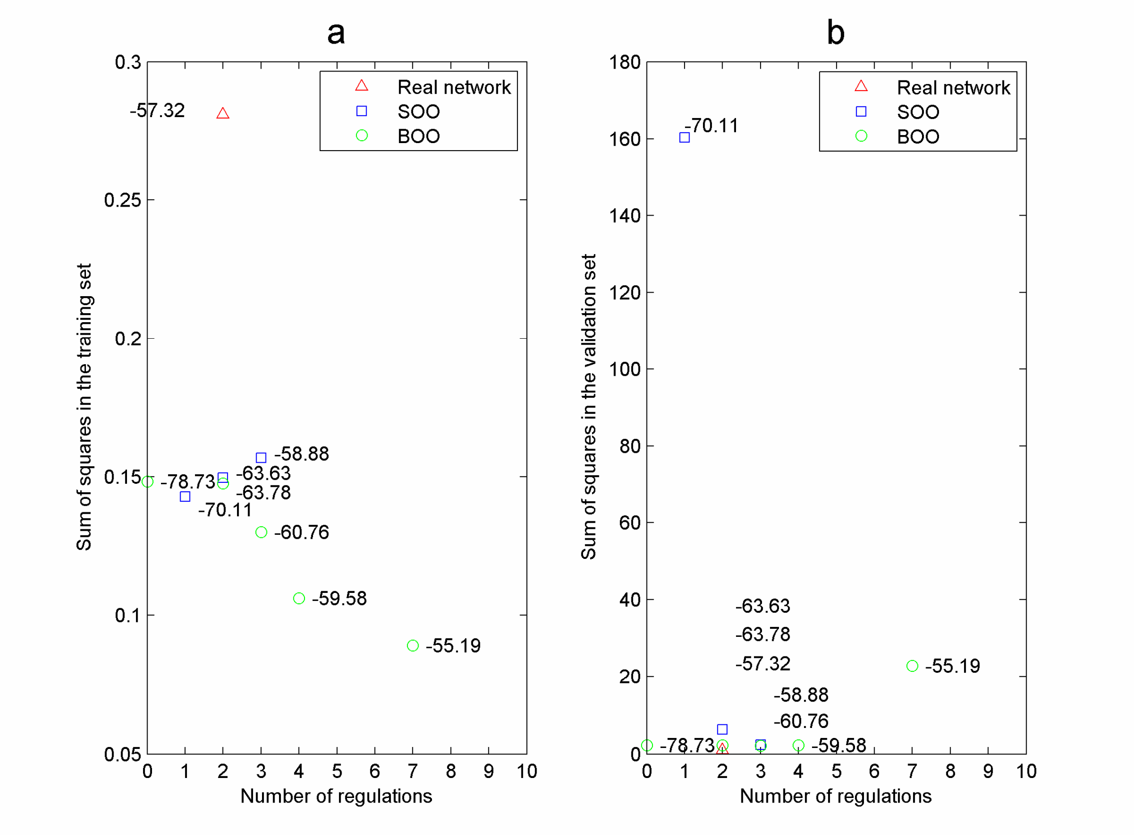

Figure 4 Number of regulations vs sum of squares in the training set (a) and in the validation set (b) for one experiment with 15% error. Blue squares refer to models generated with the single-objective optimization (SOO) approach whereas green circles represent models generated with the bi-objective optimization (BOO) scheme. Red triangles depict the real metabolic network (i.e., the one we used to generate experimental data and without noise). Each point is tagged with its corresponding AIC value.

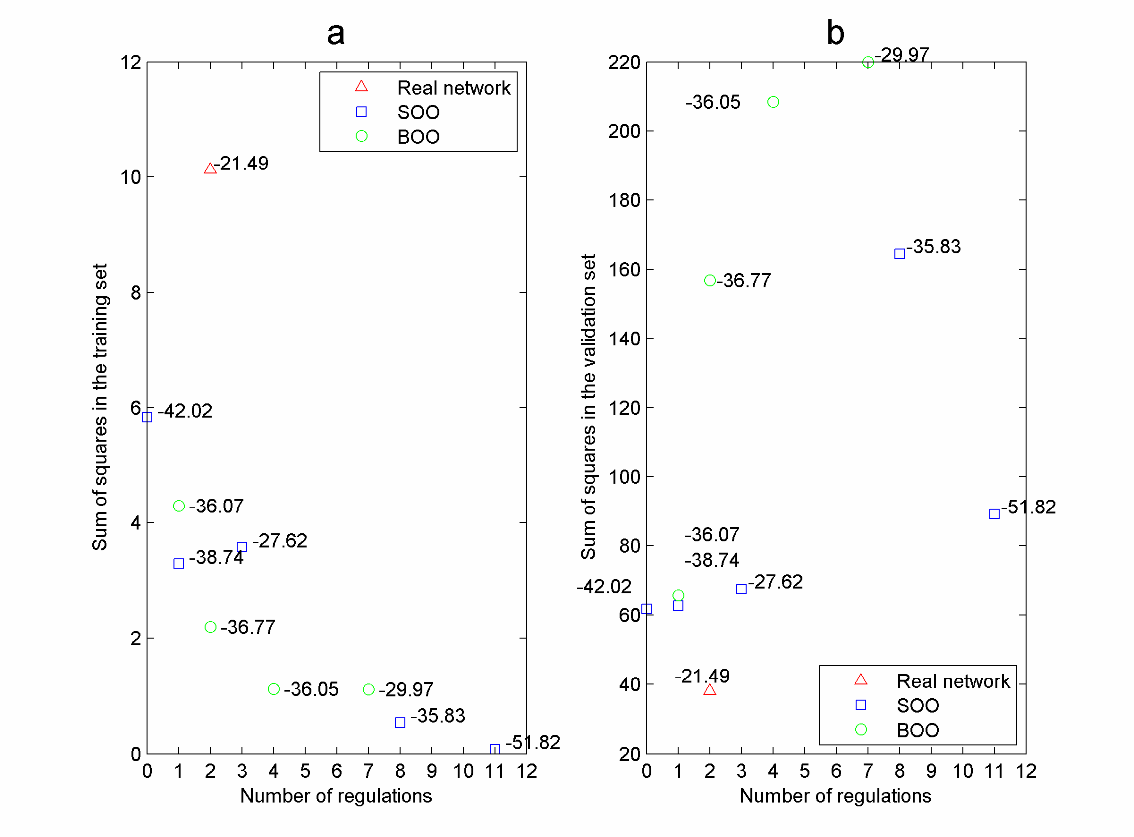

Figure 5 Number of regulations vs sum of squares in the training set (a) and in the validation set (b) for one experiment with 30% error. Blue squares refer to models generated with the single-objective optimization (SOO) approach whereas green circles represent models generated with the bi-objective optimization (BOO) scheme. Red triangles depict the real metabolic network (i.e., the one we used to generate experimental data and without noise). Each point is tagged with its corresponding AIC value.

In the case of a 5% of error (Figure 2b), however, the multi-objective optimization identifies models performing better in the validation data set than the best models obtained minimizing the AIC. In this case, the multi-objective approach identifies also a model with better AIC value than the ones produced by minimizing the AIC (See Table 1). Hence, the performance of each method depends on the case study, and there is no clear winner between the two.

Uncertainty (%)	**One Experiment**		
	Best AIC	*Approach*	*AIC real network*
5	-78.73	Epsilon constraint	-57.32
10	-65.65	AIC minimization	-43.46
15	-57.23	AIC minimization	-35.35
30	-51.82	AIC minimization	-21.49
	Best TS error		*Real network TS error*
5	0.089	Epsilon constraint	0.281
10	0.120	AIC minimization	1.126
15	0.061	AIC minimization	2.533
30	0.081	AIC minimization	10.132
	Best VS error		*Real network VS error*
5	2.17	Epsilon constraint	1.057
10	7.88	AIC minimization	4.226
15	16.53	AIC minimization	9.509
30	61.70	AIC minimization	38.034

Table 1 Summary of the main results obtained for one experiment

In general, one single experiment might not be enough for generating a reliable model. A better (*i.e.*, more accurate) model can be obtained using additional time-series data of the same system under different sets of initial conditions. Different initial concentrations of X_3 (0.2, 1.2, and 2.2) were considered to simulate three different experiments. These perturbations force the system to move across different dynamic regimes, producing additional information that constrains further the set of feasible network configurations.

The same procedure followed for one single experiment was applied. The points were split into a first partition containing 30 points (10 for each experiment in the training set) and a second one consisting of 165 points (5 replications of 11 points for each experiment in the validation set). The training set was used to produce a set of Pareto optimal models which were compared with those obtained minimizing the AIC (Figures 6-9). 10 Pareto solutions for each level of uncertainty (5, 10, 15 and 30%) were generated using the ε-constraint method (and compared with 10 models obtained minimizing the AIC values and using integer cuts).

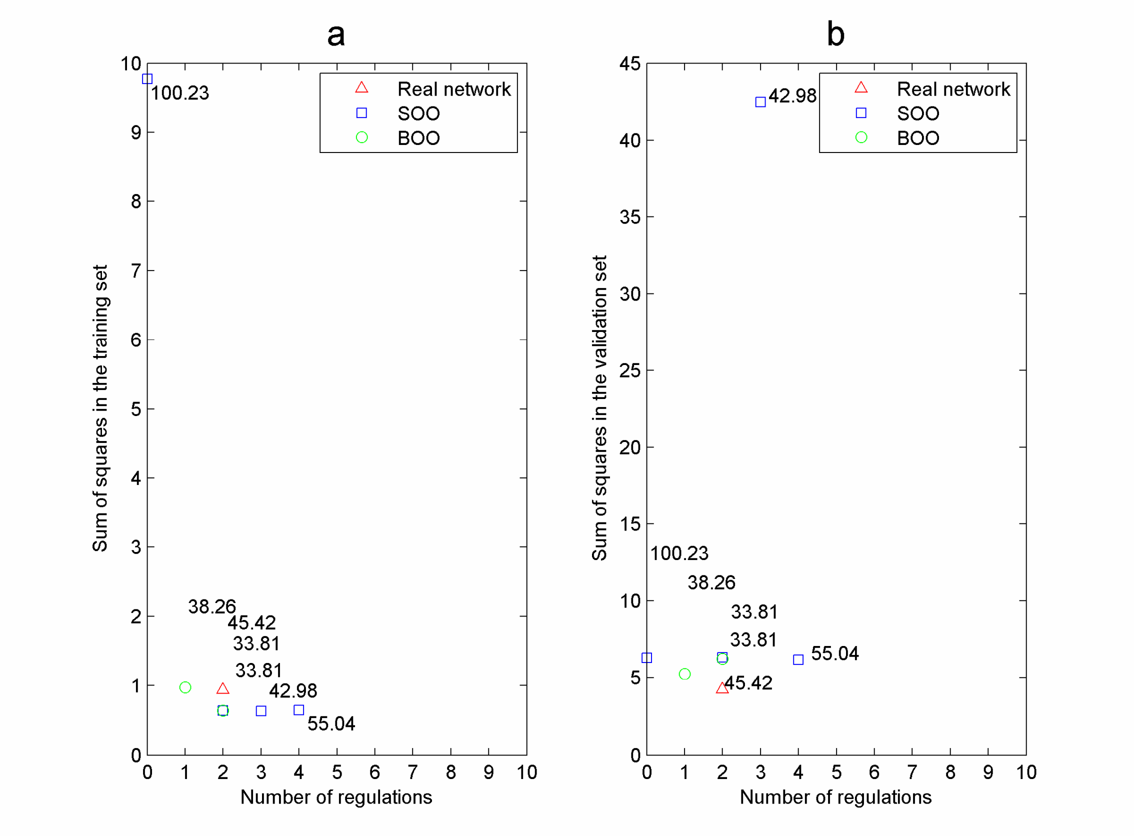

Figure 6 Number of regulations vs sum of squares in the training set (a) and in the validation set (b) for three experiments with 5% error. Blue squares refer to models generated with the single-objective optimization (SOO) approach whereas green circles represent models generated with the bi-objective optimization (BOO) scheme. Red triangles depict the real metabolic network (i.e., the one we used to generate experimental data and without noise). Each point is tagged with its corresponding AIC value.

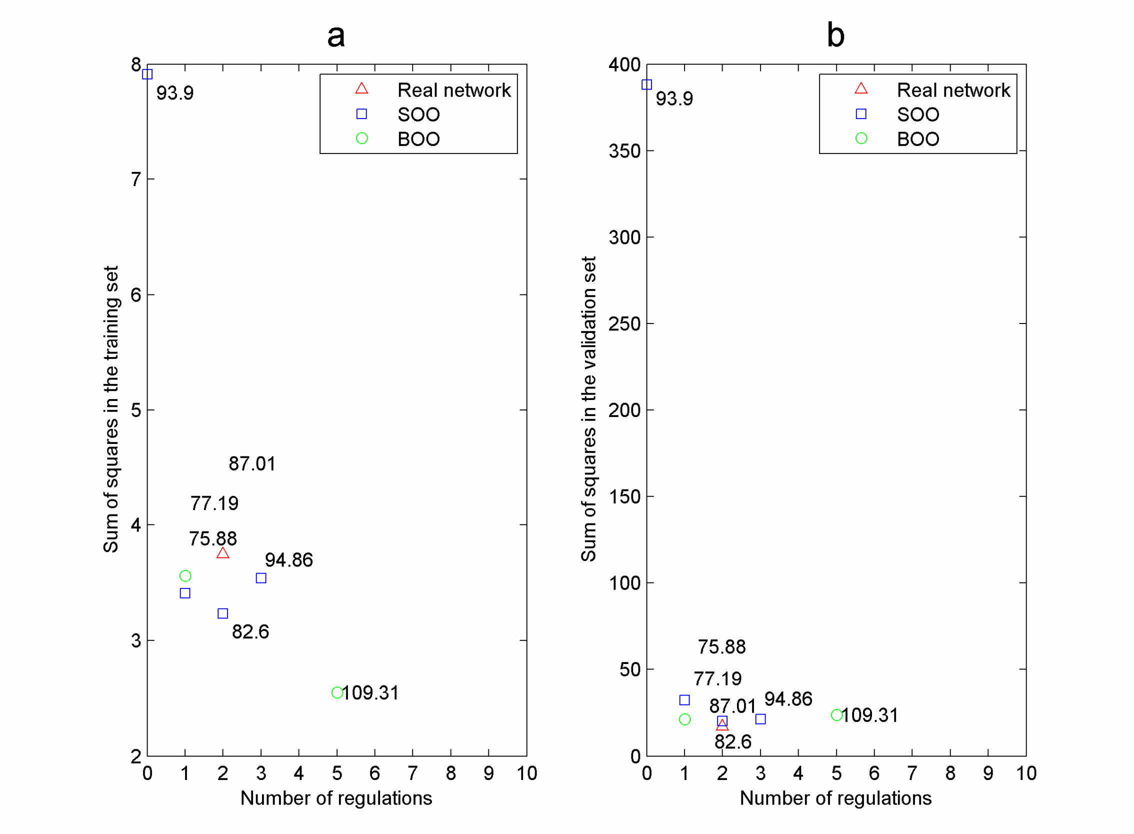

Figure 7 Number of regulations vs sum of squares in the training set (a) and in the validation set (b) for three experiments with 10% error. Blue squares refer to models generated with the single-objective optimization (SOO) approach whereas green circles represent models generated with the bi-objective optimization (BOO) scheme. Red triangles depict the real metabolic network (i.e., the one we used to generate experimental data and without noise). Each point is tagged with its corresponding AIC value.

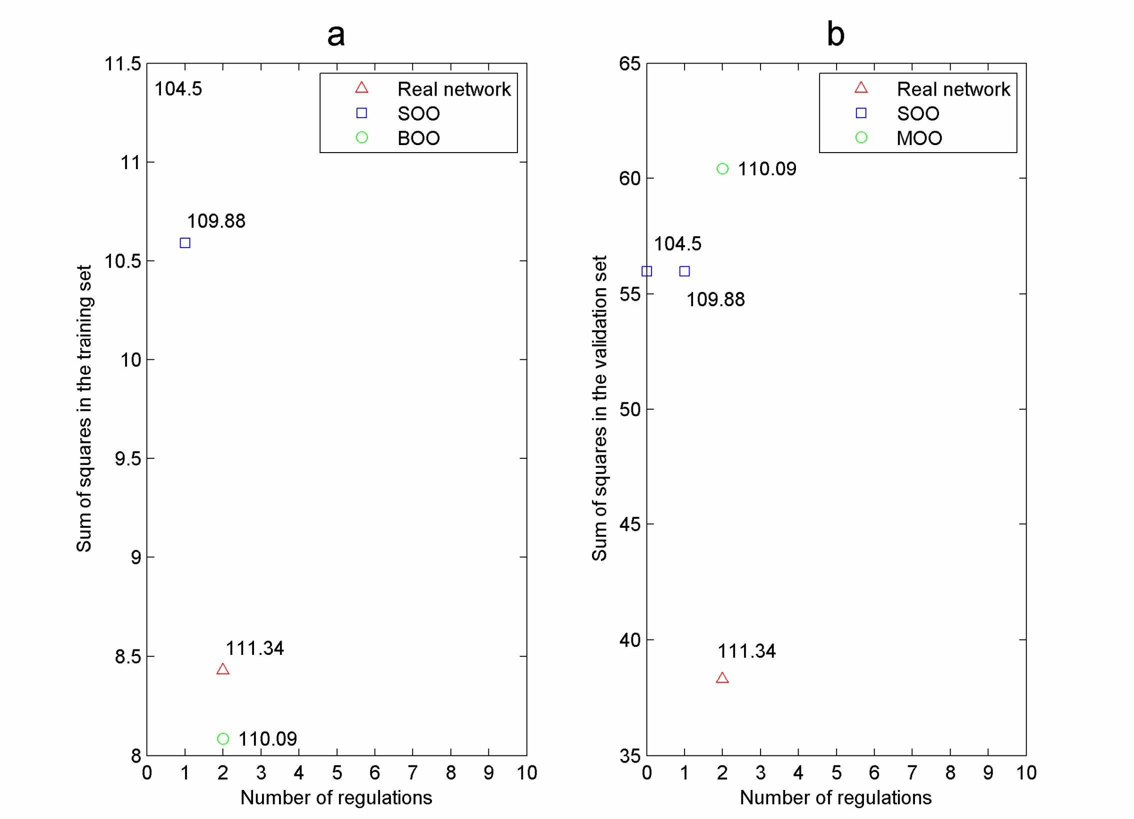

Figure 8 Number of regulations vs sum of squares in the training set (a) and in the validation set (b) for three experiments with 15% error. Blue squares refer to models generated with the single-objective optimization (SOO) approach whereas green circles represent models generated with the bi-objective optimization (BOO) scheme. Red triangles depict the real metabolic network (i.e., the one we used to generate experimental data and without noise). Each point is tagged with its corresponding AIC value.

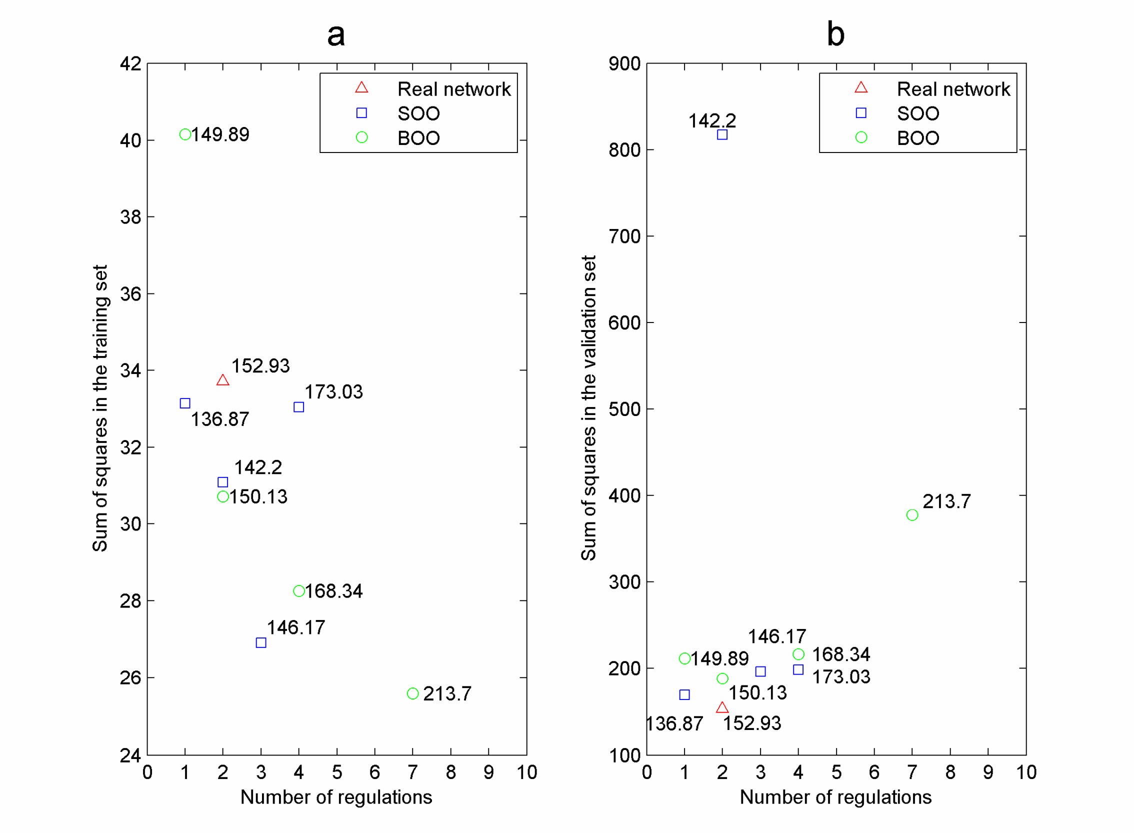

Figure 9 Number of regulations vs sum of squares in the training set (a) and in the validation set (b) for three experiments with 30% error. Blue squares refer to models generated with the single-objective optimization (SOO) approach whereas green circles represent models generated with the bi-objective optimization (BOO) scheme. Red triangles depict the real metabolic network (i.e., the one we used to generate experimental data and without noise). Each point is tagged with its corresponding AIC value.

The single-objective MINLP models solved by the epsilon constraint method were again implemented in GAMS and solved with SBB in the same computer. In this case, the MINLPs feature 67 binary variables, 967 continuous variables and 980 equations. The solution time was in the order of few minutes for each simulation.

For three experiments and 5% error (see Figure 6b), the ε-constraint method was able to identify a model with lower error in the validation set than those obtained minimizing the AIC. In the other cases, the minimization of the AIC led to models with lower residuals in the validation set (see Figures 7-9). The models identified minimizing the AIC show, as expected, better AIC values than those generated with the bi-objective model. Nevertheless, the AIC values obtained by the ε-constraint method are very close to the minimum AIC value identified by the single-objective approach. Particularly, for the case of 5% of error, (Figure 6b), the multi-objective optimization approach identifies the best solution in terms of AIC value. We summarize the results in Table 2.

Uncertainty (%)	**Three Experiment**		
	Best AIC	*Approach*	*AIC real network*
5	33.81	Both	45.42
10	75.88	AIC minimization	87.01
15	104.5	AIC minimization	111.34
30	136.87	AIC minimization	152.93
	Best TS error		*Real network TS error*
5	0.63	AIC minimization	0.937
10	2.55	Epsilon constraint	3.746
15	8.09	Epsilon constraint	8.43
30	25.59	Epsilon constraint	33.72
	Best VS error		*Real network VS error*
5	5.27	Epsilon constraint	4.257
10	19.93	AIC minimization	17.028
15	55.98	AIC minimization	38.31
30	169.11	AIC minimization	153.25

Table 2 Summary of the main results obtained for three experiments

Recall that the model yielding the minimum AIC might not be the one showing better performance under new experimental data. To further investigate this idea, the AIC values are plotted against the errors in the validation set for the models generated by both, the minimum AIC and the ε-constraint approaches. As seen, there is no clear trend between the error in the validation set and the AIC value (see Figures 10 and 11). Remarkably, the real model shows worse AIC value and better CV value than other models, confirming the fact that the model with minimum AIC might not always be the best model. For this reason, it is recommended to keep models with similar (and always as low as possible) AIC values rather than just retaining the one with minimum AIC value and discarding the others. In addition to these models, we might be interested in keeping those models showing low residuals in the validation set.

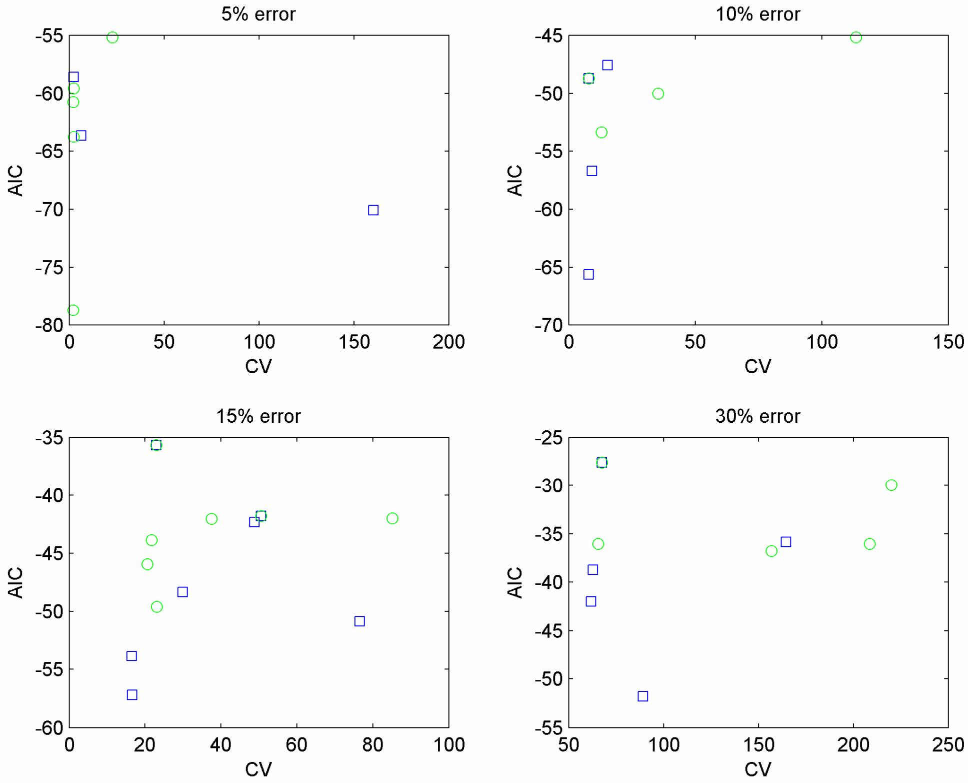

Figure 10 Relationship between the Akaike information criterion and cross validation values for one experiment. Blue squares refer to models generated with the single-objective optimization approach whereas green circles represent models generated with the bi-objective optimization scheme.

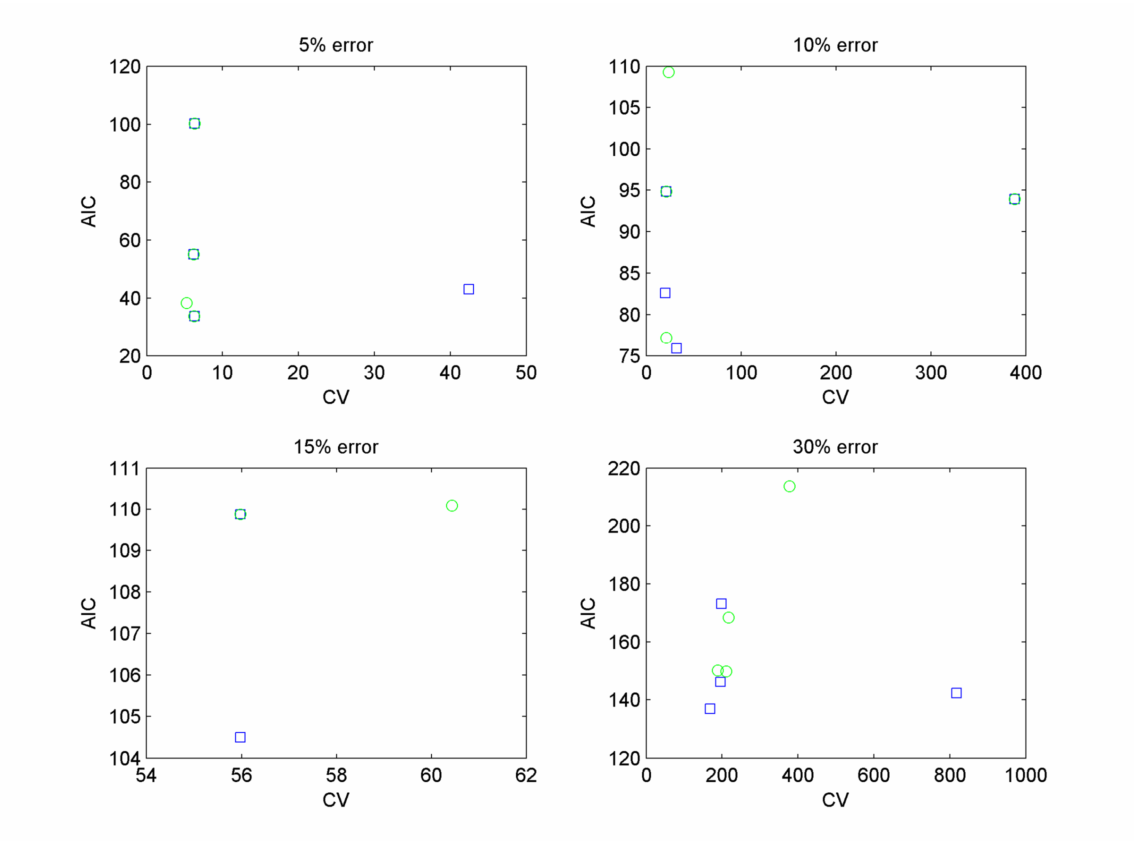

Figure 11 Relationship between the Akaike information criterion and cross validation values for three experiments. Blue squares refer to models generated with the single-objective optimization approach whereas green circles represent models generated with the bi-objective optimization scheme.

In general, models with lower AIC values contain fewer regulatory loops. In some instances, however, the single optimization approach produces overfitted models. For example, in 1 experiment with 15% of error, the algorithm identifies a model with 11 regulations (Figure 4b). This model shows a good AIC value, but it is clearly overfitted because it displays many regulations and shows a very poor predictive accuracy in the validation dataset (Figure 4b).

The minimum Akaike solution could be identified by solving a bi-criteria model with two objective functions: minimum residual and minimum complexity. Ideally, the minimum AIC solution should belong to the Pareto frontier of the corresponding bi-objective model. In our case studies, however, this occurs only in few cases. The reason for this is two-fold. On the one hand, we are using the corrected AIC (AICc) rather than the standard AIC (see the Methods section for further details). On the other hand, the models are not solved to global optimality, since a local optimizer is used in the calculations. Global optimization methods could have been applied, but in the case of MIDO problems they tend to lead to very large CPU times.

Conclusions

In this paper we have presented an alternative approach based on multi-objective optimization to simultaneously identify the regulatory interactions along with the kinetic parameters (assuming a given kinetic representation) from time series data. Our approach is based on a previous method that focused on minimizing the AIC as unique criterion. A bi-objective model is herein presented that seeks to minimize simultaneously the problem complexity and the residual, providing as output a set of Pareto optimal models representing the compromise trade-off between both objectives. This bi-objective model is solved by the epsilon constraint method, which calculates a series of single-objective MIDO models. These MIDOs are reformulated into MINLPs using orthogonal collocation on finite elements.

Ideally, the bi-objective approach should identify also the minimum AIC solution. However, this does not hold when using the corrected AIC (AICc) instead of the AIC. Furthermore, the nonconvexities of the MINLPs lead to multiple local optima in which local optimizers might get trapped during the search (on the other hand, the application of global optimization algorithms would lead to prohibitive CPU times due to the size and complexity of the model).

The bi-objective method presented provides as output alternative models with "good" AIC values and residuals in the validation set. From this set of models, experts should choose the best ones according to their biological knowledge of the system. Both approaches, the single-objective and bi-objective one show different performance depending on the case study, and there is no clear winner. Sometimes the single-objective identifies the solution with minimum AIC value, but others the bi-objective does it instead. Furthermore, the solution with minimum AIC value is not guaranteed to lead to the minimum residual. Again, sometimes the solution with minimum residual in the validation set is produced by the single-objective approach and some others with the bi-objective one.

Our approach assesses the compromise between the predictive capabilities of a model and the associated complexity. The results obtained confirm the theoretical observation that for small samples the model with minimum AIC value might not make the best predictions for new untested experimental data (i.e., other models with worst

AIC values in the sample might perform better). Hence, the AIC and other metrics based on information theory must be used with care when dealing with finite sets of points of small size. In practice, larger experimental data sets are required to produce better models, but this will eventually lead to more complex problems. In light of this, we recommend to calculate a set of models with low AICs and residuals, and further assess them taking into account their accuracy, complexity and additional biological knowledge of the system.

Methods

Given a reaction network along with a set of potential regulatory signals, and experimental time series data, the goal of the analysis is to infer the regulatory structure of a metabolic system and the associated kinetic parameter values. The concentration X_i of every metabolite *i* present in a metabolic network is assumed to vary with time *t* as a result of the action of *p* flows:

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

where X_i refers to the concentration of metabolite *i*, $\mu_{i,r}$ is the stoichiometric coefficient of metabolite *i* in process *r*, which is positive when reaction *r* produces metabolite X_i and negative when *r* consumes X_i, and v_r is the rate function of this process.

There are two key issues that arise in building an accurate model of the system under study. The first is the selection of an appropriate mathematical representation for v_r. The second is the definition of the regulatory structure of the system. In this work, an approximate formalism based on a standard mathematical representation is used to deal with these challenges [26].

Power-law models

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

$$v_r = \gamma_r \prod_{j=1}^{n+m} X_j^{f_{r,j}} \qquad r = 1,\ldots,p \tag{2}$$

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. This equation considers the effect of *n*+*m* metabolites (*n* dependent and *m* independent) on each reaction. This kinetic representation facilitates the task of simultaneously identifying the network topology and its corresponding kinetic parameters.

By substituting (2) into equation (1), we get what is known as a Generalized Mass-Action (GMA) model.

$$\dot{X}_i = \sum_{r=1}^{p} \left(\mu_{i,r} \gamma_r \prod_{j=1}^{n+m} X_j^{f_{r,j}} \right) \qquad i = 1,..,n \tag{3}$$

Our method, however, is not restricted to any particular kinetic formalism, as any general representation flexible enough to account for the regulatory interactions in a biological network can be used for the same purpose. The power law [13] and the saturable and cooperative formalisms are examples of such general kinetic representations [15].

Multi-objective optimization approach

The literature about multi-objective optimization in the context of parameter estimation is quite scarce. In these approaches more than one objective, typically conflicting, are minimized. Some authors have selected as objectives the concentration error, slope error and interaction measure [27], while others have optimized the least-squares error from dynamic or steady-state data [28].

This paper uses the ε-constraint method in order to simultaneously minimize the sum of squared deviations in the training set and the number of regulations. The tasks of parameter estimation and identification of regulatory interactions are both posed in mathematical terms as a multi-objective mixed-integer dynamic optimization (moMIDO) problem. This MIDO is reformulated into an equivalent multi-objective mixed-integer nonlinear programming (moMINLP) problem using orthogonal collocation on finite elements. This moMIDO takes the following form:

$$\begin{array}{ll}
\min\limits_{\gamma_r, f_{r,j}, y_{r,j}^-, y_{r,j}, y_{r,j}^+} & \left(\sum\limits_{u=1}^{k} \sum\limits_{i=1}^{n} \left(\hat{X}_{i,u} - \overline{X}_{i,u} \right)^2, \sum\limits_{j=1}^{n} \sum\limits_{r=1}^{p} \left(y_{r,j}^- + y_{r,j}^+ \right) \right) \\
s.t. & \dot{X}_i = \sum\limits_{r=1}^{p} \mu_{i,r} v_r \quad i = 1,\ldots,n \\
 & v_r = \gamma_r \prod\limits_{j=1}^{n+m} X_j^{f_{r,j}} \;\; r = 1,\ldots,p \\
 & \dot{X}_i \left(t_0 \right) = X_{0i} \quad i = 1,\ldots,n; \\
 & \overline{X}_{i,u} = X_i \left(t_u \right) \quad i = 1,\ldots,n; u = 1,\ldots,k; t \in \left[t_0, t_f \right] \\
 & f_{r,j} \le -\varepsilon + M \left(1 - y_{r,j}^- \right) \qquad j = 1,\ldots,n; r = 1,\ldots,p \\
 & -\varepsilon - M \left(1 - y_{r,j} \right) \le f_{r,j} \le \varepsilon + M \left(1 - y_{r,j} \right) \qquad j = 1,\ldots,n; r = 1,\ldots,p \\
 & \varepsilon \le f_{r,j} + M \left(1 - y_{r,j}^+ \right) \quad j = 1,\ldots,n; r = 1,\ldots,p \\
 & y_{r,j}^- + y_{r,j} + y_{r,j}^+ = 1 \qquad j = 1,\ldots,n; r = 1,\ldots,p \\
 & y_{r,j}^-, y_{r,j}, y_{r,j}^+ \in \left\{ 0,1 \right\}
\end{array} \tag{4}$$

The ε-constraint method entails solving a set of single objective MINLP problems in which one objective is treated as main objective function while the rest are transferred to auxiliary constraints in which upper bounds $\in_b^e$ are imposed on them using a set of epsilon parameters. In our case, the sum of squared errors is regarded as main objective, while the number of regulatory interactions is bounded using an auxiliary constraint. Hence, the single-objective mixed-integer nonlinear programming (soMIDO) problems are finally formulated as follows:

$$\begin{array}{ll}
\min\limits_{\gamma_r, f_{r,j}, y_{r,j}^-, y_{r,j}, y_{r,j}^+} & \sum\limits_{u=1}^{k} \sum\limits_{i=1}^{n} \left(\hat{X}_{i,u} - \overline{X}_{i,u} \right)^2 \\
s.t. & \sum\limits_{j=1}^{n} \sum\limits_{r=1}^{p} \left(y_{r,j}^- + y_{r,j}^+ \right) \le \varepsilon^e \quad e = 1,\ldots,E \\
 & \dot{X}_i = \sum\limits_{r=1}^{p} \mu_{i,r} v_r \quad i = 1,\ldots,n \\
 & v_r = \gamma_r \prod\limits_{j=1}^{n+m} X_j^{f_{r,j}} \;\; r = 1,\ldots,p \\
 & \dot{X}_i \left(t_0 \right) = X_{0i} \quad i = 1,\ldots,n; \\
 & \overline{X}_{i,u} = X_i \left(t_u \right) \quad i = 1,\ldots,n; u = 1,\ldots,k; t \in \left[t_0, t_f \right] \\
 & f_{r,j} \le -\varepsilon + M \left(1 - y_{r,j}^- \right) \qquad j = 1,\ldots,n; r = 1,\ldots,p \\
 & -\varepsilon - M \left(1 - y_{r,j} \right) \le f_{r,j} \le \varepsilon + M \left(1 - y_{r,j} \right) \qquad j = 1,\ldots,n; r = 1,\ldots,p \\
 & \varepsilon \le f_{r,j} + M \left(1 - y_{r,j}^+ \right) \quad j = 1,\ldots,n; r = 1,\ldots,p \\
 & y_{r,j}^- + y_{r,j} + y_{r,j}^+ = 1 \qquad j = 1,\ldots,n; r = 1,\ldots,p \\
 & y_{r,j}^-, y_{r,j}, y_{r,j}^+ \in \left\{ 0,1 \right\}
\end{array} \tag{5}$$

Single-objective optimization approach

A possible manner to avoid overfitting consists of minimizing the Akaike information criterion rather than the error between the experimental and *in silico* data. The AIC captures the trade-off between the number of parameters (*i.e.*, the complexity) of the model and its ability to accurately reproduce the experimental data. In a previous work [24], we developed an MIDO approach that minimizes the AIC, which is mathematically described as follows:

$$AIC = 2k + n\left(\log(2\pi) + \log\left(\sum_{u=1}^{k}\sum_{i=1}^{n}\left(\hat{X}_{i,u} - \bar{X}_{i,u}\right)^{2}\right) - \log \mathrm{n} + 1\right) \qquad (6)$$

Where AIC denotes the Akaike information criterion, *k* is the number of estimated parameters plus one (the standard deviation σ^2) and *n* is the number of experimental data points. Note that in the Results and Discussion section when we refer to complexity we use the number of regulations instead of *k*, with the correspondence between both being given by *k* = 2*p* + *number of regulations* + 1. Note that the factor 2 multiplying *p* is added to account for the fact that for each velocity *r* we need to find the values of two parameters: the apparent rate constant γ_r and the kinetic order of the substrate of the reaction f_{rj}.

When the ratio between the number of parameters to be estimated and the number of experimental data points is low (*i.e.*, n/k < ~40 [16]), it is recommended to use a corrective term, giving rise to the following corrected AICc expression:

$$AIC_C = AIC + \frac{2k(k+1)}{n-k-1} \qquad (7)$$

Burnham and Anderson [16] strongly recommend the use of the AICc in these instances because the standard AIC increases the probability of selecting models that have too many parameters, (*i.e.*, overfitting). In this manuscript, the AICc is used instead of the AIC, because the samples contain a small number of points that keep the MIDO in a manageable size, yet we refer to it simply as AIC for keeping the notation as simple as possible.

The single-objective parameter estimation problem can be finally posed in mathematical terms using the following MIDO (mixed-integer dynamic optimization) formulation where the objective function is the AICc:

$$\begin{aligned}
\min_{\gamma_r, f_{r,j}, y^-_{r,j}, y_{r,j}, y^+_{r,j}} \quad & 2k + n\left(\log(2\pi) + \log\left(\sum_{u=1}^{k}\sum_{i=1}^{n}\left(\hat{X}_{i,u} - \overline{X}_{i,u}\right)^2\right) - \log \mathrm{n} + 1\right) + \frac{2k(k+1)}{n-k-1} \\
s.t. \quad & \dot{X}_i = \sum_{r=1}^{p} \mu_{i,r} v_r \quad i = 1,\dots,n \\
& v_r = \gamma_r \prod_{j=1}^{n+m} X_j^{f_{r,j}} \quad r = 1,\dots,p \\
& \dot{X}_i\left(t_0\right) = X_{0i} \quad i = 1,\dots,n; \\
& \overline{X}_{i,u} = X_i\left(t_u\right) \quad i = 1,\dots,n; u = 1,\dots,k; t \in \left[t_0, t_f\right] \\
& f_{r,j} \le -\varepsilon + M\left(1 - y^-_{r,j}\right) \qquad j = 1,\dots,n; r = 1,\dots,p \\
& -\varepsilon - M\left(1 - y_{r,j}\right) \le f_{r,j} \le \varepsilon + M\left(1 - y_{r,j}\right) \qquad j = 1,\dots,n; r = 1,\dots,p \\
& \varepsilon \le f_{r,j} + M\left(1 - y^+_{r,j}\right) \quad j = 1,\dots,n; r = 1,\dots,p \\
& y^-_{r,j} + y_{r,j} + y^+_{r,j} = 1 \quad j = 1,\dots,n; r = 1,\dots,p \\
& y^-_{r,j}, y_{r,j}, y^+_{r,j} \in \left\{0,1\right\}
\end{aligned} \tag{8}$$

The MIDO algorithm is applied iteratively in order to identify a set of plausible regulatory topologies. The algorithm identifies first one solution encoded in a set of values of the binary variables. The model is then executed again, but this time adding an integer cut (valid inequality), which excludes the solutions identified so far in previous iterations. The integer cut is hence a valid inequality that takes the following form:

$$\begin{aligned}
& \sum_{(r,j)\in ONE^-_{it}} {y^-_{r,j}}^{it} + \sum_{(r,j)\in ONE_{it}} {y_{r,j}}^{it} + \sum_{(r,j)\in ONE^+_{it}} {y^+_{r,j}}^{it} \\
& \quad - \sum_{(r,j)\in ZERO^-_{it}} {y^-_{r,t}}^{it} - \sum_{(r,j)\in ZERO_{it}} {y_{r,j}}^{it} - \sum_{(r,j)\in ZERO^+_{it}} {y^+_{r,t}}^{it} \\
& \qquad \le \left|ONE^-_{it} + ONE_{it} + ONE^+_{it}\right| - 1 \\
& ONE^-_{it} = \left\{(r,j)\middle| {y^-_{r,t}}^{it} = 1 \text{ in the solution obtained in iteration } it\right\} \\
& ONE_{it} = \left\{(r,j)\middle| {y_{r,j}}^{it} = 1 \text{ in the solution obtained in iteration } it\right\} \\
& ONE^+_{it} = \left\{(r,j)\middle| {y^+_{r,j}}^{it} = 1 \text{ in the solution obtained in iteration } it\right\} \\
& ZERO^-_{it} = \left\{\left(r,j\right)\middle| {y^-_{r,j}}^{it} = 0 \text{ in the solution obtained in iteration } it\right\} \\
& ZERO_{it} = \left\{\left(r,j\right)\middle| {y_{r,j}}^{it} = 0 \text{ in the solution obtained in iteration } it\right\} \\
& ZERO^+_{it} = \left\{\left(r,j\right)\middle| {y^+_{r,j}}^{it} = 0 \text{ in the solution obtained in iteration } it\right\}
\end{aligned} \tag{9}$$

Where $\mathrm{ONE}_{\mathrm{it}}$ and $\mathrm{ZERO}_{\mathrm{it}}$ 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 to the MINLP, the algorithm is solved iteratively to obtain a given desired number of configurations. Hence, our algorithm produces as output a set of potential network configurations (encoded in the values of the binary variables) rather than a single topology. Note that the AIC values tend to increase as iterations proceed.

Theoretical connections between the minimization of the AIC and the bi-objective optimization

We discuss next the theoretical connections between our two methods. The bi-criteria model seeks to optimize simultaneously the residual and the problem complexity. The bi-objective function is as follows

$$MOF = \left(\sum_{j=1}^{n} \sum_{r=1}^{p} \left(y_{r,j}^{-} + y_{r,j}^{+} \right), \sum_{u=1}^{k} \sum_{i=1}^{n} \left(\hat{X}_{i,u} - \bar{X}_{i,u} \right)^{2} \right) \tag{10}$$

On the other hand, the single-objective model that minimizes the AIC optimizes the following objective function:

$$SOF = 2k + \mathrm{n} \log(2\pi) + n \log \left(\sum_{u=1}^{k} \sum_{i=1}^{n} \left(\hat{X}_{i,u} - \bar{X}_{i,u} \right)^{2} \right) - n \log \mathrm{n} + n \tag{11}$$

As shown, this objective function is composed of a fixed and a variable term. The fixed term $\mathrm{n} \log(2\pi) - n \log \mathrm{n} + n$ is constant that does not affect the optimization and could be therefore removed. The variable term can be further disaggregated into a term that depends on the model complexity (number of binary variables that take a value of one), and another one that depends on the residual (error between the experimental and the *in silico* data). In essence, the AIC assigns a weight to each of these terms in order to find a “good” compromise solution that properly balances both aspects of the model. Hence, the solution with minimum AIC should in principle belong to the Pareto front of the following bi-objective model:

$$\min_{\gamma_r, f_{r,j}, y_{r,j}^{-}, y_{r,j}, y_{r,j}^{+}} \left\{ 2k, n \left(\log(2\pi) + \log \left(\sum_{u=1}^{k} \sum_{i=1}^{n} \left(\hat{X}_{i,u} - \bar{X}_{i,u} \right)^{2} \right) - \log \mathrm{n} + 1 \right) \right\} \tag{12}$$

In which the model complexity and the logarithm of the residual are the objectives to be minimized. Note that minimizing the natural logarithm function is equivalent to minimizing the squared residuals, since the function is monotonically increasing function. This bi-objective model can be solved by any standard MOO algorithm, like the epsilon constraint or the weighted sum methods. The weighted sum relies on solving a set of single-objective auxiliary problems of the following form:

$$AO = w_1 O_1 + w_2 O_2 \tag{13}$$

Where the single aggregated objective AO that is minimized, consists of a linear combination of two objectives using the weights w_1 and w_2 In the case of the AIC, the objectives are O_1 = *k* and O_2 = *sum of squared residuals*, whereas the weights correspond to w_1 = *2* and w_2 = *n*. Note that the weighted sum method is unable to identify points in the nonconvex part of the Pareto set, and for this reason was not used in the calculations.

Figure 12 provides a graphical interpretation of the weighted sum method. In the figure, the weighted sum is represented by a straight line with slope $-w_2/w_1$. The minimization problem seeks to push this line towards the origin until it intersects the convex region on the boundary. The solution obtained by optimizing a given weighted combination is given by the intersection between the straight line and the curve that trades-off both objectives (i.e., the Pareto front).

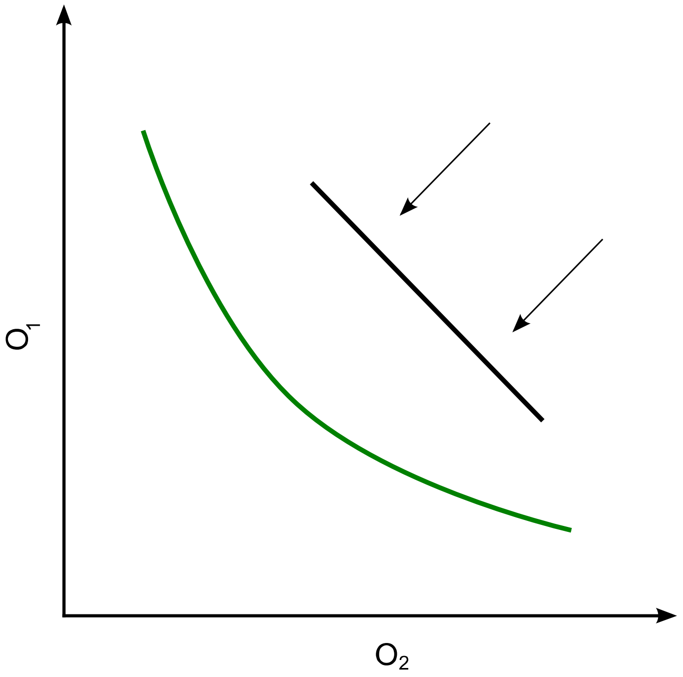

Figure 12. Weighted sum method applied to a bi-objective problem with a Pareto front of two conflicting objectives. The straight line with slope $-w_2/w_1$ is pushed towards the origin until it touches the Pareto set in at least one single point.

The AIC value is calculated using specific weight values assigned to the sum of squares and the complexity of a model. Hence, minimizing the AIC is equivalent to running one single iteration of the weighted sum method, and hence produces a Pareto solution lying on the convex part of the Pareto set.

We now show that the bi-criteria problem mentioned above is equivalent to the bi-objective model used in this paper. In fact, both models optimize the model complexity. As for the second objective, one model optimizes the residual and the other the logarithm of the residual. Since the logarithm is a monotonically increasing function, both objectives are indeed equivalent.

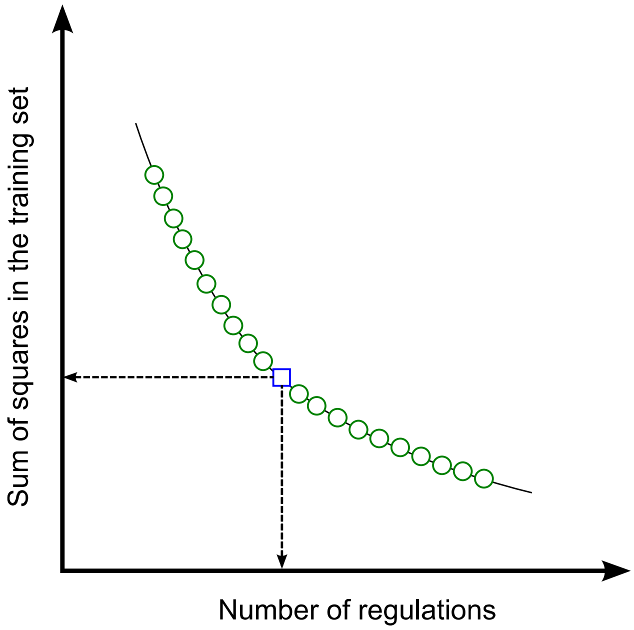

Figure 13. Pareto front of sum of squares in the training set vs. number of regulations. Green circles represent solutions obtained with the epsilon constraint method, whereas the blue square represents the solution obtained minimizing the AIC. All marks represent Pareto solutions.

Figure 13 illustrates the observation that the solution with minimum AIC value should belong to the Pareto front of the bi-objective problem solved in this paper. As observed, the minimum AIC value lies in the convex part of the Pareto front. Hence, it should be ideally identified using the bi-objective approach, which is a generalization of the single-objective one. Because of the nonconvexities present in the MINLP, local optimizers may get trapped in local optima. Hence, the solution with minimum AIC value provided by the local optimizer might not be globally optimal and, similarly, the bi-objective approach might be unable to identify the solution with global minimum AIC value. Ideally, the use of a global optimization package would avoid these problems, but unfortunately it tends to lead to large CPU times considering the size and complexity of our MINLP.

Note, however, that the above reasoning is only valid when the standard AIC formulation is used, since the term $2k + \frac{2k(k+1)}{n-k-1}$ of the AICc is not monotonically increasing in the entire domain. Hence, for the latter case, minimizing the AICc is not equivalent to simultaneously minimizing the complexity and the squared residuals. Further analysis of this term reveals that it is indeed monotonically increasing in the interval $0 \leq k \leq n$-1 (note that it has an asymptote at $k = n\text{-}1$). For this reason all the above conclusions are also valid when the AICc expression is used for k lower than n-1 or, what is the same, when *number of regulations* < n – 2p - 2. In this particularly case study, p = 6 and n = 10 for one experiment and n = 30 for three experiments. Hence, the solution of one experiment might not belong to the Pareto front k vs squared residuals when the AICc is used instead of the AIC, but it would belong for the case when *number of regulations* < -4. On the other hand, in the case of three experiments, when *number of regulations* < 16 (which are all) then we can still use the AICc with full guarantees. Note that these two conditions (i.e., *number of regulations < -4* or *< 16*) differ because so do the sample sizes.

Acknowledgments

The authors acknowledge the financial support of the following received from the Spanish Ministry of Education and Science (BES-2010-037166, CTQ2009-14420-C02-01, CTQ2009-14420-C02, CTQ2012-37039-C02, DPI2012-37154-C02-02).

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 USA 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. Chou I, Voit EO: **Recent developments in parameter estimation and structure identification of biochemical and genomic systems** Mathematical Biosciences 2009, 219 57–83.

10. M.A. Savageau, **Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions**, J. Theor. Biol. 1969, 25 365.

11. M.A. Savageau, **Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology**, Addison-Wesley Pub. Co., Advanced Book Program, Reading, Mass, 1976.

12. NV Torres, EO Voit: **Pathway Analysis and Optimization in Metabolic Engineering**, Cambridge University 2002, Cambridge, UK.

13. Voit E: Computational analysis of biochemical systems. **A practical guide for biochemists and molecular biologists.** Cambridge 2000, Cambridge University Press.

14. E.O. Voit: **Canonical Nonlinear Modeling. S-System Approach to Understanding Complexity**, Van Nostrand Reinhold, NY, 1991.

15. 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.

16. 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.

17. Mosteller F, Tukey, JW:**Data analysis, including statistics.** in G. Lindzey, and E. Aronson (eds.) Handbook of Social Psychology, Vol. 2. Addison-Wesley, Reading, MA. 1968

18. Stone M: **Cross-validatory choice and assessment of statistical predictions (with discussion).** Journal of the Royal Statistical Society 1974, Series B 39, 111–147.

19. Stone M: **An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion.** Journal of the Royal Statistical Society 1977, Series B 39, 44–47.

20. Geisser, S: **The predictive sample reuse method with applications.** Journal of the American Statistical Association 1975, 70, 320–328.

21. Akaike, H. (). **A new look at the statistical model identification.** IEEE 1974, Transactions on Automatic Control AC 19, 716–723

22. Schwarz G: **Estimating the dimension of a model.** Ann Stat 1978, 6:461–464.

23. M. Stone: **An asymptotic equivalence of choice of model by cross validation and Akaike's criterion** J. Roy. Statist. Soc. 1977, 39 (1977), pp. 44–47.

24. Guillén-Gosálbez G, Miró A, Alves R, Sorribas A, Jiménez L: **Identification of regulatory structure and kinetic parameters of biochemical networks via mixed-integer dynamic optimization** BMC Systems Biology 2013, 7:113

25. Voit EO, Almeida J: **Decoupling dynamical systems for pathway identification from metabolic profiles** 20, 11 2004, pages 1670–1681

26. 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. identification from metabolic profiles. Bioinformatics 2004, 20:1670–1681.

27. Liu PK, Wang FS: **Inference of biochemical network models in S-system using multiobjective optimization approach.** Bioinformatics Vol. 24 no. 8 2008, pages 1085–1092.

28. Barroso MFS, Takahashi RHC, Aguirre LA: **Multi-objective parameter estimation via minimal correlation criterion** Journal of Process Control 17 (2007) 321–332.

Appendices

1. List of publications

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.

Guillén-Gosálbez G, Miró A, Alves R, Sorribas A, Jiménez L: **Identification of regulatory structure and kinetic parameters of biochemical networks via mixed-integer dynamic optimization.** BMC Systems Biology 2013, 7:113.

Pozo C, Miró A, Guillén-Gosálbez G, Sorribas A, Alves R, Jiménez L: **Global optimization of hybrid kinetic/FBA models via outer-approximation.** Computers & Chemical Engineering 2014 (in press: Ms. Ref. No.: CACE-D-14-00025R1)

Miró A, Pozo C, Guillén-Gosálbez G, Jiménez L: **Bi-objective mixed-integer dynamic optimization approach for identifying the regulatory structure and kinetic parameters of biochemical networks.** BMC Systems Biology 2014 (Submitted)

2. Congress contributions

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. *European Symposium on Computer Aided Process Engineering – 22, London (UK),* 2012.

Guillén-Gosálbez G, Miró A, Alves R, Sorribas A, Jiménez L: Identification of regulatory structure and kinetic parameters of biochemical networks via mixed-integer dynamic optimization. American Institute of Chemical Engineering 2013 Annual Meeting, San Francisco (USA), 2013.