MCR-ALS on metabolic networks: obtaining more meaningful pathways

A. Folch-Fortuny[a,1] , 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

[1]Correspondence to: A. Folch-Fortuny (abfolfor@upv.es)

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.

Keywords: *Pichia pastoris*; metabolic network; grey modelling; Multivariate Curve Resolution - Alternating Least Squares

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 cells 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 reactions 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.

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 Figure 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, oxalacetate and pyruvate, are considered in compartmentalised cytosolic and mitochondrial pools [30].

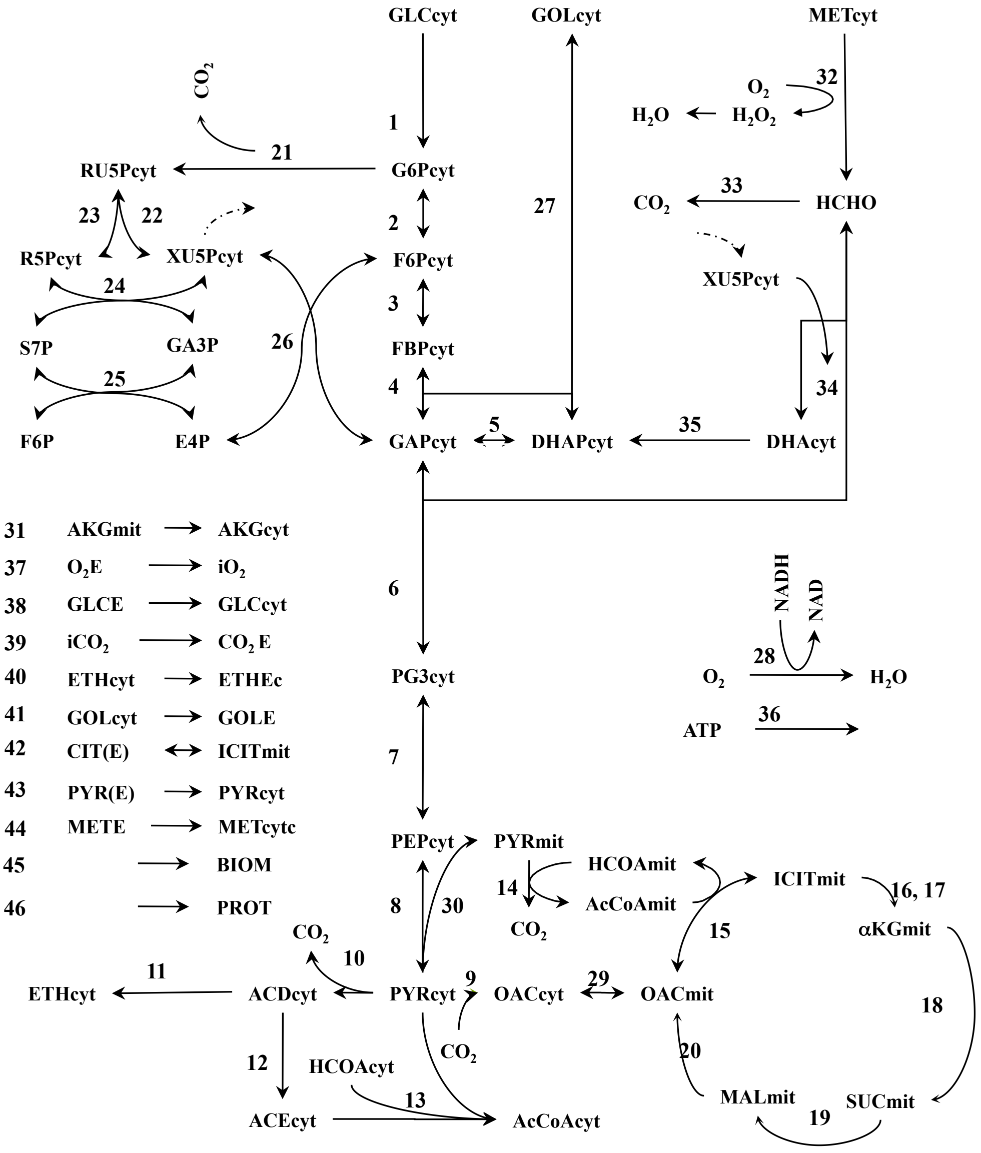
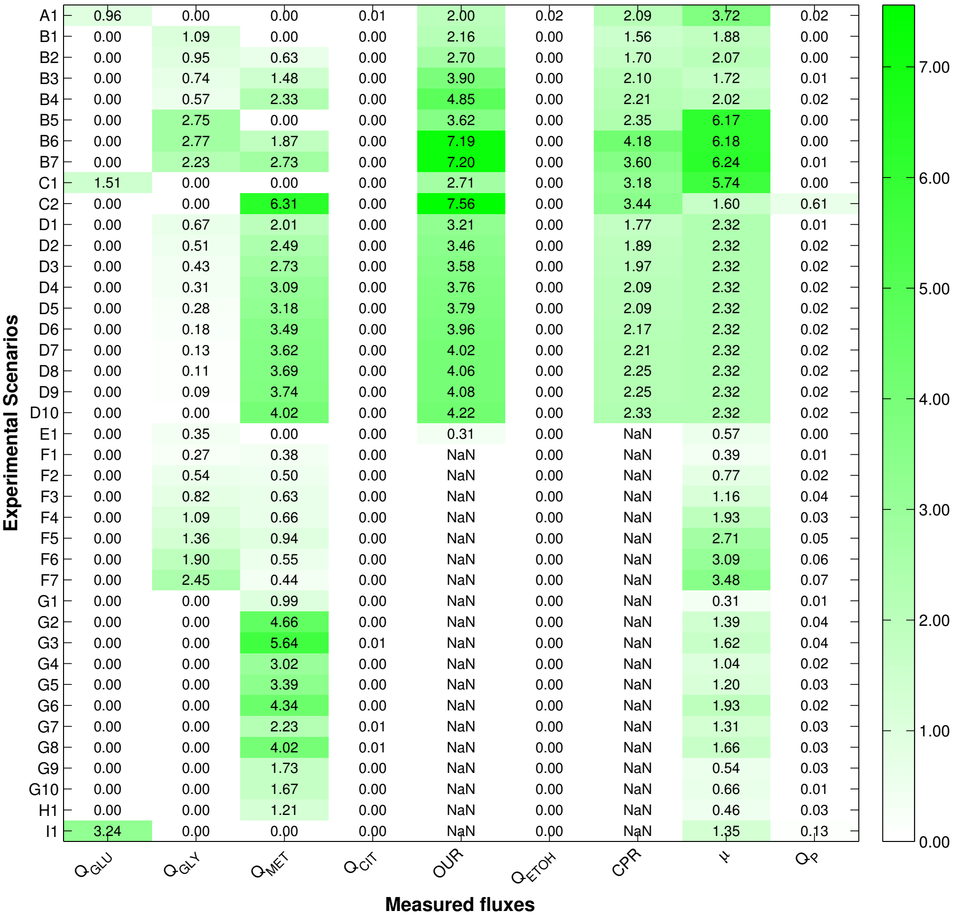


Figure 1: Metabolic network of *P. pastoris* used in this contribution, representing the central carbon metabolism of the yeast during growth on glucose, glycerol and methanol.

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 Figure 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 Figure 2.



	Q_{GLU}	Q_{GLY}	Q_{MET}	Q_{CIT}	OUR	Q_{ETOH}	CPR	μ	Q_P
A1	0.96	0.00	0.00	0.01	2.00	0.02	2.09	3.72	0.02
B1	0.00	1.09	0.00	0.00	2.16	0.00	1.56	1.88	0.00
B2	0.00	0.95	0.63	0.00	2.70	0.00	1.70	2.07	0.00
B3	0.00	0.74	1.48	0.00	3.90	0.00	2.10	1.72	0.01
B4	0.00	0.57	2.33	0.00	4.85	0.00	2.21	2.02	0.02
B5	0.00	2.75	0.00	0.00	3.62	0.00	2.35	6.17	0.00
B6	0.00	2.77	1.87	0.00	7.19	0.00	4.18	6.18	0.00
B7	0.00	2.23	2.73	0.00	7.20	0.00	3.60	6.24	0.01
C1	1.51	0.00	0.00	0.00	2.71	0.00	3.18	5.74	0.00
C2	0.00	0.00	6.31	0.00	7.56	0.00	3.44	1.60	0.61
D1	0.00	0.67	2.01	0.00	3.21	0.00	1.77	2.32	0.01
D2	0.00	0.51	2.49	0.00	3.46	0.00	1.89	2.32	0.02
D3	0.00	0.43	2.73	0.00	3.58	0.00	1.97	2.32	0.02
D4	0.00	0.31	3.09	0.00	3.76	0.00	2.09	2.32	0.02
D5	0.00	0.28	3.18	0.00	3.79	0.00	2.09	2.32	0.02
D6	0.00	0.18	3.49	0.00	3.96	0.00	2.17	2.32	0.02
D7	0.00	0.13	3.62	0.00	4.02	0.00	2.21	2.32	0.02
D8	0.00	0.11	3.69	0.00	4.06	0.00	2.25	2.32	0.02
D9	0.00	0.09	3.74	0.00	4.08	0.00	2.25	2.32	0.02
D10	0.00	0.00	4.02	0.00	4.22	0.00	2.33	2.32	0.02
E1	0.00	0.35	0.00	0.00	0.31	0.00	NaN	0.57	0.00
F1	0.00	0.27	0.38	0.00	NaN	0.00	NaN	0.39	0.01
F2	0.00	0.54	0.50	0.00	NaN	0.00	NaN	0.77	0.02
F3	0.00	0.82	0.63	0.00	NaN	0.00	NaN	1.16	0.04
F4	0.00	1.09	0.66	0.00	NaN	0.00	NaN	1.93	0.03
F5	0.00	1.36	0.94	0.00	NaN	0.00	NaN	2.71	0.05
F6	0.00	1.90	0.55	0.00	NaN	0.00	NaN	3.09	0.06
F7	0.00	2.45	0.44	0.00	NaN	0.00	NaN	3.48	0.07
G1	0.00	0.00	0.99	0.00	NaN	0.00	NaN	0.31	0.01
G2	0.00	0.00	4.66	0.00	NaN	0.00	NaN	1.39	0.04
G3	0.00	0.00	5.64	0.01	NaN	0.00	NaN	1.62	0.04
G4	0.00	0.00	3.02	0.00	NaN	0.00	NaN	1.04	0.02
G5	0.00	0.00	3.39	0.00	NaN	0.00	NaN	1.20	0.03
G6	0.00	0.00	4.34	0.00	NaN	0.00	NaN	1.93	0.02
G7	0.00	0.00	2.23	0.01	NaN	0.00	NaN	1.31	0.03
G8	0.00	0.00	4.02	0.01	NaN	0.00	NaN	1.66	0.03
G9	0.00	0.00	1.73	0.00	NaN	0.00	NaN	0.54	0.03
G10	0.00	0.00	1.67	0.00	NaN	0.00	NaN	0.66	0.01
H1	0.00	0.00	1.21	0.00	NaN	0.00	NaN	0.46	0.03
I1	3.24	0.00	0.00	0.00	NaN	0.00	NaN	1.35	0.13

Figure 2: Set of 40 experimental scenarios corresponding to *P. pastoris* chemostat cultures grown on glucose, glycerol and methanol mixtures. For each scenario, the

values of measured fluxes belonging to substrate and product specific consumption and production are shown. The substrates are glucose (Q_{GLU}), glycerol (Q_{GLYC}), methanol (Q_{MET}), citrate (Q_{CIT}) and oxygen (*OUR*). The products are ethanol (Q_{ETOH}), carbon dioxide (*CPR*), biomass (μ) and protein (Q_P). Note that NaN values stand for non-measured external fluxes.

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 Figure 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 Figure 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.

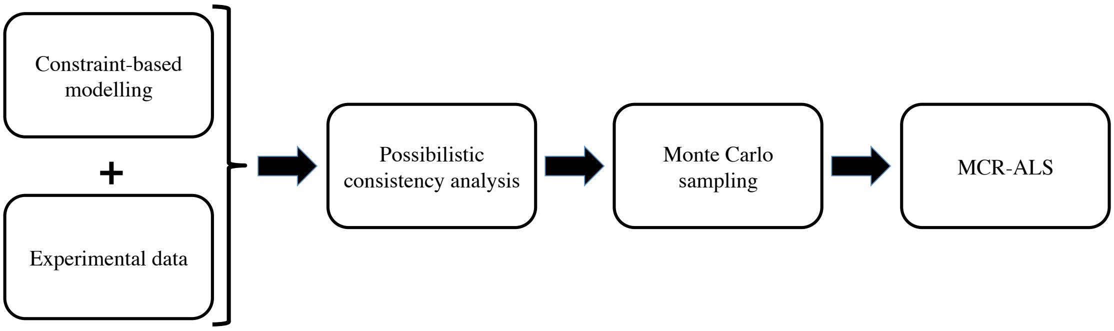

Constraint-based modelling
+
Experimental data
Possibilistic consistency analysis
Monte Carlo sampling
MCR-ALS

Figure 3: Flow diagram of the grey modelling applied in this paper.

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 Figure 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 **S** (the so-called stoichiometric matrix). Rows of this matrix represent the I metabolites, columns the J metabolic reactions and each element *(i,j)* 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 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 $\mathbf{v} = (v_1,..., v_J)$ and metabolites concentration vector $\mathbf{c} = (c_1,..., c_I)$ to represent the mass balances through the metabolic network. This equation is expressed as:

$$\mathrm{d}\mathbf{c}/\mathrm{dt} = \mathbf{Sv} = \mathbf{0} \qquad (1)$$

In stoichiometric modelling, the dynamic intracellular behaviour is disregarded on the basis assumption of pseudo-steady state for the internal metabolites [11].

In this work, the fluxes are assumed to flow in a single direction, so reversible fluxes in Figure 1 are split into two different ones. Therefore, instead of having 46 reactions (reversible and irreversible), now there are 63 irreversible ones (see the stoichiometric matrix **S** in Additional file).

Finally, a maximum value for each of the *J* fluxes is also imposed:

$$0 \leq v_j \leq v_{j,max} \tag{2}$$

The combination of the constraints imposed by Equations 1 and 2 defines a space (a bounded convex cone) of feasible steady-state flux distributions: only flux vectors that fulfil Equations 1-2 are considered valid cellular states. In this way, Equations 1-2 define our model of *P. pastoris*, following a constraint-based modelling approach [12,13,43,44].

3.2 Possibilistic consistency analysis

The simplest consistency analysis could be performed by checking that a set of measurements (or any other given flux state) fulfils the constraints imposed by the model [6] (Equations 1-2). However, this simple approach would be impractical because measurements are imprecise and do not exactly satisfy the constraints. Such difficulty is overcome by taking into account uncertainty as follows:

$$w_j = v_j + e_j \tag{3}$$

where e_j represents the deviation error between the actual fluxes v_j and the measured values w_j.

The consistency analysis can be also formulated as a possibilistic constraint satisfaction problem [45]. The basic idea is that a given flux vector compatible with the measurements and fulfilling Equations 1-2 will be considered as "possible", otherwise as "impossible". This can be refined to cope with measurements errors by introducing the notion of "degree of possibility" [46].

This degree of possibility provides an indication of the consistency between the model and the measurements. A possibility equal to one must be interpreted as complete agreement between the model and the original measurements. Lower values of possibility imply that some error in the measurements needs to be assumed to find a flux vector fulfilling the model constraints. For further details on this method, readers are referred to the original work [6,45] or the more detailed description of the Possibilistic consistency analysis performed in the previous paper [13].

3.3 Monte Carlo sampling method

The experimental data found in the literature represent partial flux solutions, because few fluxes of the metabolic network have been experimentally measured. In this context, Monte Carlo sampling methods can be used to produce complete feasible flux distributions across the cell without adding any other assumption nor biasing (*i.e.* keeping the current uncertainty) [13,47–51]. This way, the available experimental data (measured fluxes) and the first principles knowledge captured by the model (stoichiometry) are coupled together, providing a new richer dataset amenable to further analysis with a multivariate statistical method.

In order to deal with experimental errors, external fluxes are allowed to vary within a defined small range of values centred on the original measured value. Then, the unmeasured fluxes (internal) are allowed to take values within the boundaries that are imposed by putting together the constraints in Equations 1-2 and the limits imposed to the values of the measured fluxes. Further details on the sampling can be found in [13]. At this point, the feasible solutions for each scenario are obtained by sampling within the slice of the cone defined by Equations 1-2 and the experimental data constraints. The measured fluxes reduce the feasible solution space from the initial cone, which is bounded only by the constraint-based model, to the portion of it fulfilling these specific experimental measurements. The complete procedure is depicted in Figure 4.

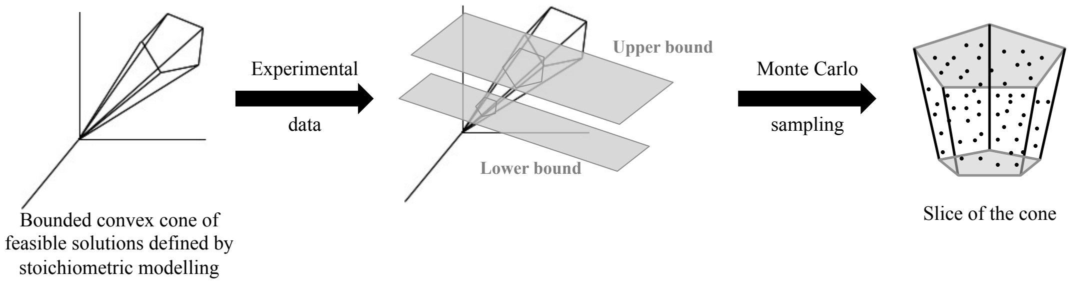


Figure 4: Monte Carlo sampling. The convex cone is obtained by Equations 1-2, the experimental measurements constrain the cone, and the sampling is performed on the resulting slice of the cone.

Notice that there are scenarios lacking measurements of some external fluxes (see Figure 2). In the Monte Carlo sampling, these fluxes are allowed to vary within the whole slice of the cone defined by the measured external ones and the constraint-based modelling.

3.4 Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS)

In this paper, a Multivariate Curve Resolution-Alternating Least Squares [19,25,52–54] (MCR-ALS) model is used. The reasons are its ability to provide physically more interpretable results by (i) imposing some *a priori* knowledge through constraints on the modelling algorithm, and (ii) avoiding the orthogonality restriction on the internal relationships between variables/pathways. The idea behind MCR, traditionally applied in analytical chemistry, can be easily expanded to flux analysis by stating that a flux distribution across the metabolic network for a particular scenario is a linear combination of the different pathways existing in it.

MCR-ALS is an iterative method that performs a bilinear decomposition of matrix **X** by means of an alternating least squares optimisation:

$$\mathbf{X} = \mathbf{C}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{4}$$

where **P** is a matrix containing in its columns each one of the metabolic pathways modelled, **C** gathers the relative contribution of each modelled pathway in each scenario, and **E** is a residual matrix.

MCR relies on the determination of the number of “real” metabolic pathways in a dataset. When some *a priori* knowledge about this number is available, it can be used as an initial guess. This knowledge can be checked by using some tool able to show up the relevant sources of information in the data. One possible way to do this is by applying PCA on the data set, and taking a look at the number of latent variables with the highest variance. Once the number of likely pathways present in the data is determined, they can be sought using some purity based algorithms [55–61]. In this work, since we have used the software available at the Multivariate Curve Resolution Homepage [62], the algorithm implemented in *Pure function* is applied.

Also, three additional constraints can be imposed to the MCR procedure. First, we have introduced a non-negativity constraint on the relative contributions and pathways. We have also imposed a closure constraint on the relative contributions [63,64]. This constraint is usually applied to closed systems, where the principle of mass balance is fulfilled. With this constraint, the sum of the contributions of the "real" pathways in each scenario (the elements in each row of the **C** matrix) is forced to be equal to a constant value, in our case, 1. Finally, depending on the biological information of each scenario, another constraint can be imposed: selectivity. This constraint forbids some pathways to be used in some scenarios, and it is applied by multiplying the relative contribution of each scenario by 0 if the corresponding pathway is not allowed to be active, and by 1 otherwise. The idea is to reduce the noise in data, avoiding inconsistent behaviours from a biological standpoint.

3.5 Software

All methods have been computed in Matlab environment (The Mathworks Inc., Natick, MA, USA). The Monte Carlo sampling method has been applied using the COBRA toolbox [65]. The MCR-ALS algorithms can be found at the Multivariate Curve Resolution Homepage [62].

4. Results

The results of the grey modelling approach of the yeast *P. pastoris* described in Methods section (see Figure 3) are discussed throughout the following subsections. The results of the Possibilistic consistency analysis and the Monte Carlo sampling are discussed here to ease the understanding of the general procedure; however, for a deeper explanation readers are referred to our previous work [13]. MCR-ALS procedure is explained here in full detail, pointing at the different models fitted and the problems found in each one.

4.1 Possibilistic consistency analysis

To perform the Possibilistic consistency analysis, the set of measured values for each scenario is compared with the stoichiometric modelling proposed in Section 3.1. Scenarios with NaN values for the external fluxes, *e.g.* F1 (see Figure 2) are analysed considering only the available values. The result of this analysis on each experimental scenario is a degree of possibility, between 0 to 1, reflecting a degree of consistency between each scenario and the biological model (Equations 1-2). With the aim of discriminating between "consistent" and "not consistent" scenarios, a minimum degree of consistency is imposed. In this way, 4 out of 40 original scenarios are classified as "not consistent", which implies that there are errors in the model regarding these scenarios, or more likely, that there are large measurement errors in those scenarios (the degree of possibility for each experimental scenario are given in [13]). These scenarios are B3, B4, C2, and E1. For this reason, these scenarios are not considered in the following analysis.

4.2 Monte Carlo sampling

Since only the external fluxes of each solution have been measured in the literature, the Monte Carlo sampling method is proposed to simulate different complete possible flux solutions. These sampled scenarios are consistent with the proposed model and the measured subset of fluxes. Once the sampling has been performed, the fluxes for each scenario are arranged by rows in a feasible flux solution matrix **X**. This matrix has the complete 3600 sampled flux solutions in its rows (36 scenarios × 100 samples) and the corresponding 63 flux values, including the protein production rate for each scenario, in its columns. The sampled fluxes for reactions 55-63 are zero for all scenarios. This implies that given the stoichiometric modelling and the available measured fluxes for each original scenario, it is not likely that these fluxes can be positive. Therefore, columns 55-63 of matrix **X** are removed.

4.3 MCR-ALS: Data considerations

In our previous work [13], the data were autoscaled (each column was mean-centred and divided by its standard deviation) in order to be analysed using PCA. Here, the data are not autoscaled. Considering that the aim is to estimate the percentage of usage of each pathway in each scenario, the columns of the dataset are scaled by its maximum value. Therefore, the flux values have a value 0 if the flux is not used in this particular scenario, and 1 if the flux is used at its maximum.

4.4 MCR-ALS: Initial estimation

Since the MCR-ALS method is an iterative approach of MCR, it needs an initial estimation of either pathways or relative contributions matrix to start the alternating least squares estimation of both matrices. MCR-ALS Toolbox [19] has implemented the *Pure* estimations method, which uses the most different rows or columns to estimate **P** or **C** matrix, respectively. Here, the pathways matrix is (initially) estimated using the most different scenarios in the dataset.

4.5 MCR-ALS: solution as PCA-MEDA approach

MCR-ALS needs, unlike PCA, the number of components (or pathways) to be extracted before running the algorithm. Since our main objective is to compare the results of the PCA+MEDA approach and the results of MCR-ALS, it makes sense to start the MCR-ALS algorithm with three pathways, which was the number of principal components in the previous paper. Additionally, the SVD estimation of the number of components indicates that 3-4 components describe well the data set.

As explained above, different constraints are imposed in the MCR-ALS algorithm to achieve the solution in the way defined in the previous subsections. The first constraint is that both the pathways and their relative contributions have to be positive. This is attained by the non-negativity constraint. Secondly, for each scenario, the relative contributions of pathways are forced to sum one, in order to represent a percentage of usage. This is the closure constraint, which is applied in the contributions direction (**C** matrix).

The variance in data explained by the MCR-ALS model is 78.5%. The pathways obtained in this first approach are represented graphically in Figure 5. Each row represents the weights of the original variables in each pathway: the clearer is the corresponding square the higher is the weight. These pathways are represented on the metabolic network in Figure 6.

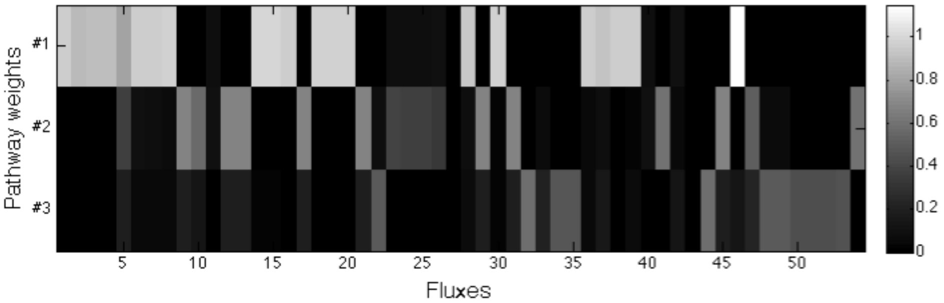


Figure 5: Pathways obtained extracting three components in the MCR-ALS method.

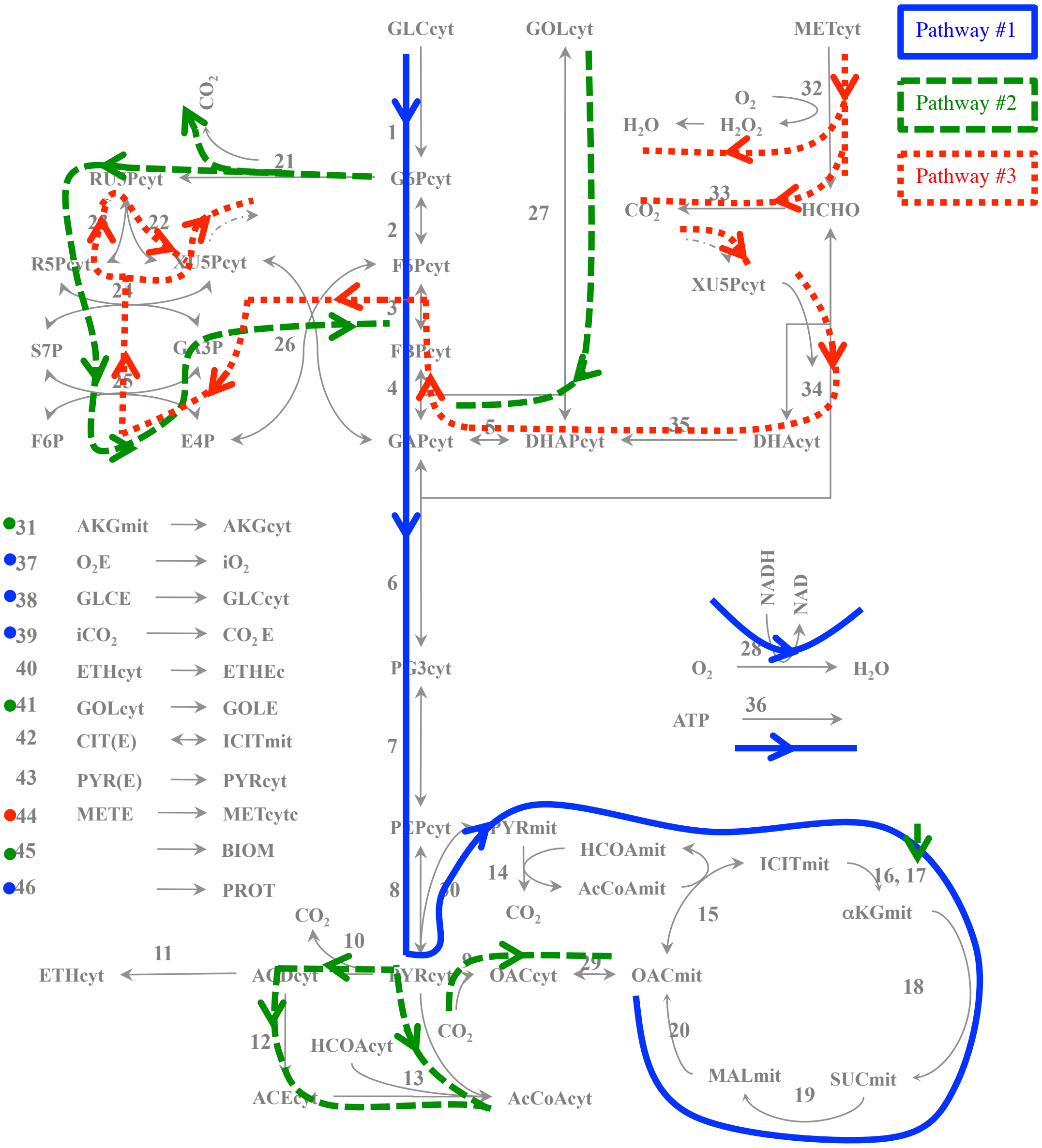


Figure 6: Metabolic network of *P. pastoris* with the three pathways obtained in the MCR-ALS method. The solid blue lines represent the first pathway, the dashed green lines the second one, and the dotted red lines the third one.

These results are in fact similar to the ones obtained in [13]: the first pathway is related to energy generation, in the form of ATP equivalents, mostly provided by glucose consumption through glycolysis and oxidative phosphorylation. The second pathway identified can be related to anabolism, and particularly to NADPH and AcCoA generation (thus indirectly to biomass growth) from glycerol. Finally, the third pathway seems to identify methanol consumption. Note that protein production is directly related to the first pathway as ATP is used as its single precursor in reaction 46. These pathways do not correspond exactly to the ones obtained in [13], especially in the case of the green (#2) and red (#3) pathways on the pentose phosphate route (reactions 21-26 in Figure 6), because the stoichiometric model was slightly different in that approach (*i.e.* the reversible reactions were not split into two irreversible ones).

The MCR-ALS approach allows studying the relationship between each scenario and the pathways obtained. This relationship is depicted in Figure 7. This figure shows three plots, the first one represents the percentage of usage of the first pathway in each of the 3600 scenarios. As well, the other two plots represent the percentage of usage of the second and third pathways, respectively. The first pathway is surprisingly not strongly associated to some scenarios (1-200) in which glucose is the only carbon source. In an analogous way, the third pathway is used nearly at 100% in scenarios in which methanol is consumed. The second pathway is contributing both to scenarios in which only glucose or glycerol are used as a substrate, despite the fact that (as shown in Figure 6) this pathway does not consume glucose.

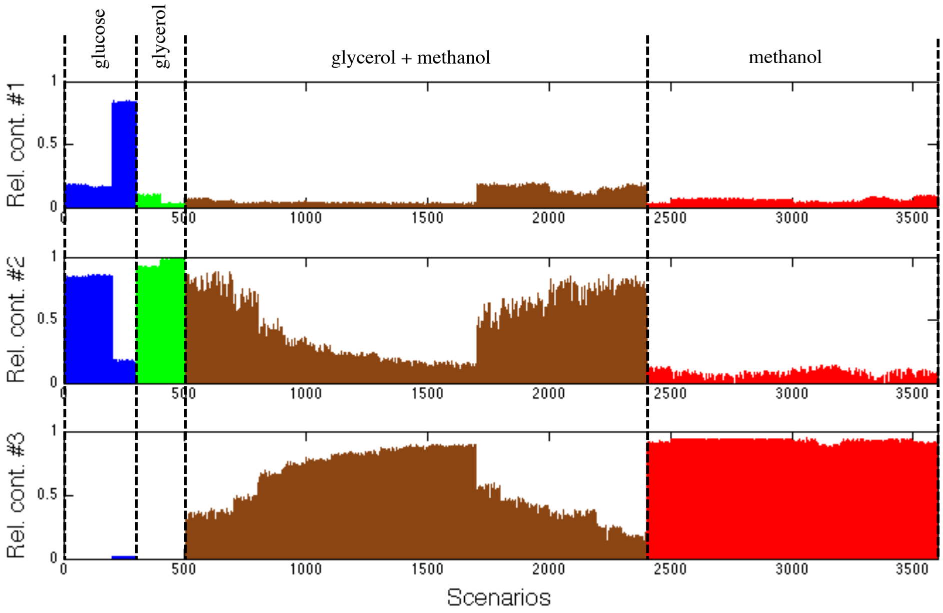


Figure 7: Relative contributions of the three pathways. The blue columns (scenarios 1-300) represent the percentage of usage of each pathway in glucose scenarios. The green columns (scenarios 301-500) represent the percentage of usage in glycerol scenarios. The brown ones (501-2400) are the scenarios with a glycerol-methanol mixture. The red columns (2401-3600) represent scenarios with only methanol as a substrate.

Once the relative contributions and the pathways have been visualised (Figures 5-7) some comments can be drawn. The second pathway depicted in Figure 6 does not flow in a thermodynamically feasible way through the metabolic network. The dashed green line crosses the pentose phosphate zone (reactions 22-26) and reaches reactions 3-4, where the glycerol consumption (reaction 27) ends in the opposite direction. This result, in addition with the poor contribution of the first pathway (solid blue) to the first two scenarios with glucose (1-200 in Figure 7), and the percentage of usage of the second pathway in glucose scenarios, indicates that the current model does not fully comprehend the behaviour of the scenarios analysed.

4.6 MCR-ALS: solution with four pathways

The results shown previously lead us to think that the actual MCR-ALS model may be improved by extracting another pathway, in order to discover if some of the pathways can be refined or if there is another pattern in the data that is not explained at the moment. So a new model with four pathways is fitted.

The model explains 82.4% of variance in data. The pathways obtained in this model are directly represented onto the metabolic network in Figure 8. The current first, third and fourth pathways are similar to the ones obtained in the previous MCR-ALS model (3 pathways). However, the second pathway represents a new metabolic route across the network.

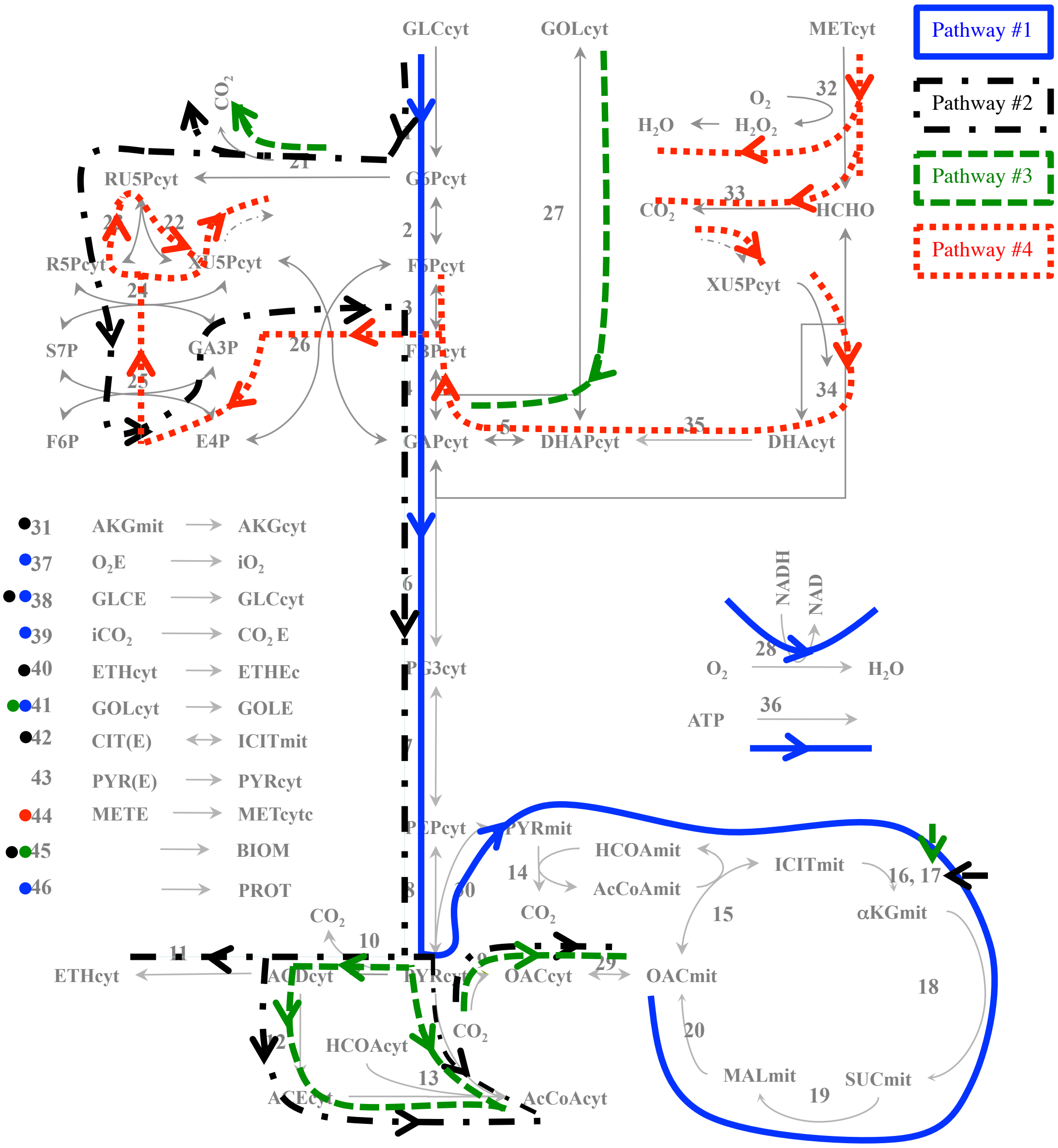


Figure 8: Metabolic network with four pathways. The solid blue lines represent the first pathway, the dash-dotted black lines the second pathway, the dashed green lines the third one, and the dotted red lines the fourth one.

The relative contribution of each pathway is plotted in Figure 9. Again, there is a plot for each pathway extracted from data. The first and second pathways seem to be associated mainly to glucose scenarios. The third pathway is widely used in the glycerol and glycerol+methanol scenarios, being the highest contribution attained in scenarios where glycerol is used as single carbon source. Finally, scenarios with only methanol use nearly at 100% the fourth pathway, and so do mixed scenarios with higher amount of this substrate.

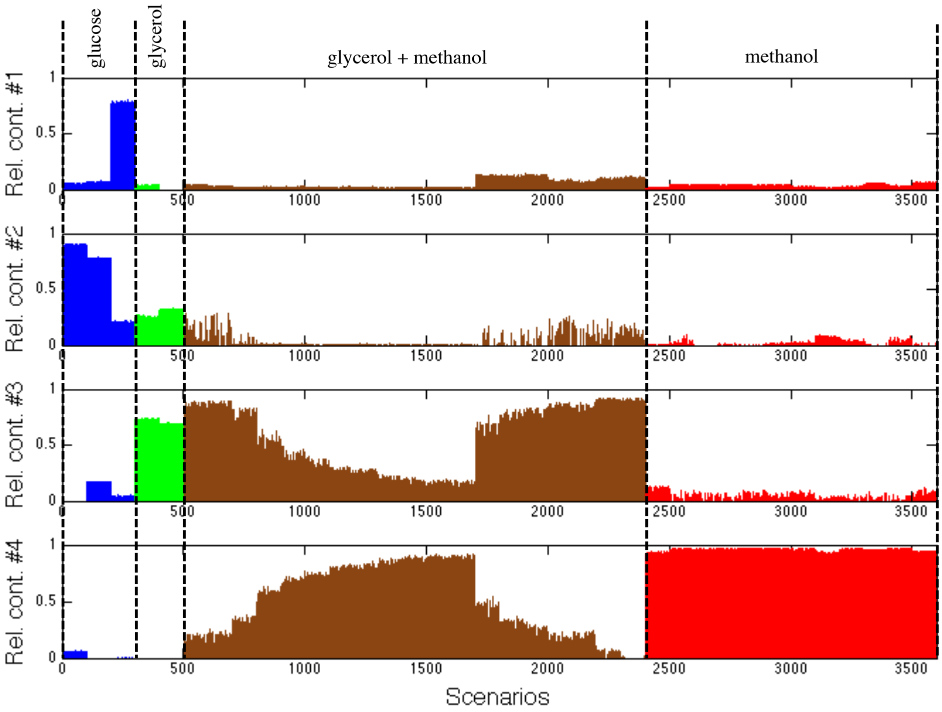


Figure 9: Relative contributions of the four pathways. The blue columns (scenarios 1-300) represent the percentage of usage of each pathway in glucose scenarios. The green columns (301-500) represent the percentage of usage in glycerol scenarios. The brown ones (501-2400) are the scenarios with a glycerol-methanol mixture. The red columns (2401-3600) represent scenarios with only methanol as a substrate.

The percentages of usage of the four pathways depicted in Figure 9 suggest that each one is dominating in a single type of scenario, *i.e.* in those where the substrate is glucose, glycerol, glycerol+methanol or only methanol.

As explained above, the flexibility of MCR-ALS method allows including different kind of constraints during the optimisation process. One of the most used constraints is selectivity. In this context, selectivity allows to constrain each pathway to not be used or "expressed" in all scenarios. By visual inspection of Figure 9 it seems that the first two pathways are related mainly to the glucose scenarios, the third one to glycerol and glycerol+methanol ones, and the last one to glycerol+methanol and methanol. This hypothesis is supported by the fact that *P. pastoris* cannot consume a substrate that is not present initially in the culture, so it makes sense to avoid this unrealistic metabolic behaviour through the statistical modelling.

The percentage of variance explained by including the selectivity constraint in the MCR-ALS model is 81.6%, which is only slightly lower than the percentage explained without this constraint (an admissible loss of explained variance). The variances explained by each pathway are: 11.8% (1st pathway), 9.6% (2nd pathway), 26.8% (3rd one) and 39.3% (4th one). The sum is 87.5%. Since the variance explained by the MCR model with 4 components using selectivity is 81.6%, the pathways have a degree of orthogonality of 93.2%.

The relative contributions of the pathways extracted with this model are plotted in Figure 10. The pathways obtained with this extra constraint in the model are basically the same as the ones represented in Figure 8 (results not shown).

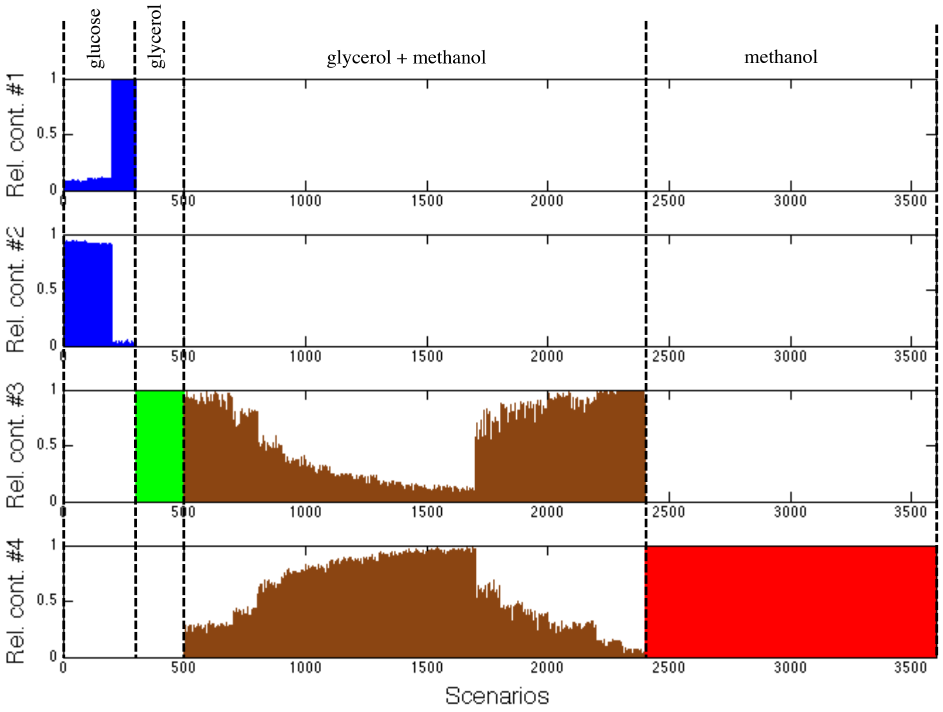


Figure 10: Relative contributions of the four pathways, including selectivity constraint. The blue columns (scenarios 1-300) represent the percentage of usage of each pathway in glucose scenarios. The green columns (301-500) represent the percentage of usage in glycerol scenarios. The brown ones (501-2400) are the scenarios with a glycerol-methanol mixture. The red columns (2401-3600) represent scenarios with only methanol as a substrate.

Nevertheless, the inclusion of the selectivity constraint on the model produces a more clear usage of each pathway. In this way, the first two pathways explain the glucose scenarios, and the third and fourth pathways explain the glycerol and methanol ones, respectively, including their mixtures.

5. Discussion

Our PCA (with the MEDA improvement) and MCR-ALS models of *P. pastoris* deserve some discussion here. The final model of MCR-ALS includes all 36 possible experimental scenarios, while in the PCA method scenario C1 (sampled scenarios 101-200) were discarded. The reason was that this scenario, in fact the hundred simulated ones, widely exceeds the 99% control limit for the Squared Prediction Error (SPE) of PCA. However, this scenario is clearly described in MCR-ALS by the first and second pathways. Moreover, the second pathway, which is describing scenario C1 up to 90% (Figure 10), consumes glucose and produces biomass. This pathway was not described by the PCA model, because biomass was only associated to glycerol consumption, while glucose consumption was only associated to TCA cycle, ATP and protein production. PCA associates a source of variability to a single principal component, so biomass cannot be explained by two orthogonal components. However, it is obviously possible for the microorganism to grow using glucose as the only carbon source, as can be seen in Figure 2 (μ values of scenarios A1 and C1). Actually, this is highly desirable as the biomass yield on this substrate is the highest. This situation illustrates the main advantage of using MCR-ALS: a source of variability can be associated to more than one pathway —in the present case, biomass growth, which appears in the second pathway (associated to glucose consumption) and the third one (associated to glycerol consumption)—. This is also related to the degree of orthogonality of the MCR-ALS pathways. They are highly orthogonal (and that is the reason why some of its pathways are similar to the PCA ones), but without imposing this constraint a new biologically meaningful metabolic route (Pathway 2) can be isolated.

The ability to include constraints during the optimisation is an advantage of MCR-ALS over PCA. Different types of biological knowledge can be included in a multicomponent model by using MCR-ALS. In the present case, non-negativity and closure are very useful in order to clearly identify and associate pathways to scenarios, while selectivity permits to avoid inconsistent behaviours related to known experimental conditions. The closure constraint allows us to explain the percentage of usage of each pathway in each scenario, but the total amount of flux flowing through a pathway cannot be compared between scenarios. This represents a disadvantage of the MCR-ALS model over a classical PCA, in which the more related is a scenario with a pathway the more flux is flowing through it.

6. Conclusions

Investigate the metabolic phenomena occurring within microorganisms is mandatory to really understand their observed behaviour. The knowledge derived from these studies is also relevant for biotechnological industries, which exploit these microbial cultures to produce top quality biochemicals. In the present work, the use of a grey modelling approach combining a first principles-based model with experimental information, followed by multivariate statistical techniques, such as MCR-ALS, provides an insight on the main metabolic relationships underlying on actual *P. pastoris* cultures.

In this way, the new approach presented here relates experimental substrates, metabolic pathways and biological functions of the yeast. Four pathways, from the bunch of possible routes, seem to be particularly relevant to represent the cellular state of a given culture. The first two pathways describe glucose consumption, but the first one is describing its use to produce a recombinant protein, while the second one addresses biomass growth. The third pathway also describes biomass growth, but using glycerol as substrate instead of glucose. Finally, the fourth pathway represents methanol consumption and the related pentose phosphate pathway.

The new methodology presented here leads to biologically more meaningful metabolic pathways than the previous approach that was using PCA-MEDA [13]. Additionally, the flexible modelling of MCR-ALS, which permits to include many sources of biological knowledge in the model, opens a new framework of collaboration between statistical and biological modellers. This framework, which can be considered as a two-step grey modelling (first step: experimental data + constraint-based model, second step: statistical models + additional biological knowledge) leads to a better understanding of these complex systems, and thus allows us to constrain the models into the desired direction and exploit all the available knowledge —first-principles, experimental data, *etc.*— in a suitable way.

Acknowledgements

Research in this study was partially supported by the Spanish Ministry of Science and Innovation and FEDER funds from the European Union through grants DPI2011-28112-C04-01 and DPI2011-28112-C04-02. The authors are also grateful to Biopolis SL for supporting this research.

Appendix. Supplementary data

The stoichiometric matrix **S** of the constraint-based modelling of *P. pastoris* can be found online.

References

[1] I.A. Isidro, A.R. Ferreira, J.J. Clemente, A.E. Cunha, J.M.L. Dias, R. Oliveira, Design of Pathway-Level Bioprocess Monitoring and Control Strategies Supported by Metabolic Networks, in: C.-F. Mandenius, N.J. Titchener-Hooker (Eds.), Meas. Monit. Model. Control Bioprocesses, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012: pp. 193–215.

[2] J.M. Otero, J. Nielsen, Industrial systems biology, Biotechnol. Bioeng. 105 (3) (2010) 439–460.

[3] M.J.T. Carrondo, P.M. Alves, N. Carinhas, J. Glassey, F. Hesse, O.-W. Merten,

et al., How can measurement, monitoring, modeling and control advance cell culture in industrial biotechnology?, Biotechnol. J. 7 (12) (2012) 1522–1529.

[4] F. Llaneras, Interval and Possibilistic Methods for Constraint-Based Metabolic Models, Ph.D. thesis, Universitat Politècnica de València, 2010.

[5] A. Kayser, J. Weber, V. Hecht, U. Rinas, Metabolic flux analysis of Escherichia coli in glucose-limited continuous culture. I. Growth-rate-dependent metabolic efficiency at steady state, Microbiology. 151 (3) (2005) 693–706.

[6] M. Tortajada, F. Llaneras, J. Pico, Validation of a constraint-based model of Pichia pastoris metabolism under data scarcity, BMC Syst. Biol. 4 (115) (2010) 1–11.

[7] T. Benyamini, O. Folger, E. Ruppin, T. Shlomi, Flux balance analysis accounting for metabolite dilution, Genome Biol. 11 (4) (2010) 1-9.

[8] M.W. Covert, E.M. Knight, J.L. Reed, M.J. Herrgard, B.O. Palsson, Integrating high-throughput and computational data elucidates bacterial networks, Nature. 429 (2004) 92–96.

[9] M.W. Covert, C.H. Schilling, B. Palsson, Regulation of gene expression in flux balance models of metabolism, J. Theor. Biol. 213 (1) (2001) 73–88.

[10] M. Åkesson, J. Förster, J. Nielsen, Integration of gene expression data into genome-scale metabolic models, Metab. Eng. 6 (4) (2004) 285–293.

[11] G.N. Stephanopulos, A.A. Aristidou, J. Nielsen, Metabolic engineering: principles and methodologies, Academic Press, San Diego, 1998.

[12] F. Llaneras, J. Picó, Stoichiometric modelling of cell metabolism, J. Biosci. Bioeng. 105 (1) (2008) 1–11.

[13] J.M. González-Martínez, A. Folch-Fortuny, F. Llaneras, M. Tortajada, J. Picó, A. Ferrer, Metabolic flux understanding of Pichia pastoris grown on heterogenous culture media, Chemom. Intell. Lab. Syst. 134 (2014) 89–99.

[14] M.D. Mesarovic, S.N. Sreenath, J.D. Keene, Search for organising principles: understanding in systems biology, Syst. Biol. IEE 1 (1) (2004) 19–27.

[15] K.J. Kauffman, P. Prakash, J.S. Edwards, Advances in flux balance analysis, Curr. Opin. Biotechnol. 14 (5) (2003) 491–496.

[16] S. Feyo De Azevedo, B. Dahm, F.R. Oliveira, Hybrid modelling of biochemical processes: A comparison with the conventional approach, Comput. Chem. Eng. 21 (1997) S751–S756.

[17] H.J. Ramaker, E.N.M. Van Sprang, S.P. Gurden, J.A. Westerhuis, A.K. Smilde, Improved monitoring of batch processes by incorporating external information, J. Process Control. 12 (4) (2002) 569–576.

[18] J.E. Jackson, A User's Guide to Principal Components, Wiley Series in Probability and Statistics, 1991.

[19] J. Jaumot, R. Gargallo, A. De Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: A new tool for multivariate curve resolution in MATLAB, Chemom. Intell. Lab. Syst. 76 (2005) 101–110.

[20] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, Crit. Rev. Anal. Chem. 36 (2006) 163–176.

[21] N. Spegazzini, I. Ruisánchez, M.S. Larrechi, MCR-ALS for sequential estimation of FTIR-ATR spectra to resolve a curing process using global phase angle convergence criterion, Anal. Chim. Acta. 642 (1-2) (2009) 155–162.

[22] T. Azzouz, R. Tauler, Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples, Talanta. 74 (5) (2008) 1201–1210.

[23] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics, Chem. Rev. 110 (8) (2010) 4582–4605.

[24] J.M. Prats-Montalbán, J.I. Jerez-Rozo, R.J. Romañach, A. Ferrer, MIA and NIR Chemical Imaging for pharmaceutical product characterization, Chemom. Intell. Lab. Syst. 117 (2012) 240–249.

[25] J. Jaumot, R. Tauler, R. Gargallo, Exploratory data analysis of DNA microarrays by multivariate curve resolution, Anal. Biochem. 358 (1) (2006) 76–89.

[26] J.M. Prats-Montalbán, R. Sanz-Requena, L. Martí-Bonmatí, A. Ferrer, Prostate functional magnetic resonance image analysis using multivariate curve resolution methods, J. Chemom. (2013) *(available online)* doi:10.1002/cem.2585.

[27] G. Potvin, A. Ahmad, Z. Zhang, Bioprocess engineering aspects of heterologous protein production in Pichia pastoris: A review, Biochem. Eng. J. 64 (2012) 91–105.

[28] J.L. Cereghino, J.M. Cregg, Heterologous protein expression in the methylotrophic yeast Pichia pastoris, FEMS Microbiol. Rev. 24 (2000) 45–66.

[29] M. Tortajada, F. Llaneras, D. Ramón, J. Picó, Estimation of recombinant protein production in Pichia pastoris based on a constraint-based model, J. Process Control. 22 (6) (2012) 1139–1151.

[30] M. Dragosits, J. Stadlmann, J. Albiol, K. Baumann, M. Maurer, B. Gasser, et al., The effect of temperature on the proteome of recombinant Pichia pastoris, J. Proteome Res. 8 (3) (2009) 1380–1392.

[31] A. Solà, P. Jouhten, H. Maaheimo, F. Sánchez-Ferrando, T. Szyperski, P. Ferrer, Metabolic flux profiling of Pichia pastoris grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates, Microbiology. 153 (1) (2007) 281–290.

[32] Solà, A., Estudi del metabolisme central del carboni de Pichia pastoris, Universitat Autònoma de Barcelona, Ph.D. thesis 2004.

[33] C. Jungo, I. Marison, U. von Stockar, Mixed feeds of glycerol and methanol can improve the performance of Pichia pastoris cultures: A quantitative study based on concentration gradients in transient continuous cultures, J. Biotechnol. 128 (4) (2007) 824–837.

[34] H.T. Ren, J.Q. Yuan, K.-H. Bellgardt, Macrokinetic model for methylotrophic Pichia pastoris based on stoichiometric balance, J. Biotechnol. 106 (1) (2003) 53–68.

[35] M.C. d' Anjou, A.J. Daugulis, A rational approach to improving productivity in recombinant Pichia pastoris fermentation, Biotechnol. Bioeng. 72 (1) (2001) 1–11.

[36] S. Curvers, J. Linnemann, T. Klauser, C. Wandrey, R. Takors, Recombinant protein production with Pichia pastoris in continuous fermentation - Kinetic analysis of growth and product formation, Chem. Eng. Technol. 25 (8) (2002) 229–235.

[37] W. Zhang, C.-P. Liu, M. Inan, M.M. Meagher, Optimization of cell density and dilution rate in Pichia pastoris continuous fermentations for production of recombinant proteins, J. Ind. Microbiol. Biotechnol. 31 (7) (2004) 330–334.

[38] B.M. Schilling, J.C. Goodrick, N.C. Wan, Scale-up of a high cell-density continuous culture with Pichia pastoris X-33 for the constitutive expression of rh-chitinase, Biotechnol. Prog. 17 (4) (2001) 629–633.

[39] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C.M. Perou, et al., Adjustment of systematic microarray data biases, Bioinformatics. 20 (1) (2004) 105–114.

[40] W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray

expression data using empirical Bayes methods, Biostatistics. 8 (1) (2007) 118–127.

[41] J. Luo, M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, et al., A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data, Pharmacogenomics J. 10 (4) (2010) 278–291.

[42] S.E. Reese, K.J. Archer, T.M. Therneau, E.J. Atkinson, C.M. Vachon, M. de Andrade, et al., A New Statistic for Identifying Batch Effects in High-Throughput Genomic Data that uses Guided Principal Components Analysis, Bioinformatics 29 (22) (2013) 2877-2883.

[43] B.Ø. Palsson, Properties of Reconstructed Networks, Cambridge University Press, New York, USA, 2006.

[44] N.D. Price, J.A. Papin, C.H. Schilling, B.O. Palsson, Genome-scale microbial in silico models: The constraints-based approach, Trends Biotechnol. 21 (4) (2003) 162–169.

[45] F. Llaneras, A. Sala, J. Picó, A possibilistic framework for constraint-based metabolic flux analysis, BMC Syst. Biol. 3 (79) (2009) 1–22.

[46] L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets Syst. 1 (1) (1978) 3–28.

[47] B. Sariyar, S. Perk, U. Akman, A. Hortaçsu, Monte Carlo sampling and principal component analysis of flux distributions yield topological and modular information on metabolic networks, J. Theor. Biol. 242 (2) (2006) 389–400.

[48] C.L. Barrett, M.J. Herrgard, B. Palsson, Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation, BMC Syst. Biol. 3 (30) (2009) 1–8.

[49] S.J. Van Dien, S. Iwatani, Y. Usuda, K. Matsui, Theoretical analysis of amino acid-producing Escherichia coli using a stoichiometric model and multivariate linear regression, J. Biosci. Bioeng. 102 (1) (2006) 34–40.

[50] F. Hadlich, K. Nöh, W. Wiechert, Determination of flux directions by thermodynamic network analysis: Computing informative metabolite pools, Math. Comput. Simul. 82 (3) (2011) 460–470.

[51] D. Machado, R.S. Costa, E.C. Ferreira, I. Rocha, B. Tidor, Exploring the gap between dynamic and constraint-based models of metabolism, Metab. Eng. 14 (2)

(2012) 112–119.

[52] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemom. 9 (1) (1995) 31–58.

[53] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146.

[54] A. De Juan, R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution, Anal. Chim. Acta. 500 (2003) 195–210.

[55] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (14) (1991) 1425–1432.

[56] N.B. Gallagher, J.M. Shaver, E.B. Martin, J. Morris, B.M. Wise, W. Windig, Curve resolution for multivariate images with applications to TOF-SIMS and Raman, Chemom. Intell. Lab. Syst. 73 (2004) 105–117.

[57] W. Windig, Two-Way Data Analysis: Detection of Purest Variables, in: Compr. Chemom., 2010: pp. 275–307.

[58] R. Leardi, Nature-inspired methods in chemometrics: genetic alogorithms and artificial neural networks, Elsevier, Amsterdam, 2003.

[59] Y. Chtioui, D. Bertrand, D. Barba, Feature selection by a genetic algorithm. Application to seed discrimination by artificial vision, J. Sci. Food Agric. 76 (1) (1998) 77–86.

[60] M. Wang, X. Zhou, R.W. King, S.T. Wong, Context based mixture model for cell phase identification in automated fluorescence microscopy, BMC Bioinformatics. 8 (32) (2007) 1-12.

[61] F. Cuesta Sánchez, J. Toft, B. Van den Bogaert, D.L. Massart, Orthogonal projection approach applied to peak purity assessment, Anal. Chem. 68 (1) (1996) 79–85.

[62] Multivariate Curve Resolution Homepage (http://www.mcrals.info/) [18 April 2014].

[63] R. Tauler. Multivariate curve resolution applied to second order data. Chemom. Intell. Lab. Syst. 30 (1995) 133-146.

[64] R. Tauler, A.K. Smilde, B.R. Kowalski. Selectivity, local rank, three-way data

analysis and ambiguity in Multivariate Curve Resolution. J. Chemom. 9 (1995) 31-58.

[65] S.A. Becker, A.M. Feist, M.L. Mo, G. Hannum, B.Ø. Palsson, M.J. Herrgard, Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox, Nat. Protoc. 2 (3) (2007) 727–738.