Computers in Biology and Medicine 63 (2015) 301–309

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

# Modeling multiple experiments using regularized optimization: A case study on bacterial glucose utilization dynamics

András Hartmann [a,b], João M. Lemos [a], Susana Vinga [b,*]

[a] INESC-ID/Instituto Superior Técnico, Universidade de Lisboa, Portugal – R Alves Redol 9, 1000-029 Lisboa, Portugal
[b] LAETA, IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Portugal – Av. Rovisco Pais, 1049-001 Lisboa, Portugal

ARTICLE INFO

*Article history:*
Received 14 April 2014
Accepted 28 August 2014

*Keywords:*
Identification
Optimization
Regularization
Modeling
Particle swarm optimization
Biochemical systems
Bacterial metabolism

ABSTRACT

The aim of inverse modeling is to capture the systems' dynamics through a set of parameterized Ordinary Differential Equations (ODEs). Parameters are often required to fit multiple repeated measurements or different experimental conditions. This typically leads to a multi-objective optimization problem that can be formulated as a non-convex optimization problem. Modeling of glucose utilization of *Lactococcus lactis* bacteria is considered using in vivo Nuclear Magnetic Resonance (NMR) measurements in perturbation experiments. We propose an ODE model based on a modified time-varying exponential decay that is flexible enough to model several different experimental conditions. The starting point is an over-parameterized non-linear model that will be further simplified through an optimization procedure with regularization penalties. For the parameter estimation, a stochastic global optimization method, particle swarm optimization (PSO) is used. A regularization is introduced to the identification, imposing that parameters should be the same across several experiments in order to identify a general model. On the remaining parameter that varies across the experiments a function is fit in order to be able to predict new experiments for any initial condition. The method is cross-validated by fitting the model to two experiments and validating the third one. Finally, the proposed model is integrated with existing models of glycolysis in order to reconstruct the remaining metabolites. The method was found useful as a general procedure to reduce the number of parameters of unidentifiable and over-parameterized models, thus supporting feature selection methods for parametric models.


## 1. Introduction

In the area of biochemical engineering, emerging analytical techniques such as Nuclear Magnetic Resonance (NMR) deliver increasing amount of experimental biological data containing important information about the system of interest. A great technological advancement is in vivo NMR [1], which makes possible to monitor metabolism in living cells by tracking the concentration of specially marked metabolites,[1] allowing to better understand their complex physiology. Indeed, dynamic modeling of the metabolism became one of the main research areas of systems biology due to the expected impact in areas such as metabolic and genetic engineering.

A typical but yet unsolved problem is the modeling of glucose utilization in bacteria, a highly regulated process, in which the external sugar is transported through the membrane into the cell [2]. Accurately modeling this first step is of high relevance since it is usually the first reaction of the bacterial metabolism pathway and to which all the other metabolites are highly dependent [3,4]. Most modelers are focusing on the inverse problem, namely to identify the parameters of a set of differential equations that best fit available experimental datasets [5]. However, the majority of these models lack generalization capabilities, i.e., even if a perfect fit to a single experiment is achieved, they cannot explain the systems' behavior in different experimental conditions.

In this context, multi-objective PSO (MOPSO) was extended with term-wise decomposition, using a generalized mass action (GMA) model [6]. This ODE model approximates the reaction rates with power-laws, which in this case are decoupled into an equation system by replacing derivatives with the values of the observed slopes. However, it is known that this type of decomposition is highly sensitive to noise [7], which might lead to poor estimates in real noisy settings. Another approach is to apply MOPSO directly for global modeling [8], and combining the multiple objectives with dynamic weighting. The present work greatly expands this previous proposal but a more throughout

* Corresponding author. Tel.: +351 218 419 504; fax: +351 218 498 097.
*E-mail addresses:* ahartmann@kdbio.inesc-id.pt (A. Hartmann), jlml@inesc-id.pt (J.M. Lemos), svinga@dem.ist.utl.pt (S. Vinga).

[1] Usually, the 6th carbon in glucose is switched to ($^{13}C$) isotope, which is tracked through the metabolic network.

http://dx.doi.org/10.1016/j.compbiomed.2014.08.027
0010-4825/

analysis of the parameter solution space is performed, along with an improvement on the utilization dynamic model. Moreover, regularization is introduced in the optimization procedure, thus leading to a contrasting perspective regarding the overall fitting methodology.

In this paper the multiple objectives are extended with regularization, which penalize the deviances between the parameters on different experiments. The aim is to develop and identify a model that can accurately describe different experimental conditions of bacterial glucose utilization and simulate/predict novel experiments. This paper is organized as follows: in Section 2 we introduce the dataset, the model and our method of identification, in Section 3 the results of bacterial glucose utilization is described in detail. The methods and the results are further discussed in Section 4 and finally in Section 5, conclusions are drawn.

## 2. Methods

Here we first describe the dataset, then the model is introduced, and finally we briefly review the particle swarm optimization (PSO) algorithm used for the identification together with the objective functions including the regularization approach.

### 2.1. The dataset

In vivo NMR measurements opened new horizons for systems biology, allowing measurement of metabolite concentrations in the living cell [1]. In the case of *Lactococcus lactis* the extracellular glucose concentration is measured and the glucose transport has to be modeled. Three perturbation datasets were used, where a bolus of 20, 40 and 80 mM ($^{13}$C) labeled glucose was given to starving bacteria in anaerobe conditions. It was observed that the multiple bolus experiments do not differ much from the single bolus regarding the shape of the glucose decay. The data was made publicly accessible in BGFit, a biological data management and curve fitting system [9] (http://kdbio.inesc-id.pt/bgfit). Some models using similar datasets were proposed previously [4,10–14], integrated in wider dynamic models for the glycolytic pathway.

### 2.2. Identification method

Numerous methods were proposed in the literature for fitting ODE models to biochemical and genomic systems [5]. Because of global optimality and its stochastic nature, here the particle swarm optimization (PSO) [15] was chosen for parameter identification. PSO is a population based stochastic optimization method inspired by the collective intelligence of simple interacting individuals. The traditional example for such systems is a bird flock seeking for food. The birds do not know the explicit location of the food, but their distance from it, this corresponds to the objective function. Communication is a key issue of the method, sharing knowledge with the other members of the flock allows them to follow the bird closest to the food.

In practice, PSO is initialized with a set of possible solutions, called particles $S_i(0)$ and associated random velocities $v_i(0)$. In every iteration $k$ the speed $v_i(k)$ and location $S_i(k)$ of each particle in the parameter space is updated as

$$v_i(k) = wv_i(k-1) + c_1 r_1 (pbest_i - S_i(k-1)) + c_2 r_2 (gbest - S_i(k-1)) \quad (1)$$

$$S_i(k) = S_i(k-1) + v_i(k), \quad (2)$$

where $w$ is the inertia describing the impact of the previous velocity to the current one. The positive constants $c_1$ and $c_2$ correspond to the acceleration rate towards the local and global optima respectively. $r_1$ and $r_2$ are independent, uniformly distributed random variables on the interval [0…1] ensuring the stochastic behavior of the method, $pbest_i$ is the best solution discovered by the $i$th particle and $gbest$ is the best global solution found. The particle velocities are lower and upper bounded as $v_{min} < v_i < v_{max}$. The method can be summarized in the following steps.

**Algorithm 1.** Particle swarm optimization (PSO) algorithm.

1. Initialize a set of particles of cardinality $N$
2. Evaluate the objective function for all the particles
3. Update $pbest_i$ for each particle $i = [1…N]$ and $gbest$
4. Compute the new velocities using Eq. (1)
5. Update the particles' position using Eq. (2)
6. Repeat from step 2 until the desired precision or the limit of iterations is reached

#### 2.2.1. Objective function

The objective function in Algorithm 1 can be expressed in terms of Mean Squared Error (MSE). This method was already successfully put into practice for inferring metabolic networks [16]. Single experiments are fitted using an objective function based on MSE, that corresponds to the normalized $\ell_2$ norm:

$$\mathcal{L}_d = \frac{1}{n_d} \sum_{t=1}^{n_d} \left(y_d(t) - \hat{y}_d(t)\right)^2, \quad (3)$$

where $y_d$ denotes the measured time-series with length of $n_d$, and $\hat{y}_d(t)$ is the estimated (reconstructed) measurements using the model $f(\cdot)$ and an estimated parameter set $\hat{\theta}$:

$$\hat{y}(t) = f(\hat{\theta}, t) \quad (4)$$

The index $d$ identifies the different experiments, here $d = 1,2$ and 3 represents the 20, 40 and 80 mM initial glucose concentrations, respectively. Here we do not aim at identifying all the metabolites dynamics, focusing only on glucose utilization using data from three different experiments simultaneously. In this context, it is known that the least absolute shrinkage and selection operator (Lasso) regularization [17] promote sparsity [18]. The idea behind introducing regularization is to impose sparsity in the sense that most of the parameters should not vary across the experiments and, therefore, can be considered global parameters. The remaining parameters should be dependent on the initial condition and will be described as a function of the initial sugar concentration.

The general form of the objective function including both the MSE terms and the regularization is

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \lambda(|\hat{\theta}_1 - \hat{\theta}_2|_1 + |\hat{\theta}_2 - \hat{\theta}_3|_1 + |\hat{\theta}_3 - \hat{\theta}_1|_1) \quad (5)$$

where $|\cdot|_1$ denotes for the $\ell_1$ norm, i.e. for a vector $x = [x_1 \ldots x_n]^T$, it is the sum of the absolute values of the elements $|x|_1 = \sum_{i=1}^{n} |x_i|$. The vector $\hat{\theta}_j$ with $j = 1,2,3$ represents the parameter vector estimated from experiment $j$. The regularization constant, $\lambda$, is arbitrary positive scalar, which is a parameter of the method. In this application $\lambda$ was chosen to be 10 since this value was found to balance optimally between the MSE and the regularization.

### 2.3. The model

Glucose utilization can be modeled using ODEs where the observations $y(t)$ correspond to the extracellular sugar concentration. One of the simplest model for this decay process involves the use of a pure exponential function, whose ODE formulation states that the derivative $\dot{y}(t)$ is proportional to $y(t)$ through a constant $k$:

$$\dot{y}(t) = -ky(t) \quad (6)$$

Other possibilities involve the use of logistic functions under a statistical non-linear mixed effects models framework [19]. In