Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Reduction of metabolic models by polygons optimization method applied to Bioethanol production with co-substrates

C.E. Robles-Rodriguez[1, 2, 3], C. Bideaux[1, 2, 3], S. Gaucel[4], B. Laroche[5], N. Gorret[1, 2, 3], C.A. Aceves-Lara[1, 2, 3]

[1]Université de Toulouse ; UPS, INSA, INP, LISBP ; F-31077 Toulouse, France
[2]INRA, UMR792, Ingénierie des Systèmes Biologiques et des Procédés, Toulouse, France
[3]CNRS, UMR5504, Toulouse, France 135 Avenue de Rangueil, Toulouse Cedex, F-31077
(e-mail : [roblesro, bideaux, ngorret, aceves]@insa-toulouse.fr)
[4]UMR 782 Génie et Microbiologie des Procédés Alimentaires, AgroParisTech – INRA, BP 1, 1 Avenue Lucien Brétignières, 78850 Thiverval-Grignon, France
(e-mail : sebastien.gaucel@agroparistech.fr)
[5] UR341 Mathématiques et Informatique Appliquées MIA-Jouy INRA Domaine de Vilvert 78352 Jouy en Josas Cedex (beatrice.laroche@jouy.inra.fr)

Abstract: In literature metabolic stoichiometric matrix reduction is based on convex analysis by choosing the greatest triangle. This paper proposes a new methodology for the reduction of metabolic networks based on the concept of convex hull by optimization methods. Different polygons are tested to conjointly minimize the squared error (convex hull - experimental data) and maximize the convex hull area in order to reduce the set of metabolic reactions involved in the model. The advantage of this method relies on its ability to select different geometries in a simple manner with the knowledge of the elementary modes. A cybernetic model implementing the proposed optimization method is tested with data for bioethanol production by *Saccharomyces cerevisiae* growing on four substrates. Parameter estimation and model validation allow comparing the performance of the chosen polygons for reduction of metabolic pathways.

Keywords: Optimization methods, Elementary Modes, Yield Analysis, co-substrate ethanol production.

1. INTRODUCTION

Mathematical modelling of biological processes has confronted an ample variety of difficulties that have motivated the study of biological kinetics and its analysis in different manners.

Macroscopic modelling provides dynamical models which have proven enormous interest in bioengineering for the design of the on-line algorithms for bioreactor monitoring, control, and optimisation (Bastin and Dochain, 1990). These models have been extended for their use for metabolic modelling.

The kinetic approach relates metabolites concentrations with their corresponding rates. Nonetheless, they require a detailed understanding of reaction mechanisms and regulatory interaction leading to an increasing set of adjustable parameters as models grow more sophisticated (Palsson, 2000). To overcome these difficulties, stoichiometric models assume that all intracellular concentrations were at steady state, which led to algebraic equations. The inclusion of pseudo stoichiometric matrix has permitted to lump together the set of intracellular metabolic reactions of the involved microbial species (Bernard and Bastin, 2005). The necessary condition to reach steady state is that the rates of the initial and final reactions (or, equivalently, the concentrations of the initial and final metabolites) must be constant simplifying the mass balance of metabolites (Stephanopoulos *et al,* 1998).

Another approach is cybernetic modelling (Kompala *et al.,* 1984, Ramkrishna, 1982) which have the aim to include regulatory effects at the level of enzymes in a way to enrich kinetic and stoichiometric models (Young *et al.,* 2008). The complexity of this approach relies on decomposing metabolic networks into elementary pathways for which the standard cybernetic control laws can be applied.

Under the steady state condition, the metabolic network can be decomposed into a set of sub-networks called Elementary Modes (EMs) which are a set of non-decomposable pathways consisting of a minimal set of reactions that function in steady state (Shuster *et al.,* 2002). Even the computation of EMs due to external metabolites reduces the metabolic network; the complete set can still be large to be employed in metabolic modelling, so that other considerations must be applied (Song and Ramkrishna, 2009).

Useful tools based on convex mathematics have drawn to the introduction of convex hull, which main characteristic is its ability to reconstruct any data point inside the convex hull data points (Thurau *et al*., 2009). The points (modes) located in the convex hull are commonly named as generating modes.

Several efforts for reduction of elementary flux modes have been presented by Wagner and Urbanczik (2005) by using Flux Balance Analysis (FBA), which represents an important step towards making full quantitative use of stoichiometry for

maximizing biomass yield with information about the metabolic state of the organism (Geng *et al.,* 2012).

In addition to FBA, Song and Ramkrishna (2009) reported a reduction method of generating modes (GMs) based on yield analysis (YA) allowing the simplification of the convex hull and its 2-D visualization. YA implies the reduction of convex hull by (i) selecting the active modes (AMs) that form the greatest triangle and (ii) increasing the area of the convex hull by adding one vertex (AMs) at a time to achieve the 99% of the total area of the convex hull. Other works based on YA have been tested to find the 99% of the area (e.g. Ant Colony Algorithms) (Aceves-Lara *et al.,* 2011).

Following the approach of YA, the objective of this paper is the proposition of an optimization method to reduce the set of GMs by minimizing error with data and maximizing the area. The reduced set of GMs, now called AMs, will be used into the cybernetic model. The methodology is evaluated in yield space to compare different 2-D geometrical configurations (polygons) constructed by the selection of GMs, which only requires the knowledge of EMs, and the calculation of surfaces instead of regarding the distances between them.

2. MODEL REDUCTION METHOD

2.1 Cybernetic Model

Cybernetic approach is concerned with modelling regulatory processes. It views metabolic regulation as an attempt by cells to make optimal adjustments continually in response to changes in the environment by controlling both the synthesis and the activities of enzymes (Young *et al.,* 2008).

Dynamic mass balances of extracellular metabolites in a batch process can be given as:

$$\frac{dx}{dt} = S_x rc \tag{1}$$

where x is the vector of the n_x concentrations of extracellular components including biomass c. S_x is the n_x x n_r stoichiometric matrix, and r is the vector of n_r intracellular and exchange fluxes expressed per gram of biomass. Under the quasi-steady state approximation, the flux vector r can be represented by a convex combination of EMs (Schuster *et al.,* 2000).

$$r = Zr_M \tag{2}$$

where Z is the n_r x n_Z EMs matrix, and r_M is the vector of the n_Z elementary flux modes, such that (1) can be rewritten as:

$$\frac{dx}{dt} = S_x Zr_M c \tag{3}$$

The expression of r_M fluxes are based on the product of (i) a cybernetic variable $v_{M,j}$ controlling the enzyme activity, (ii) a term of relative level $e_{M,j}^{\text{rel}}$ of enzyme in relation with its maximum value, (iii) and the kinetic term $r_{M,j}^{kin}$ considered for the j^{th} elementary mode. In most of the cases, the latter term is given by the Michaelis-Menten kinetics,

$$r_{M,j}^{kin} = k_j^{\max} \prod_i \frac{x_i}{K_{j,i} + x_i} (i = 1,..., n_x; j = 1,..., n_Z) \tag{4}$$

where $k_j^{\max}$ corresponds to the reaction constant of substrates, and x_i is the concentration of external metabolites. The consumption constant of Michaelis-Menten is represented by $K_{j,i}$ for external metabolites. The kinetic term for enzymes $r_{ME,j}^{kin}$ can be computed from equation (4) by replacing the reaction and consumption constants for $k_{E,j}^{\max}$ and $K_{E,j,i}$ respectively.

The enzyme level is determined considering the following dynamic equation,

$$\frac{de_{M,j}}{dt} = \alpha_{M,j} + u_{M,j} r_{ME,j}^{kin} - (\beta_{M,j} + \mu) e_{M,j} \tag{5}$$

The parameters $\alpha_{M,j}$ and $\beta_{M,j}$ represent the constitutive synthesis and degradation rate respectively, meanwhile $u_{M,j} r_{ME,j}^{kin}$ is the inducible synthesis rate term regulated by the cybernetic variable $u_{M,j}$, and μ is the dilution rate due to the growth of enzymes.

Following the description of hybrid cybernetic modeling (Song and Ramkrishna, 2010), this work focus on seeking different options to reduce the matrix $S_xZ = Z_y$ expressed in function of external metabolites.

2.2 Reduction method

The optimization aims finding the convex hull of the normalized EMs to one compound of the metabolic pathways (e.g. biomass/substrate yield). Following the methodology described by Song and Ramkrishna (2009) the yield vector can be represented by:

$$y = Z_y h, \quad h \geq 0, \ \|h\|_1 = 1 \tag{6}$$

where Z_y is the normalized S_xZ expressed as yields, and h indicates the weight vector implying the contribution of each GM to the total area of the convex hull.

Assuming that data (experimental yield) is available, two cases are taken into account in our method.

[case 1] Data inside the convex hull. The method takes all possible combinations of GMs for the predetermined geometry which include experimental points inside the polygons. In this work we consider polygons of 3, 4, or 5 vertexes

[case 2] Data outside the convex hull. The method looks for the projection of the experimental data into the two closest GMs. Those modes are taken as reference fixed points to construct the set of combinations for the predetermined geometry reducing the number of possible combinations with respect to case 1.

Once all the possibilities of geometric configurations are known, the next optimization problem is established in order to find the best values of $Z_y h$ that,

$$\min_{Z_y h} \sum \left| Z_y h - \bar{y} \right|^T W \left| Z_y h - \bar{y} \right| \quad (7)$$
$$s.t. \max \ area$$

From (7), $Z_y h$ and $\bar{y}$ are the N_y polygon generated and experimental yield vector respectively, and W is an N_y *x* N_y weighting matrix. *area* is the area of the current polygon. For the sake of evaluating the different polygons, the maximum possible area was considered to be sufficient to represent all the set of elementary modes. A case of study is presented in the following section to evaluate whether it is possible or not to use this assumption.

If experimental data are not available, the minimization of the sum of square residuals is computed by replacing $\bar{y}$ in equation (7) for $y_{theorical}$ that represents the theorical yield generated for all the GMs ensuring surface maximization.

3. CASE OF STUDY

3.1 Bio-ethanol production from co-substrates.

In the last decades, bioethanol production from lignocellulosic biomass arouses increasing the interest to avoid the use of food crops. After the hydrolysis of lignocellulosic biomass, the remaining materials are glucose, xylose, mannose, galactose, and arabinose (Taherzadeh *et al.,* 1997). For this work, the metabolic network considered correspond to the proposed by Geng *et al* (2009) to produce ethanol from four substrates: glucose, xylose, mannose and galactose. Notice that oxygen concentration has been omitted for model simplification. Table 1 shows the 40 reactions involved in the fermentation. The network is mainly composed by glycolytic and pentose phosphate pathways, citric acid cycle, pyruvate metabolism, xylose metabolism, mannose metabolism and galactose metabolism.

- Four substrates: glucose (GLC), xylose (XYL), manosse (MAN), and galactose (GAL).
- Five products: ethanol (ETH), glycerol (GOLx), xylitol (XOLx), carbon dioxide (CO_2), and acetate (ACTx).
- Biomass (Biom) and consumption of excess ATP for maintenance (MAINT). All of them considered as external metabolites.
- Thirty one compounds are considered as internal metabolites.

Saccharomyces cerevisiae is the most traditional microorganism used for bioethanol production. Even its popularity this yeast is not able to ferment xylose (Kotter and Ciriacy, 1993). So in this work, xylose will not be considered for calculations.

The first step was to determine the Elementary modes, which were calculated with METATOOL 2005 (Kamp and Shuster, 2006). A total set of 602 EMs were obtained. Three groups were identified for the consumption of each substrate individually consisting of 33 EMs for glucose, 33 EMs for mannose, and 33 EMs for galactose, as it was mentioned by Geng *et al* (2009).

3.2 Method implementation

Each group of elementary modes is normalized for yield analysis according to their respective substrate. Experimental yield data for ethanol and biomass are taken from individual experiments for each substrate reported by Rouhollah *et al.* (2007) for the implementation of the method. Seven out of thirty tree elementary modes were identified as GMs for each substrate (Figure 1).

Minimization of square errors and maximization of area is implemented for each substrate. Figures 2 – 4 show the resulting optimized polygons and provide the value of the sum of square errors and the normalized area as a percentage of the total area of the convex hull. In order to appreciate the simplification of this method in 2-D yield space, the phase plane (adding GOLx/GLC yield) is presented in Figures 2 - 4 (d) for each geometry. Results for other substrates reported similar structures (Results not shown).

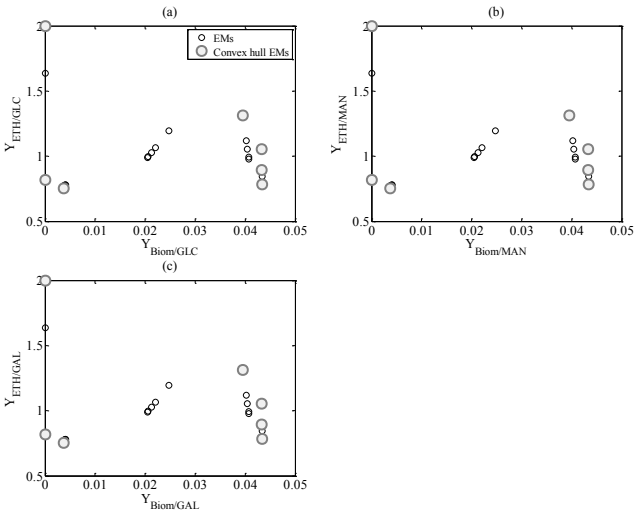

Figure 1. Elementary Modes representing the metabolic network for (a) Glucose, (b) Mannose, and (c) Galactose.

The calculation of the selected geometry is made base on [case 1] for glucose and galactose where the yield experimental point is observed inside the convex hull, meanwhile [case 2] is applied for mannose.

Regarding the number of possible polygons that can be computed for polygons, it is possible to find 35, 35, and 21 different combinations considering three, four, and five vertexes. In contrast, our optimization method calculates only (8, 5, and 5) polygons with three, (16, 10, and 10) with four, and (14, 10, and 10) with five AMs or vertexes for glucose, mannose, and galactose respectively (Results not shown).

As it is stated in Figure 2, Mannose is not optimized reflecting an area of 10.3% out of the total; however this result can be explained because the experimental yield point is placed outside the convex hull. Similar results are presented by Geng *et al.* (2012) where they only chose two elementary extreme modes to describe mannose.

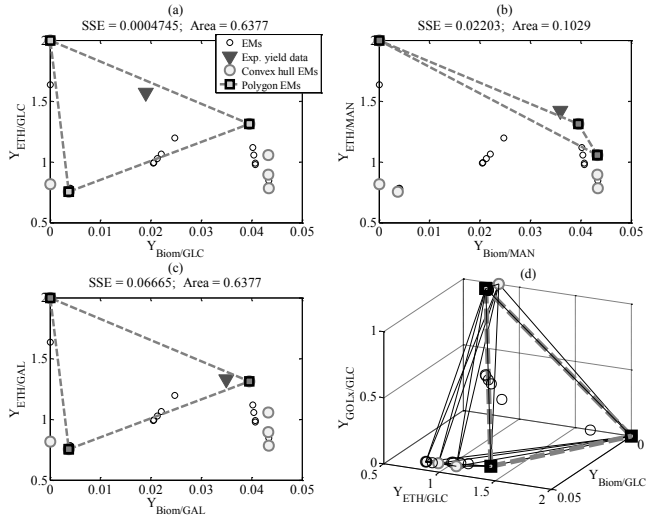

Figure 2. Optimized three AMs polygon (a) Glucose, (b) Mannose, and (c) Galactose. (d) Phase plane of yields.

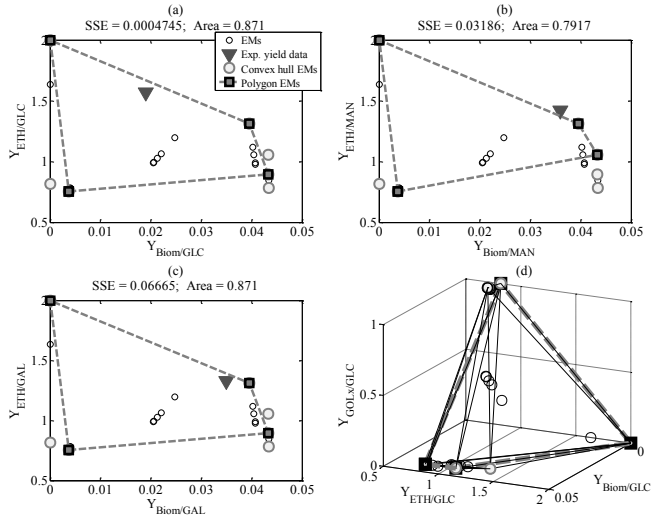

Figure 3. Optimized four AMs polygon for (a) Glucose, (b) Mannose, and (c) Galactose. (d) Phase plane of yields

Even though, glucose and galactose are represented by the 63.8% of the area which reflects the well applicability of the method here proposed.

Regarding Figures 3 and 4, areas are maximized up to more than 79% and 87% for polygons formed out of four and five points respectively. Glucose reflects the best optimization for both criteria.

3.3 Parameter estimation and validation

The reduced model consists of 9, 12, and 15 reactions instead of 602 (full model) for polygons of three, four, and five vertexes respectively. These reactions are employed to compute the $Z_y = S_x Z$ matrix of equation (3).

Table 1. Metabolic network reactions for ethanol production

v1: GLC + ATP => G6P
v2: G6P = F6P
v3: F6P + ATP = DHAP + GAP
v4: DHAP = GAP
v5: DHAP + NADH => GOL
v6: GOL => GOLx
v7: GAP => PG3 + NADH + ATP
v8: PG3 = PEP
v9: PEP = PYR + ATP
v10: PYR => ACD + CO2
v11: ACD + NADH => ETH
v12: ACD + NADHm => ETH
v13: ACD => ACT + NADPH
v14: ACT => ACTx
v15: ACT + 2 ATP => AcCoA
v16: PYR + ATP + CO2 => OAA
v17: G6P => Ru5P + CO2 + 2 NADPH
v18: Ru5P = X5P
v19: Ru5P = R5P
v20: R5P + X5P = S7P + GAP
v21: X5P + E4P = F6P + GAP
v22: S7P + GAP = F6P + E4P
v23: PYR => AcCoAm + CO2 + NADHm
v24: OAA + NADH = OAAm + NADHm
v25: OAAm + AcCoAm => ICT
v26: ICT => AKG + CO2 + NADHm
v27: ICT => AKG + CO2 + NADPHm
v28: AKG => SUC + ATP + CO2 + NADHm
v29: SUC = MAL + 0.5 NADHm
v30: MAL = OAAm + NADHm
v31: XYL + 0.5 NADH + 0.5 NADPH => XOL
v32: XOL => XOLx
v33: XOL => XUL + NADH
v34: XUL + ATP = X5P
v35: 1.04 AKG + 0.57 E4P + 0.11 GOL + 2.39 G6P + 1.07 OAA + 0.99 PEP + 0.57 PG3 + 1.15 PYR + 0.74 R5P + 2.36 AcCoA + 0.31 AcCoAm + 11.55 NADPH + 1.51 NADPHm + 30.48 ATP + 0.43 CO2 => BIOM + 2.68 NADH + 0.53 NADHm
v36: ATP => MAINT
v37: NADH =>
v38: MAN + ATP => M6P
v39: M6P => F6P
v40: GAL + ATP => G6P

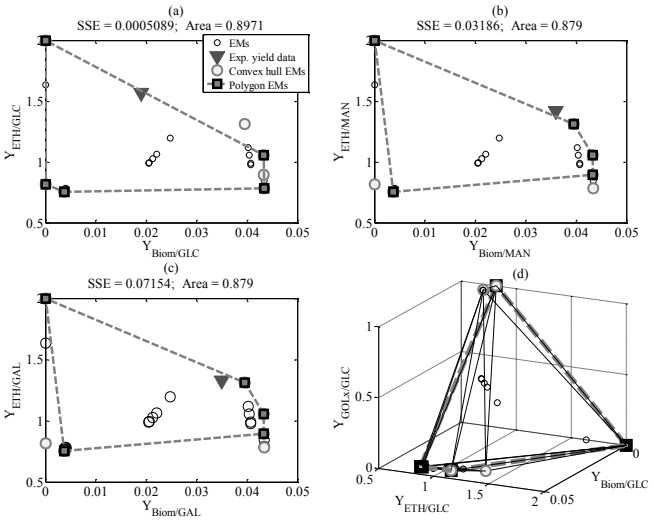

Figure 4. Optimized five AMs polygon for (a) Glucose, (b) Mannose, and (c) Galactose. (d) Phase plane of yields.

The experiments reported by Rouhollah *et al.* (2007) for co-substrates using *Saccaromyces cerevisiae* are taken as data for model validation. Rouhollah *et al.* (2007) proposed a batch experiment initiated with four substrates glucose, xylose, mannose, and galactose with initial concentrations of 30, 30, 12, 8 g/L respectively. As it was mentioned before, xylose is neither considered in the polygons nor for parameter estimation.

Equation (4) is taken from Geng *et al.* (2012) describing metabolites and enzymes as,

$$r_M^{kin} = k_j^{\max} \frac{x_i}{K_{j,i} + x_i} \frac{1}{1 + x_{ETH} / K_{I,i} + x_i} \quad (8)$$

$$r_{ME}^{kin} = k_{E,j}^{\max} \frac{x_i}{K_{j,i} + x_i} \frac{1}{1 + x_{ETH} / K_{I,i} + x_i} \quad (9)$$

$(i = 1,..., n_x ; j = 1,..., n_Z)$

The Michaelis-Menten constants $K_{j,I}$ and $K_{E,j,i}$ for equations (8) and (9) have the corresponding values employed by Geng *et al.* (2012). Notice that as all the parameters are available, thus model is fully identifiable. The optimal values of $k_j^{\max}$ and $k_{E,j}^{\max}$ (Mean values in Table 2) are estimated by the Rosenbrock method implemented in MATLAB® to fit the model to the experimental data set described above (Table 1).

Table 2. Parameters for the different polygons

AMs	$k_j^{\max}$[1/h]	$k_{E,j}^{\max}$[1/h]
3	19.344 ± 15.82	1.131 ± 0.230
4	26.08 ± 40.32	1.231 ± 0.424
5	23.19 ± 37.86	1.429 ± 0.574

Similar results were found for the estimated values of $k_{E,j}^{\max}$ for each vertex of the convex hull (AM) reflecting the conservation of these values within a low threshold of variation. The values of $k_j^{\max}$ present large standard deviations due to the coupling of different substrates, nonetheless the order of variation is analogous and intrinsically related to the chosen AMs.

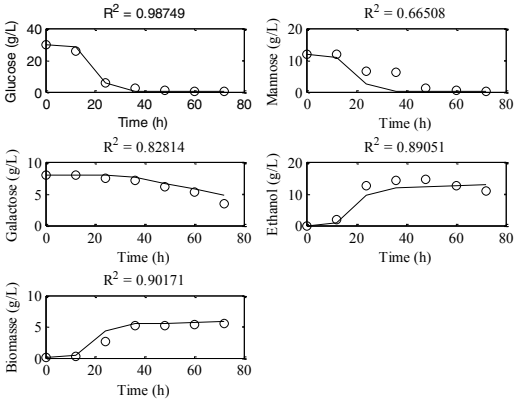

Figure 5. Performance of the polygon with three vertexes reduced metabolic model fitted to experimental data.

Validation of the model with experimental data is displayed in Figures 5 – 7, which are obtain after the estimation of 18, 24, and 30 parameters for polygons with three, four and five vertexes respectively. The coefficient of determination (R^2) is determined for each substrate and product of the metabolic pathway.

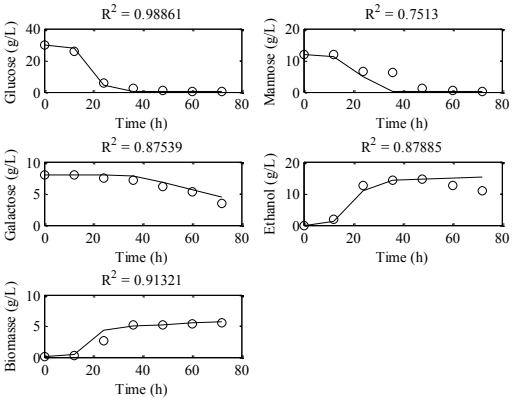

Figure 6. Performance of the polygon with four vertexes reduced metabolic model fitted to experimental data.

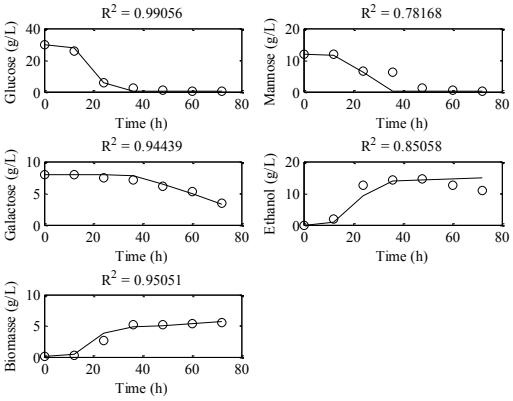

Figure 7. Performance of the polygon with five vertexes reduced metabolic model fitted to experimental data.

Concerning validity, the Root Mean Square Residuals (RMSE) is evaluated taking into account the number of the p estimated parameters as reported by Dochain and Vanrolleghem (2001),

$$RMSE = \frac{(x - \bar{x})^T W (x - \bar{x})}{(N - p)} \quad (12)$$

N represents the number of data, W the weighting matrix, x and $\bar{x}$ are the data and estimated data respectively. Besides, a coefficient of determination (R^2_T) is calculated by involving all the set of experimental data. Table 3 reports the performance evaluation of each polygon reduced metabolic model where the polygon with 5 vertexes (larger area) demonstrate the better R^2_T, but its well-fitting capacity is questioned by the great number of parameters estimated as it is reflected in the RMSE. However, the different polygon

configurations exhibit a similar behaviour in Figures 5 – 7 and analogous results in enforcement evaluation.

Notice that Mannose is the substrate with the inferior coefficient of determination due to the position of its experimental yield placed outside of the convex hull. Even though, the maximization of surface improved the fitting capacity of its parameters in the cybernetic model.

Table 3. Enforcement evaluation of the different polygon based models

Reduced Models by Polygons	RMSE	R^2_T
3 vertexes	0.1970	0.9406
4 vertexes	0.2235	0.9489
5 vertexes	0.3098	0.9508

Even the enforcement evaluation makes difficult to establish the best polygon configuration for model reduction, it is observed that all of them can represent properly the experimental data if parameter estimation is well executed.

4. CONCLUSIONS

Regarding the optimization of bioprocesses, robust models are needed to guaranty their proper use for several applications (e.g. process control). Great number of modelling approaches used metabolic networks information which involves many Elementary Modes that should be reduced. Several methods are proposed in literature, considering complicated calculations and the knowledge of experimental data and the metabolic network. The optimization method presented in this work is a useful tool to simply identify the elementary modes through polygons representing a part of the convex hull of Elementary Modes. This method can be implemented assuming that data is available or not by switching a simple parameter. Even though, it always requires the knowledge of the metabolic network.

The development of the method used in this work has been implemented on data published in the literature. In this application, the simplest polygon is the best choice. Indeed, it reproduces available data and provides the reduced model with a minimum number of parameters. Further studies will contemplate searching for a trade-off between complexity of the reduced model (analysis of the reactions of the metabolic network chosen by the polygons) and quality of the fitness considering or not the availability of experimental yield data.

REFERENCES

Aceves-Lara, C. A., Bideaux, C., Molina-Jouve, C., and Roux, G. (2011). Determination of stoichiometric matrix for ethanol production from xylose by reduction of elementary modes with ant colony systems. *IFAC 2011*. Milano, Italy.

Bastin, G., and Dochain D. (1990). On-line estimation and adaptive control of bioreactors. *Elsevier*. Amsterdam.

Bernard, O.; Bastin, G. (2005). Identification of reaction networks for bioprocesses: determination of a partially unknown pseudo-stoichiometric matrix. *Bioporcess Biosyst Eng.* **27**. 293-301.

Dochain, D., and Vanrolleghem, P. (2001) Dynamical modelling and estimation in wastewater treatment processes. *IWA Publishing*. Padstow, Cornwall, UK

Geng, J., Song, HS., Yuan, J., and Ramkrishna, D. (2012). On enhancing productivity of bioethanol with multiple species. *Biotechnol Bioeng.* **109**. 1508-1517.

Kompala, DS., Ramkrishna, D., Jansen, NB., and Tsao, GT. (1986). Investigation of bacterial growth on mixed substrates: Experimental evaluation of cybernetic models. *Biotechnol Bioeng.* **28**. 1044-1055.

Kotter, P., and Ciriacy, M. (1993). Fermentation by *Saccharomyces cerevisiae. Applied Microbiology and Biotechnology*. **38**. 776-783.

Palsson, B. (2000). The challenges of in silico biology. *Nat Biotechnol.***18**. 1147-1150.

Provost, A., and Bastin, G. (2004). Dynamic metabolic modelling under the balanced growth condition. *Process Control*. **14(7)**. 717-728.

Ramkrishna, D. (1982). A cybernetic perspective of microbial growth. In. Foundations of biochemical engineering: Kinetics and thermodynamics in biological systems. (Papoutsakis E, Stephanopoulus GN, Balnch HW, (Ed)). 161-178. *American Chemical Society.* Washington, U.S.A.

Shuster, S., Hilgetag, C., Woods, J.H., and Fell, D.A. (2002). Reaction routes in biochemical reaction systems: Algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J Math Biol.* **45**. 153 – 181.

Song, HS., and Ramkrishna, D. (2009). Reduction of a set of elementary modes using yield analysis. *Biotechnol Bioeng.* **102**. 554-568.

Song, HS., and Ramkrishna, D. (2010). Prediction of a metabolic function from limited data: Lumped hybrid cybernetic modelling (L-HCM). *Biotechnol Bioeng.* **106**. 271-284.

Stephanopoulos, G., Nielsen, J., and Aristidou, A. (1998). Metabolic Engineering: Principles and Methodologies. *Academic Press, San Diego*.

Taherzadeh, M.J., Eklund, R., and Gustafsson, L. (1997). Characterization and fermentation of dilute-acid hydrolyzates from wook. *Ind. Eng. Chem.* **36**. 4659-4665.

Thurau, C., Kersting, K., and Bauckhage, C. (2009). Convex non-negative matrix factorization in the wild. *IEEE International Conference on data mining ICDM.* 523-532.

Wagner, C., and Urbanczik, R. (2005). The geometry of the flux cone of a metabolic network. *Biophys*. **89(6)**. 3837-3845.

Young, J.D., Henne, K.L., Morgan, J.A., Konopka, A.E., and Ramkrishna, D. (2008). Integrating cybernetic modelling with pathway analysis provides a dynamic, systems-level description of metabolic control. *Biotechnol Bioeng.* **100**. 542 – 559.