Chemometrics and Intelligent Laboratory Systems 142 (2015) 293–303

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

MCR-ALS on metabolic networks: Obtaining more meaningful pathways ☆

A. Folch-Fortuny [a,*], M. Tortajada [b], J.M. Prats-Montalbán [a], F. Llaneras [c], J. Picó [d], A. Ferrer [a]

[a] Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, Edificio 7A, 46022 Valencia, Spain
[b] Biopolis, S.L., Parc Científic Universitat de València, c/ Catedrático Agustín Escardino Benlloch 9, 46980 Paterna, Spain
[c] Department of Electrical, Electronic and Automatic Engineering, Universitat de Girona, Campus de Montilivi, 17071 Girona, Spain
[d] Institut Universitari d'Automàtica i Informàtica Industrial, Universitat Politècnica de València, Camino de Vera s/n, Edificio 5C, 46022 Valencia, Spain

ARTICLE INFO

Article history:
Received 30 August 2014
Received in revised form 17 October 2014
Accepted 24 October 2014
Available online 1 November 2014

Keywords:
Pichia pastoris
Metabolic network
Grey modelling
Multivariate Curve Resolution-Alternating Least Squares

ABSTRACT

With the aim of understanding the flux distributions across a metabolic network, *i.e.* within living cells, Principal Component Analysis (PCA) has been proposed to obtain a set of orthogonal components (pathways) capturing most of the variance in the flux data. The problems with this method are (i) that no additional information can be included in the model, and (ii) that orthogonality imposes a hard constraint, not always reasonably. To overcome these drawbacks, here we propose to use a more flexible approach such as Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) to obtain this set of biological pathways through the network. By using this method, different constraints can be included in the model, and the same source of variability can be present in different pathways, which is reasonable from a biological standpoint. This work follows a methodology developed for *Pichia pastoris* cultures grown on different carbon sources, lately presented in González-Martínez *et al.* (2014). In this paper a different grey modelling approach, which aims to incorporate *a priori* knowledge through constraints on the modelling algorithms, is applied to the same case of study. The results of both models are compared to show their strengths and weaknesses.

1. Introduction

Systems Biology has become very popular during the last decade. Scientists with different backgrounds are nowadays working together in order to reach a systematic understanding of organisms. The impact of Systems Biology in biotechnological processes is so great that the term "industrial systems biology" is today very common within this kind of industries [1,2]. Measurement, monitoring, modelling and control (the so-called M3C methodology) are critical for obtaining high value-added biochemicals [3].

First principles-based models of microbial systems can be developed to describe the cell behaviour and achieve a predictive understanding of how they operate [4]. At a lower-intermediate degree of details, a cell can be roughly described as a collection of metabolites, which are consumed and produced dynamically by a set of biochemical reactions occurring within the cell and also being exchanged with their environment. These systems can be represented as directed graphs, or, in fact, directed hypergraphs, which are called metabolic networks.

Metabolic networks are used to represent an organism metabolism and its growth [5,6]. These networks are modelled assuming that certain constraints rule at steady-state, such as environmental constraints [7], regulatory constraints [8,9], gene expression data [10], mass balances or reaction irreversibilities [11] (the so-called *constraint-based perspective*) [12,13]. The imposed constraints define a solution space that encloses all the possible states of the network (*i.e.* flux distributions through the reactions).

A limitation of these type of models based solely on the fundamental information available is that other aspects will remain unknown, and some of their underlying assumptions (*e.g.* specific kinetics of the reaction system, unknown dynamics, values of the model parameters, objective functions) may not be valid for all the metabolic possible states of the network [13–15]. To face this limitation, hybrid (grey) models can be useful [16]. They combine knowledge-based models (which fit the theoretical, well-known phenomena), and empirical models (which fit any remaining systematic variation).

In the context of grey modelling, there are different approaches to decompose the data into the three types of variation (known causes, unknown causes and residuals) [17]. In the previous work [13], a model based on known constraints was imposed. In this way, the first principles-based model of the yeast *Pichia pastoris* was combined with experimental measurements of the external fluxes found in the literature. Defining the flux across each reaction in the network as a variable, Principal Components Analysis [18] (PCA) was used to obtain a set of uncorrelated components, representing groups of reactions, associated to the relevant biological functions of the cell. However, two problems arise when one applies PCA on metabolic networks: (i) no extra, available knowledge can be included in the model, and (ii) the components (pathways) have to be orthogonal among them.

☆ Selected paper presented at the 3rd European Conference on Process Analytics and Control Technology, 6–9 May 2014, Barcelona, Spain.

* Corresponding author.
E-mail address: abfolfor@upv.es (A. Folch-Fortuny).

http://dx.doi.org/10.1016/j.chemolab.2014.10.004
0169-7439/

In order to overcome these drawbacks, a different grey modelling approach is presented here, based on incorporating the fundamental knowledge through constraints on the modelling algorithms using the Multivariate Curve Resolution (MCR) technique. This is a flexible method for multivariate modelling, being its Alternating Least Squares version [19] (MCR-ALS) one of its most used iterative versions. MCR focuses on describing the evolution of the experimental multicomponent measurements through their underlying component contributions [20], without imposing hard-to-accomplish constraints from a chemical, physical or biological point of view, as orthogonality in the components. This methodology has been applied to other different types of data, such as spectral data [21,22], chromatographic data [23], hyperspectral data for multivariate image analysis [24], microarray data [25] or dynamic MRI data [26].

This paper completes the work developed for *P. pastoris* cultures grown on different carbon sources [13] by using MCR-ALS to obtain the set of biological pathways through the cell. This method permits to include modelling constraints, both from biological and mathematical points of view, in the optimisation algorithm. Another advantage of MCR-ALS is that, as opposed to PCA, the obtained pathways can share a single source of variability, which is reasonable from a biological standpoint. The paper is organised as follows. Section 2 presents the metabolic network reconstruction of the yeast *P. pastoris* and the different scenarios used in the study. Section 3 describes the grey modelling approach, explaining briefly the common part with [13] and deeply the new methodology proposed here. This procedure is applied to the available data from *P. pastoris* in Section 4. MCR-ALS results are compared to PCA ones [13] in Section 5. Finally, some conclusions on the use of MCR-ALS method are shown in Section 6.

2. Materials

2.1. Metabolic network reconstruction

The methylotrophic yeast *P. pastoris* has become one of the most widely studied microorganisms, since its development in the early 1970s, as it is reportedly one of the most useful and versatile systems for heterologous protein expression [27]. Many factors have contributed to the increasing interest in this yeast: (i) its easy molecular genetic manipulation, (ii) its ability to produce foreign proteins at high levels, (iii) its capability to perform many eukaryotic post-translational modifications, and (iv) its commercial availability [28].

A constraint-based model, whose corresponding metabolic network is shown in Fig. 1, has been used throughout this work. The model represents the most significant features of *P. pastoris* metabolism, including the main catabolic pathways of the yeast, such as glycolysis, the citric acid (TCA) cycle, glycerol and methanol oxidation and fermentative pathways [29]. Anabolism is introduced through the pentose phosphate pathway and a general lumped biomass equation, according to which growth is assumed to depend exclusively on key biochemical precursors. Branch-point metabolites, such as NADH, NADPH, AcCoA, oxaloacetate and pyruvate, are considered in compartmentalised cytosolic and mitochondrial pools [30].

2.2. P. pastoris experimental data set

In this work, experimental data from several fermentation runs with different *P. pastoris* strains have been taken from the literature, defining the different scenarios considered for the subsequent statistical analysis. The 40 scenarios under study show different uptake rates of the substrates glucose, glycerol and methanol (see Fig. 2). Scenario A1 corresponds to a *P. pastoris* culture expressing the Fab fragment of the human anti-HIV antibody 3H6 [30]. Scenarios B1–B7 and C1–C2 correspond to cultures producing a lipase from *Rhizopus oryzae* (ROL) [31,32]. Scenarios D1–D10 have been taken from *P. pastoris* cultures expressing and secreting recombinant avidin [33]. Scenario E1 has been obtained from a macrokinetic model for *P. pastoris* expressing recombinant human serum albumin (HSA) [34]. Scenarios F1–F7 correspond to cultures of a *P. pastoris* strain genetically modified to produce sea raven antifreeze protein [35]. Scenarios G1–G10 have been extracted from *P. pastoris* cultures producing recombinant human chymotrypsinogen B [36]. Scenario H1 corresponds to the continuous fermentation of a *P. pastoris* strain for the extracellular production of a recombinant ovine interferon protein [37]. Finally, scenario I1 comes from the culture of a genetically modified *P. pastoris* strain to produce recombinant chitinase [38]. The experimental data for all these scenarios are given in Fig. 2.

At this point, a comment regarding the so-called "batch effects" is in order. These are defined as systematic non-biological variation between groups of samples (or batches) caused by experimental artefacts [39–42], which can be present when experimental data are collected. If replicates of the same scenario are available (*i.e.* several experimental runs with the same strain and same uptake rates for each substrate), the presence of batch effects could be removed. Otherwise, the bias introduced by the non-biological nature of this kind of effects may confound true biological differences [41], affecting the results of statistical analysis. In this study, the scenarios have no replicates (see Fig. 2). Hence, the variation observed among scenarios with the same strains will be (at least partially) due to variations in the substrate uptake rates, and will be of biological relevance. This fact, jointly with the scarcity of information about other experimentation conditions (temperature, media, *etc.*), does not allow us to straightforwardly confirm actual batch effects in data.

3. Methods

The methodology applied in this paper is detailed in Fig. 3. First, the constraint-based model of *P. pastoris* is combined with the experimental information found in the literature. These two sources of information are unified applying a Possibilistic consistency analysis. Then, Monte Carlo sampling is performed to obtain a large dataset of feasible flux distributions across the metabolic network. Finally, the MCR-ALS is applied on the dataset.

The main objective of this article is to compare the results between the grey modelling approach presented in a previous work [13], where Principal Component Analysis (PCA) and Missing data method for Exploratory Data Analysis (MEDA) were applied, and the new approach presented here, which is based on MCR-ALS modelling (see Fig. 3). Both approaches share the Monte Carlo sampling and the Possibilistic consistency analysis, as well as their results. However, these methods are described here for the sake of completion (see [13] for details in these methods).

3.1. Stoichiometric modelling

To build a constraint-based model, the stoichiometric information embedded in the metabolic network (*i.e.* metabolites or cofactors involved in each reaction) must be arranged into an $I \times J$ matrix $\mathbf{S}$ (the so-called stoichiometric matrix). The rows of this matrix represent the I metabolites, the columns represent the J metabolic reactions and each element (i,j) is the stoichiometric coefficient $S_{i,j}$ of the ith metabolite in the jth reaction. A value of $S_{i,j} = -1$ indicates that the ith metabolite is consumed by the jth reaction. In contrast, a $S_{i,j} = 1$ indicates that the ith metabolite is produced by the jth reaction. Finally, a value of $S_{i,j} = 0$ stands for the ith metabolite is not involved in the jth reaction.

The stoichiometric matrix is used in combination with the flux vector $\mathbf{v} = (v_1,\ldots, v_J)$ and the metabolite concentration vector $\mathbf{c} = (c_1,\ldots, c_I)$ to represent the mass balances through the metabolic network. This equation is expressed as:

$$\frac{d\mathbf{c}}{dt} = \mathbf{S}\cdot v = \mathbf{0} \tag{1}$$