Inference of dynamic macroscopic models of cell metabolism based on elementary flux modes analysis

Thomas Abbate\textsuperscript{a}, Sofia Fernandes de Sousa\textsuperscript{a}, Laurent Dewasme\textsuperscript{a}, Georges Bastin\textsuperscript{b}, Alain Vande Wouwer\textsuperscript{a}

\textsuperscript{a}Automatic Control Laboratory, University of Mons, 31 Boulevard Dolez, 7000 Mons, Belgium
\textsuperscript{b}Université catholique de Louvain, ICTEAM, Department of Mathematical Engineering, av. G. Lemaître 4, B1348 Louvain-La-Neuve, Belgium

Abstract
The objective of this study is to derive dynamic models of cell cultures based upon the concept of elementary flux modes to translate the metabolic network into macroscopic bioreactions linking extracellular substrates to products. The main steps in the development of such models are (1) the determination of the minimum number of macroreactions (2) the selection of adequate elementary flux modes, among the large number of candidates, to represent these macroreactions and (3) the identification of kinetic laws to fully explain the observed data. In this study, these questions are addressed based on both data- and knowledge-driven approaches. The methodology is illustrated with an experimental case-study, including four different hybridoma cell perfusion cultures. The resulting model reproduces well the experimental data, and could be used for process monitoring or control.

Keywords: Mathematical modeling, parameter estimation, model reduction, biotechnology, mammalian cells

1. Introduction
Mammalian cell cultures are a representative source of a number of biopharmaceutical products, including monoclonal antibodies [1, 2], viral vaccines [3], and hormones [4]. Predicting the behavior of mammalian cells during cell culture processes under different culture conditions is highly desirable for both industrial and scientific reasons [5]. The efficiency of these processes can be further increased by model-based optimization and control strategies. Therefore, a reliable kinetic model for the cell metabolism is required to identify the parameters which have a significant impact on cell viability and on protein production and to understand their effects on the cellular phenotype.

Kinetic models can be categorized as structured or unstructured models. Structured models correspond to a relatively detailed representation of the cells, including the interaction between intracellular components [6]. They attempt to explicitly describe the physiological state of the cells, its composition or its regulatory adaptation to the environmental changes, resulting in a high number of equations which are difficult to handle. Instead, unstructured models do not consider the intracellular activity of the cells. They are regarded as a black-box, where substrates are consumed and converted into products. These models are known to describe the evolution of extracellular culture variables, such as biomass, substrates and products by a small set of macroscopic reactions connecting the initial components to the final products. They are simpler than structured models, thus making the model easier to identify, to use for optimization, or design of on-line algorithms for process monitoring and control.

Despite the differences between structured and unstructured models, recent studies have highlighted the strong link between both modeling approaches. Several methods have been devised in order to deduce macroscopic reaction schemes from metabolic networks of intracellular reaction pathways thanks to model reduction procedures [7, 8, 9, 10, 11, 12, 13, 14].

In this study, a simple procedure to infer macroscopic bioreaction models from metabolic networks is proposed. As
in previous reports, the concept of elementary flux modes (EFMs) is used to translate the metabolic network into macroscopic bioreactions linking extracellular substrates to products. However, even considering a small metabolic network, the number of EFMs can be quite large, and a selection of specific EFMs as candidate macro-reactions for the model is needed [13, 15, 14].

In this connection, several approaches have been investigated, such as the one of Hebing et al., 2016 [9], where a method based upon dynamic metabolic flux analysis (DMFA) is used to find an optimal selection of EFMs based on concentration measurements instead of reaction rates. In the study of Naderi et al., 2010 [11], the metabolic network is simplified by eliminating insignificant fluxes. Moreover, Hybrid Cybernetic Models (HCM) [16] have been used to reduce the number of EFMs by projection of the modes into the yield space. Another approach is the class of Lumped Hybrid Cybernetic Models (L-HCM) [17], which consist in grouping EFMs into clusters and creating an average EFM for each cluster. Finally, Flux Balance Analysis (FBA) and its extension Dynamic Flux Balance Analysis (DFBA) can also be seen as a method to reduce the number of EFMs using optimization [7]. Indeed, a solution of FBA corresponds to a positive linear combination of EFMs and the solution for any optimal product/substrate ratio always coincide with an elementary mode [18].

In this study, a linear optimization problem is formulated to select the best candidate EFMs that describe the macroscopic bioreactions. This strategy proceeds in three steps:

• first, a data-driven approach is exploited, where principal component analysis, in the spirit of [19], is used to infer the minimum number of bioreactions that are required to explain the observed experimental data;
• next, the knowledge of the metabolic network is exploited to select a corresponding number of EFMs that are a priori the best candidates to describe the bioreaction network;
• finally, simple kinetic laws, e.g., Monod and inhibition factors, are formulated and nonlinear parameter estimation is used to fit the model to the experimental data. At that stage, the range of operation of the bioreactor (range of concentrations of the main components) and the dynamical changes within this range are determinant in the possibility to estimate kinetic parameters and to possibly model additional phenomena, such as inhibition.

This paper is organized as follows. The next section presents the methodology for determining the minimum number of macoreactions and the most representative EFMs. Information on the experimental databank, i.e. data collected from cultures of hybridoma cell line HB58 in 2L bioreactor operated in perfusion mode is provided in section 3. The considered metabolic reaction network is introduced in section 4. Then, the results of the experimental application are discussed and a dynamic macroscopic model is developed and validated in section 5. Conclusions are finally drawn in section 6.

2. Determining minimal sets of macroscopic bioreactions

Based on the method originally developed in [19], this section proposes a methodology to determine a minimal set of macrobioreactions which is consistent with the available experimental data.

The general mass-balance model equations are given by:

\[
\frac{dC_m}{dt} = K\varphi + D(C_{m_{in}} - C_m) - Q
\]

(1)

where \(C_m\) is the vector of concentrations of the \(n_m\) measured species in the culture medium. The full rank (equal to the number of reactions \(n_r\)) matrix \(K\) represents the stoichiometry of the reaction network. \(\varphi\) is the vector of reaction rates and \(D\) is the dilution rate. The term \(C_{m_{in}}\) represents the inflow concentrations. The gaseous flow rate \(Q\) is the exchange of matter in gaseous form between the surrounding and the reaction medium. At each measurement time \(t_j\), the vector of (measurable) biological rates of consumption/production of each species can thus be defined as:

\[
u(t_j) = \frac{dC_m(t_j)}{dt} - D(t_j)(C_{m_{in}}(t_j) - C_m(t_j)) + Q(t_j)
\]

(2)
The computation of the vector \( u(t_j) = (u_1(t_j), \ldots, u_n(t_j)) \) is based on the combination of smoothing B-splines (function `spaps` in Matlab) applied to the noisy concentration measurements and the differentiation of these splines (function `fnder` in Matlab). Equation (1) therefore leads to:

\[
u(t_j) = K \phi(t_j).
\]  

(3)

This equation can be written in matrix form as follows:

\[
U = KW
\]  

(4)

where matrix \( U^{n \times N} \) is made of the vectors \( u(t_j) \) at \( N (> n_m) \) measurement times \( t_1, \ldots, t_N \). In the same way, the associated matrix of reaction rates \( W \) is made of \( N \phi(t_j) \) vectors. We assume that the reaction rates are linearly independent and that the experimental conditions provide data that are sufficiently informative to have a full rank matrix \( W \) (i.e. none of the lines of \( W \) can be written as a linear combination of the others). Then, the following classical result can be used (see [19]):

**Property 1** For a matrix \( K \) of rank \( n_r \), if \( W \) has full rank, then the \( n_m \times n_m \) matrix \( M = U U^T = K WW^T K^T \) has rank \( n_r \). Since it is a symmetric matrix, it can be written:

\[
M = P \Sigma P^T
\]

where \( P \) is an orthogonal matrix \( (P^T P = I) \) and

\[
\Sigma = \begin{pmatrix}
\sigma_1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
0 & \sigma_2 & \ldots & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \sigma_{n_r} & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 0 & 0 & \ldots & 0 
\end{pmatrix}
\]

with \( \sigma_{i-1} \geq \sigma_i > 0 \) for \( i = 2, \ldots, n_r \). The number of macroscopic reactions corresponds to the number of non zero eigenvalues of \( U U^T \).

In practice, due to data sampling and measurement noise, there will be no zero eigenvalues, and the question on how to determine the number of eigenvalues that must be taken into account in order to reasonably approximate the experimental data is answered in [20] by selecting the \( n_r \) first principal axes, which represent a total variance larger than a fixed threshold. To avoid conditioning problems, the components \( \tilde{u}_k(t_j) \) of the vector \( u(t_j) \) are normalised as follows:

\[
\tilde{u}_k(t_j) = \frac{u_k(t_j)}{\|u(t_j)\|}, \quad k = 1, \ldots, n_m
\]  

(5)

for the computation of the matrix \( U U^T \). Having determined the number \( n_r \) of macro-reactions, the next step is to select a subset of \( n_r \) reactions among the \( n_e \) biological macro-reactions derived from the elementary flux modes, to represent the experimental data.

Due to measurement errors and smoothing spline approximations, it may happen, in practical applications, that the sign of small fluxes in equation (3) becomes uncertain. Therefore, in order to deal with the situation where the fluxes \( \phi(t_j) \) are known to be of constant sign along the data of the concerned experiment, we introduce also an alternative linear program to solve the linear system (3). This linear program aims at finding the smallest tolerance corridor \( \epsilon \) for the satisfaction of the system (3) with a prescribed sign of the fluxes. The linear programming problem is expressed
as follows in the case where all fluxes are assumed to be positive:

\[
\forall j, j = 1, \ldots, N;
\]

Let \( \varepsilon(t_j) = [\varepsilon_1(t_j), \ldots, \varepsilon_{\nu_m}(t_j)]^T \); 
\( \varphi(t_j) \geq 0 \)

\[
\min_{\varepsilon(t_j)} \sum_{k=1}^{\nu_m} \varepsilon_k(t_j)
\]

s.t. \[
K_e \varphi(t_j) \leq u(t_j) + \Gamma \varepsilon(t_j) \\
K_e \varphi(t_j) \geq u(t_j) - \Gamma \varepsilon(t_j)
\]

where

- \( \varepsilon(t_j) \) can be interpreted as a positive vector of minimal tolerances that allow computing an approximate solution to the system of equations \( K_e \varphi(t_j) \approx u(t_j) \) (this system is over-determined when \( n_r < n_m \)).
- \( \varphi(t_j) \geq 0 \) imposes a positivity constraint on the solutions,
- \( K_e \) is the stoichiometry matrix of the selected set of macroreactions (Table 2),
- the weighting matrix \( \Gamma \) is defined as follows:

\[
\Gamma = \begin{pmatrix}
g_1 & 0 & \ldots & 0 \\
0 & g_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & g_{\nu_m}
\end{pmatrix},
\]

with

\[
g_k = \max_{t_j} |\tilde{u}_k(t_j)|, \ k = 1, \ldots, \nu_m.
\]

An interesting indicator of the capacity of the model associated with \( K \) to accurately describe the external data is given by:

\[
R = \sum_{j=1}^{N} \sum_{k=1}^{\nu_m} \varepsilon_k(t_j)
\]

Obviously the model precision increases when the quantity (9) decreases and the model-candidates can be selected accordingly. Besides, system (6) also provides the values of the fluxes \( \varphi(t_j) \) at any instant \( t_j \) \( (j = 1, \ldots, N) \) which allows identifying kinetic models. In section 5, this methodology is applied to a small metabolic network proposed in [21] and experimental data from perfused hybridoma cell cultures.

3. Material and methods

3.1. Cell line and media

This study is illustrated with experimental data from cultures of hybridoma cell line HB58 (ATCC), which produces antibodies type IgG1, anti-CD54, specific for mouse kappa light chain. These experiments have been performed at the State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology (ECUST), Shanghai [12]. Serum-free medium chemically defined with 1:1 mixture of DMEM and F12 (Gibco) was used and supplemented with 10 mg of bovine insulin, 10 mg of transferring-selenite (Fe-saturated), 500 \( \mu \)mol of ethanolamine and other property additives. Glucose, glutamine and other amino acids were added to the cultures. Table 1 describes
the operating conditions of the cultures and the glucose, glutamine and alanine concentrations in the feed stream. It must be stressed that the experiments are designed to cover a wide operating range and capture information about glucose and glutamine saturation and limitation levels.

3.2. Bioreactor operation mode and analysis methods

Cultures were conducted in a 2-L stirred bioreactor (B. Braun Biostat B-DCU) and were settled in a working volume of 1.8 liters during the whole cultures. All cultures started in batch mode and were inoculated to reach an initial concentration between $0.2 - 0.5 \times 10^9 \text{cells/L}$. Temperature was kept at $36.8^\circ C$; the gases air, $CO_2$, $O_2$ and $N_2$ were mixed to maintain DO at 40% air saturation and bicarbonate solution (0.75 mol/L $Na_2CO_3$ and 0.5 mol/L $NaHCO_3$) was used for pH control around 7.0 ± 0.2. Data acquisition and process control were performed using the supervisory software MFCS/Win 3.0.

Perfusion phase started after 36-48 hours of batch culture with a constant dilution rate of 0.0197h⁻¹. Cells were retained by a spin-filter (20 µm) and the stirring speed was kept at 200 rpm.

Cells were counted with hemocytometer using the trypan blue exclusion method. Glucose, lactate and ammonia concentrations were determined using YSI 7100 biochemical analyzer (Yellow Springs Instruments). The amino acids were analyzed by reverse-phase High Performance Liquid Chromatography (HPLC) with a UV-visible detector.

4. Metabolic network presentation

A small metabolic network is used to represent the central metabolism of hybridoma cells (see figure 1). This metabolic network describes only a part of the metabolism concerned with the utilization of the main energetic nutrients (glucose and glutamine), the secretion of the main extracellular products (lactate, ammonia, alanine and $CO_2$) and the production of the intracellular nucleotides.

This network is quite small in comparison with more detailed networks (e.g. [22]) but this case study is used to develop and illustrate the concepts in a clear, traceable way. This small-size example is therefore more to be seen as a "proof-of-concept".

From this metabolic network, eleven elementary flux modes are obtained (see [21] for details). They induce the ten macroscopic input/output bioreactions given in Table 2 (Note that EFMs $e_9$ and $e_{10}$ give the same input/output reaction). In the sequel of this study, the macroreactions of Table 2 will be labelled with the notations $e_i$ of the corresponding EFMs. The stoichiometry matrix corresponding to these ten macroreactions is given in table 3.

5. Application to hybridoma cell cultures in perfusion

5.1. Selection of the candidate macroreactions

Our case study includes the experimental data of four different perfused cell cultures, from which on-line measurements of glucose, glutamine, lactate, ammonia and alanine are available. The vectors are computed by fitting smoothing B-splines to the data and computing their derivatives:

$$u(t_j) = \begin{bmatrix} \frac{dGlc}{dt} & -D(t_j)(Glc_{in}(t_j) - Glc(t_j)) \\ \frac{dLac}{dt} & +D(t_j)Lac(t_j) \\ \frac{dGln}{dt} & -D(t_j)(Gln_{in}(t_j) - Gln(t_j)) \\ \frac{dN}{dt} & +D(t_j)N(t_j) \\ \frac{dAla}{dt} & -D(t_j)(Ala_{in}(t_j) - Ala(t_j)) \end{bmatrix}$$  (10)
Then, using the normalization (5), the eigenvalues of the matrix $UU^T$ are computed and the principal component analysis (PCA) indicates, as shown in Figure 2, that three reactions are sufficient to explain 95% of the observed variance, four reactions explain 99% of the total variance and five reactions obviously allow explaining 100%. In this application, we know by assumption that, whatever the selected macroreactions in Table 2, the fluxes $\varphi(t_j)$ must be positive. Interestingly, as shown in Table 4, the solution of the linear program (6) provides results that are not completely in line with the preliminary PCA. This table gathers the residual cost $R$, obtained from (6) and (9), of the best candidate-combinations of two to ten macroreactions covering the whole network, i.e., which allow describing the evolution of all the network species. It shows that the minimal residual cost is first obtained with a combination of 7 elementary fluxes (and not of 5 EFMs as expected). Moreover, if a combination was able to describe 100% of the variance, the residual cost of that combination would be equal to zero. This is not the case even when including all elementary fluxes of the studied network.

This counter-intuitive result is due to the positivity constraint imposed on the solutions $\varphi(t_j)$ of system (6) which transform a problem of linear algebra into a convex analysis. When removing the positivity constraint, the results are in agreement with the PCA outcome and the total data variance is then explained by all combinations of 5 elementary fluxes leading to $K_e$ matrices of rank 5 (not shown in the paper). This interesting observation induces that the several candidate-combinations corresponding to a specific number of macro-reactions usually do not lead to the same level of accuracy in representing the data. System (6) therefore constitutes a useful tool to discriminate among candidates, even for combinations of more than 5 reactions. This is confirmed by figure 3 which plots the time evolution of the concentrations corresponding to all the combinations of 5 EFMs obtained from the computation of (6). It clearly shows that some EFM combinations are better suited to describe the available data. In fact, 4 combinations seem particularly well suited to reproduce the data, which are presented in table 5.

The corresponding kinetics, inferred from the computation of (6), have almost identical shapes. As shown in figure 4a, which presents the kinetics of combination 2 (table 5), these complex shapes might be difficult to reproduce using simple kinetic laws with Monod or inhibition factors. In order to check the consistency of these kinetic signals, they are smoothed using B-splines, integrated and compared to the measurements. It is interesting to observe that the concentration evolutions obtained by integration of the smoothed kinetic signals for the four model-candidates are similar and satisfactorily fit the experimental data (figure 4b), despite the loss of information associated with the smoothing approximation. However, the reproduction of alanine in dataset 3 is less satisfactory, whatever the number of macroreactions considered. We can only conjecture about this observation: (a) the level of alanine is dataset 3 is lower than in the other experiments and might be close to the detectability levels of the HPLC technique, (b) alanine is provided in the feed stream and quantification errors as well as small perturbations can have a direct impact on the concentrations in the culture medium, (c) alanine consumption could occur in some phases of the culture, but as all fluxes are assumed irreversible, this phenomena is not modelled here (we have not pursued this route, which seems unlikely, because of the limited information in dataset 3 and the lack of further evidence).

Any of the combinations of table 4 might thus be considered for model identification. As the results of PCA indicate that 3 reactions explain 95% of the variance, it is proposed to also investigate models with 4 and 3 reactions. Combination 1-3-5-9-11 (i.e. combination 2 of table 5) seems to be a good base to achieve this objective since the residual cost of combinations 1-3-5-11 and 1-5-11 are the second and first best combinations of 4 and 3 elementary fluxes, respectively.

The mass balance ordinary differential equations, for which kinetic evolution have now to be postulated, are given
by:

\[
d\text{Glc}/dt = -\varphi_1 X + 2 \varphi_9 X - 2 \varphi_{11} + D(\text{Glc}_{\text{in}} - \text{Glc}) \tag{11}
\]

\[
d\text{Lac}/dt = 2 \varphi_1 X - \varphi_9 X - DLac \tag{12}
\]

\[
d\text{Gln}/dt = -\varphi_3 X + 3 \varphi_{11} + D(\text{Gln}_{\text{in}} - \text{Gln}) \tag{13}
\]

\[
dN/dt = 2 \varphi_3 X + 2 \varphi_9 X + \varphi_{11} X - DN \tag{14}
\]

\[
d\text{Ala}/dt = \varphi_3 X + D(\text{Ala}_{\text{in}} - \text{Ala}) \tag{15}
\]

where Glc_{in}, Gln_{in} and Ala_{in} are the concentrations of glucose, glutamine and alanine in the feed stream and \(D\) is the dilution rate. The parameter identification procedure is described in the following section.

### 5.2. Nonlinear parameter estimation

The model identification performed in this work is achieved using classical nonlinear parameter estimation techniques [23]. Since the stoichiometric coefficients of the macroreactions are provided by the elementary fluxes modes, only the kinetic parameters are estimated by minimizing a weighted nonlinear least-square criterion (sum of squared differences between model predictions and experimental measurements) of the form:

\[
\min_{\theta} J = \sum_{k=1}^{N} (y_k - y_{m,k}(\hat{\theta}))\Sigma^{-1}(y_k - y_{m,k}(\hat{\theta}))^T
\]

where \(\hat{\theta}\) is the set of parameters to identify, \(J\) is the value of the cost function, \(y_k\) are the measurements, \(y_{m,k}\) are the model predictions, \(\Sigma^{-1}\) is a normalization matrix in which diagonal elements are chosen as the squares of the maximum measurement values of each component concentration, and \(N\) is the number of sampling times. Biomass is considered as a known variable, whose values are obtained by smoothing B-spline.

Once an estimation \(\hat{\theta}^*\) of the hypothetical true parameters \(\theta^*\) is found, an a posteriori estimation of the measurement error variance \(\hat{\sigma}^2\) is obtained by:

\[
\hat{\sigma}^2 = J/N - P
\]

where \(P\) is the number of parameters (dimension of the vector \(\hat{\theta}^*\)). This relation is then used to build the a posteriori covariance matrix of the measurement errors \(\Sigma\):

\[
\Sigma = \hat{\sigma}^2 W
\]

which can be used to evaluate the Fisher information matrix (FIM):

\[
\text{FIM} = \sum_{k=1}^{N} Y_{\theta,k}\Sigma^{-1}Y_{\theta,k}^T
\]

where \(Y_{\theta,k}\) is the sensitivity matrix of the model with respect to the parameters at time \(k\). These sensitivities are computed through integration of the parametric sensitivity equations:

\[
\frac{dY_{\theta}}{dt} = \frac{\partial f}{\partial Y}Y_{\theta} + \frac{\partial f}{\partial \theta}
\]

A lower bound of the parameter estimation error covariance matrix is then given by:

\[
Q = \text{FIM}^{-1}
\]
It must be stressed that due to possible measurement errors affecting the initial concentrations of the experiments, the latter are introduced in the identification problem as unknown parameters. This allows alleviating the bias that could introduce wrong initial concentrations.

6. Results and discussion

In order to start the identification with a model with some flexibility, all kinetics are first set to the same general structure:

\[
\varphi_i = \mu_{\text{max},i} \frac{\text{Glc}}{\text{Glc} + K_{\text{Glc},i}} \frac{\text{Gln}}{\text{Gln} + K_{\text{Gln},i}} \frac{\text{K Lac}_i}{\text{Lac} + K_{\text{Lac},i}} \frac{N}{N + K_N} \quad i = 1, \ldots, 5
\]

where Glc, Lac, Gln, N and Ala stands for the concentrations of glucose, lactate, glutamine, ammonium and alanine, respectively, KGlc and KGln are the half-saturation constants of Monod laws with respect to glucose and glutamine, and Klac and KN are inhibition constants associated with lactate and ammonium. This kinetic structure can subsequently be adapted depending on the parameter identification results as well as parameter sensitivity analysis.

Moreover, since nonlinear parameter estimation problems are subject to local minima, a multi-start strategy is developed in order to ensure the exploration of the solution space and the identification of the optimal parameters $\bar{\theta}^\star$. This strategy consists in performing repeated optimization runs starting from initial guesses $\theta_0$ randomly chosen among predefined admissible intervals (A.I.) in which the parameters are likely to be found.

Since the purpose of the present work is to test the capacity of system (6) to select sets of macroreactions able to reproduce the available data, the identification procedure is first achieved using all datasets at hand for identification and direct validation.

The application of the first round of optimization runs yields a first model (model 1) where the kinetics have the following structure:

\[
\begin{align*}
\varphi_1 &= \mu_{\text{max},1} \frac{\text{Glc}}{\text{Glc} + K_{\text{Glc},1}} \frac{\text{Gln}}{\text{Gln} + K_{\text{Gln},1}} \quad \text{Glc} \\
\varphi_3 &= \mu_{\text{max},3} \frac{\text{Glc}}{\text{Glc} + K_{\text{Glc},3}} \frac{\text{Gln}}{\text{Gln} + K_{\text{Gln},3}} \frac{\text{K Lac}_3}{\text{Lac} + K_{\text{Lac},3}} \frac{N}{N + K_N} \\
\varphi_5 &= \mu_{\text{max},5} \frac{\text{Gln}}{\text{Gln} + K_{\text{Gln},5}} \\
\varphi_9 &= \mu_{\text{max},9} \frac{\text{Glc}}{\text{Glc} + 0.01} \\
\varphi_{11} &= \mu_{\text{max},11} \frac{\text{Glc}}{\text{Glc} + 0.01} \frac{\text{Gln}}{\text{Gln} + K_{\text{Gln},11}} \frac{\text{K Lac}_{11}}{\text{Lac} + K_{\text{Lac},11}} \frac{N}{N + K_N}
\end{align*}
\]

where arbitrary small half-saturation coefficients have been introduced in the Monod factors for glucose activation in $\varphi_9$ and $\varphi_{11}$ to ensure the positivity of the model prediction. Indeed, these parameters cannot be estimated precisely from the available data, and have to be fixed in order to resolve identifiability issues. It is notorious that half-saturation coefficients can be delicate to estimate and require enough data in specific concentration ranges (actually data points where the glucose concentration values are relatively close to the half-saturation coefficients). On the other hand, the positivity of the predicted concentrations is an essential feature of any candidate model, which explains why these coefficients cannot be simply eliminated from the model. It is also interesting to note that Monod factors for glutamine and glucose appear in $\varphi_1$ and $\varphi_3$ respectively, while these components are not consumed in the corresponding reactions. Glucose and glutamine are indeed well-known to be interconnected in the cellular metabolism through the TCA cycle and the related anaplerotic reactions [24, 25, 26, 27], which could explain the latter observation.

Identification results are provided in table 6 which gather the estimation results of the multistart strategy with a residual cost function $J$ below specified cut-off values. The minimum and maximum parameter values are listed alongside with the best estimate $\hat{\theta}^\star$ (leading to the lowest residual cost). Coefficients of variation (standard deviations expressed in percentage of the parameter values) are also evaluated based on the Fisher Information Matrix. The normalized residual mean squared error (RMSE), scaled with the square of the maximum measured values, is also
provided. This information is given per dataset and per variable and expressed in % of J (obviously, the sum of these relative errors is 100%).

From 200 optimization runs, 33 present a residual cost lower than 10 and are considered as leading to interesting local minima (potentially in the neighbourhood of the global minima), while others are rejected. The precision of the best estimate $\hat{\theta}^*$ is satisfactory since the coefficients of variations are mostly below 50%. The model predictions associated with $\hat{\theta}^*$ are given in figure 5a where the error bars represent the a posteriori 95% confidence intervals on the measurements. Figure 5b shows the comparison between the fluxes obtained after identification of the kinetic structures and the smoothing of the fluxes obtained by applying LP (6) with the combination 1-3-5-9-11. Clearly, the evolution of the concentrations obtained with model 1 are in good agreement with the data sets. Although the fluxes predicted by the identified model do not exactly match the results of the smoothing splines, they present similar shapes and share the same order of magnitude. This supports the initial choice of simple kinetic structures. It is also interesting to note that 22% of the residual cost J is due to alanine measurements in dataset 3. Yet, as mentioned in section 5.1, the alanine evolution cannot be reproducible given the positivity constraints imposed on the fluxes and the possible errors in the measured feed concentrations.

To further test the capacity of the proposed method to select appropriate model candidates, models 2 and 3, which are described by combinations 1-3-5-11 and 1-5-11, respectively, are now identified. Even though the residual costs increase with the model reduction, the measurement prediction remain acceptable (figure 5a). The apparent precision of the parameter estimates also increases as a side effect of the reduction of the number of parameters to estimate from a given amount of data information. These results are remarkable, as identification only targets reaction kinetics, while the stoichiometry is imposed by the selection of EFMs.

The next step consists in challenging the capacity of the model to predict unseen experiments, and to this end, extra identification runs are conducted using only a subset of the available data sets, while keeping the rest for cross-validation.

As mentioned in section 3.1, the experiments were designed to highlight the limitation and saturation associated with glucose and glutamine levels. Experiments 2 and 3 point out limitations with respect to glucose or glutamine, while experiments 1 is expected to show both. In contrast, experiment 4 aims at avoiding these limitations. In consequence, extra identification runs are performed using the 3 datasets where limitations are foreseen (data sets 1, 2 and 3) whereas data set 4 is kept for cross-validation. Models 4, 5 and 6 are basically the same as model 1, 2 and 3, respectively, but they are identified based on only 3 data sets. The results are shown in figure 6 which confirms that satisfactory results can be obtained using 5 reactions (model 4) both in direct and cross-validation (although the alanine behaviour is not very well reproduced for reasons already mentioned). Other combination of 3 datasets have been tested but none of them leads satisfactory cross-validation results (not shown in this work), which stresses the importance of the information content of the data sets to achieve model selection and identification. As shown in table 6, the parameter estimates of model 4 are close to the ones of model 1, which supports the robustness of the approach.

7. Conclusions

In this study, a simple procedure to deduce a macroscopic model of cell cultures is developed based on the concept of elementary flux modes. This procedure relies on (1) a data-driven approach where a principal component analysis is used to determine the minimum number of bioreactions required to explain the observed experimental data, (2) the selection of appropriate elementary flux modes based on the solution of a linear programming problem and (3) the identification of the model kinetic parameters using a nonlinear weighted least square criterion.

The positivity constraints imposed on the solutions generally induce a larger number of reactions to explain the data than the minimum number suggested by PCA, and the models based on the selected combinations of elementary flux modes provide a satisfactory prediction of the concentration evolution, even when using simple kinetic structures.

The apparent main limitation of the present study is the use of a very simple metabolic network resulting in only eleven elementary flux modes. The extension of the method to a larger network would imply a significant increase in the computational load to explore the possible combinations among a much larger set of EFMs. However, as the main purpose of the approach is to deduce low-order dynamic models of cell cultures, that could be used for optimization or control, one could question the interest of initiating the procedure with large, detailed, networks.

It must also be noticed that the proposed macroscopic models do not include biomass production reactions and that the cell concentrations are directly deduced from smoothing-splines applied to the experimental measurements. If
biomass production must be included in the model, it can be achieved by imposing the cell growth rate as an external flux in the linear programming problem.

References


List of Figures

1 Metabolic network for mammalian cell central metabolism. ............................... 12
2 PCA results giving the total variance explained with respect to the number of reactions. ......................... 13
3 Time evolution of the concentrations obtained by integration of the solutions of LP (6) for all combinations of 5 EFMs covering the whole network. ......................................................... 14
4 Smoothing splines of the fluxes $\Phi$ for combination 2 (a) and time evolution of the concentration obtained by integration of the smoothed $\Phi$ signals related to the 8 best combinations of 5 EFMs (b). .............................. 15
5 Direct validation of models 1, 2 and 3 (a) and comparison of the predicted fluxes with the smoothing results of figure 4a (b). ............................................................... 16
6 Direct validation of models 4, 5 and 6 ........................................................................ 17
Figure 1: Metabolic network for mammalian cell central metabolism.
Figure 2: PCA results giving the total variance explained with respect to the number of reactions.
Figure 3: Time evolution of the concentrations obtained by integration of the solutions of LP (6) for all combinations of 5 EFMs covering the whole network.
Figure 4: Smoothing splines of the fluxes $\phi$ for combination 2 (a) and time evolution of the concentration obtained by integration of the smoothed $\phi$ signals related to the 8 best combinations of 5 EFMs (b).
Figure 5: Direct validation of models 1, 2 and 3 (a) and comparison of the predicted fluxes with the smoothing results of figure 4a (b).
Figure 6: Direct validation of models 4, 5 and 6
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Operating conditions for both cultures.</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>Macroscopic Reactions</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>Stoichiometry of the measured species for the macroreactions of Table 2</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>Success combination of macroreactions passing test (6), where $R(9)$ represents the sum of minimal errors necessary for LP (6) to allow feasible solutions. Note that $R$ has no unit.</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>Best combinations of 5 elementary fluxes selected as candidates to build a macroscopic model</td>
<td>23</td>
</tr>
<tr>
<td>6</td>
<td>Parameters values identified with three perfusion cell cultures (dataset 2, 3 and 4) for Model 1, Model 2 and Model 3. VC corresponds to the variation coefficient expressed in %.</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>Parameter estimation results and related confidence intervals.</td>
<td>24</td>
</tr>
<tr>
<td>Initial viable cells</td>
<td>Feed stream</td>
<td>Duration of Batch</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------</td>
<td>------------------</td>
</tr>
<tr>
<td>units [10⁹ cells/L]</td>
<td>Glc in [mM]</td>
<td>Gln in [mM]</td>
</tr>
<tr>
<td>Cult.1</td>
<td>0.19</td>
<td>11</td>
</tr>
<tr>
<td>Cult.2</td>
<td>0.23</td>
<td>15</td>
</tr>
<tr>
<td>Cult.3</td>
<td>0.36</td>
<td>28</td>
</tr>
<tr>
<td>Cult.4</td>
<td>0.36</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 1: Operating conditions for both cultures.
Table 2: Macroscopic Reactions

<table>
<thead>
<tr>
<th>EFM</th>
<th>Macroscopic Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_1$</td>
<td>$Glc \rightarrow 2 \text{ Lac}$</td>
</tr>
<tr>
<td>$e_2$</td>
<td>$2 \text{ Glc} + 3 \text{ Gln} \rightarrow \text{ Ala} + \text{ Nucleotide} + 9 \text{ CO}_2$</td>
</tr>
<tr>
<td>$e_3$</td>
<td>$\text{ Gln} \rightarrow \text{ Lac} + 2 \text{ N} + 2 \text{ CO}_2$</td>
</tr>
<tr>
<td>$e_4$</td>
<td>$\text{ Gln} \rightarrow 2 \text{ N} + 5 \text{ CO}_2$</td>
</tr>
<tr>
<td>$e_5$</td>
<td>$\text{ Gln} \rightarrow \text{ Ala} + \text{ N} + 2 \text{ CO}_2$</td>
</tr>
<tr>
<td>$e_6$</td>
<td>$2 \text{ Glc} + 3 \text{ Gln} \rightarrow \text{ Lac} + \text{ Ala} + \text{ Nucleotide} + 6 \text{ CO}_2$</td>
</tr>
<tr>
<td>$e_7$</td>
<td>$3 \text{ Glc} \rightarrow 5\text{ Lac} + 3 \text{ CO}_2$</td>
</tr>
<tr>
<td>$e_8$</td>
<td>$2 \text{ Glc} + 3 \text{ Gln} \rightarrow 2 \text{ Lac} + \text{ N} + \text{ Nucleotide} + 6 \text{ CO}_2$</td>
</tr>
<tr>
<td>$e_9, e_{10}$</td>
<td>$\text{ Glc} \rightarrow 6 \text{ CO}_2$</td>
</tr>
<tr>
<td>$e_{11}$</td>
<td>$2 \text{ Glc} + 3 \text{ Gln} \rightarrow \text{ N} + \text{ Nucleotide} + 12 \text{ CO}_2$</td>
</tr>
</tbody>
</table>
Table 3: Stoichiometry of the measured species for the macroreactions of Table 2

<table>
<thead>
<tr>
<th></th>
<th>e_1</th>
<th>e_2</th>
<th>e_3</th>
<th>e_4</th>
<th>e_5</th>
<th>e_6</th>
<th>e_7</th>
<th>e_8</th>
<th>e_9/10</th>
<th>e_11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glc</td>
<td>-1</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>Lac</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gln</td>
<td>0</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>-3</td>
<td>0</td>
<td>-3</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Ala</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nuc</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CO2</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>
Table 4: Success combination of macroreactions passing test (6), where $R$ (9) represents the sum of minimal errors necessary for LP (6) to allow feasible solutions. Note that $R$ has no unit.

<table>
<thead>
<tr>
<th># reactions</th>
<th>Macroreactions combination</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$e_5, e_8$</td>
<td>76.0</td>
</tr>
<tr>
<td>3</td>
<td>$e_1, e_5, e_{11}$</td>
<td>23.9</td>
</tr>
<tr>
<td>4</td>
<td>$e_1, e_4, e_5, e_{11}$</td>
<td>12.6</td>
</tr>
<tr>
<td>5</td>
<td>$e_1, e_3, e_5, e_8, e_9$</td>
<td>7.2</td>
</tr>
<tr>
<td>6</td>
<td>$e_1, e_3, e_5, e_6, e_9, e_{11}$</td>
<td>6.5</td>
</tr>
<tr>
<td>7</td>
<td>$e_1, e_3, e_5, e_6, e_8, e_9, e_{11}$</td>
<td>6.37</td>
</tr>
<tr>
<td>8</td>
<td>$e_1, e_3, e_4, e_5, e_6, e_8, e_9, e_{11}$</td>
<td>6.37</td>
</tr>
<tr>
<td>9</td>
<td>$e_1, e_2, e_3, e_4, e_5, e_6, e_8, e_9, e_{11}$</td>
<td>6.37</td>
</tr>
</tbody>
</table>
Table 5: Best combinations of 5 elementary fluxes selected as candidates to build a macroscopic model

<table>
<thead>
<tr>
<th>Combination number</th>
<th>Macroreaction combinations</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$e_1, e_3, e_5, e_8, e_9$</td>
<td>7.2</td>
</tr>
<tr>
<td>2</td>
<td>$e_1, e_3, e_5, e_9, e_{11}$</td>
<td>7.5</td>
</tr>
<tr>
<td>3</td>
<td>$e_1, e_4, e_5, e_8, e_9$</td>
<td>7.6</td>
</tr>
<tr>
<td>4</td>
<td>$e_1, e_4, e_5, e_9, e_{11}$</td>
<td>8.0</td>
</tr>
</tbody>
</table>
Table 6: Parameters values identified with three perfusion cell cultures (dataset 2, 3 and 4) for Model 1, Model 2 and Model 3. VC corresponds to the variation coefficient expressed in %.

<table>
<thead>
<tr>
<th>Models</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macroreactions involved</td>
<td>1-3-5-9-11</td>
<td>1-3-5-11</td>
<td>1-8-11</td>
<td>1-3-5-9-11</td>
</tr>
<tr>
<td>Datasets for direct validation</td>
<td>1-2-3-4</td>
<td>1-2-3-4</td>
<td>1-2-3-4</td>
<td>1-2-3-4</td>
</tr>
<tr>
<td>Number of multistarts</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Multistart cut-off</td>
<td>J &lt; 10</td>
<td>J &lt; 14</td>
<td>J &lt; 18</td>
<td>J &lt; 9</td>
</tr>
<tr>
<td>Remaining identifications</td>
<td>33</td>
<td>64</td>
<td>80</td>
<td>103</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Param.</th>
<th>Units</th>
<th>A.I.</th>
<th>min θ</th>
<th>θ*</th>
<th>C.V.</th>
<th>max θ</th>
<th>min θ</th>
<th>θ*</th>
<th>C.V.</th>
<th>max θ</th>
<th>min θ</th>
<th>θ*</th>
<th>C.V.</th>
<th>max θ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{max,1}$</td>
<td>h⁻¹</td>
<td>[0.1 – 1]</td>
<td>0.080</td>
<td>0.468</td>
<td>2%</td>
<td>0.959</td>
<td>0.169</td>
<td>0.634</td>
<td>2%</td>
<td>0.975</td>
<td>0.231</td>
<td>0.800</td>
<td>1%</td>
<td>1.000</td>
</tr>
<tr>
<td>$K_{Gl/c,1}$</td>
<td>mM</td>
<td>[0.1 – 30]</td>
<td>1.429</td>
<td>8.112</td>
<td>16%</td>
<td>29.011</td>
<td>5.796</td>
<td>25.961</td>
<td>11%</td>
<td>29.989</td>
<td>5.746</td>
<td>23.912</td>
<td>9%</td>
<td>29.985</td>
</tr>
<tr>
<td>$K_{Gl/n,1}$</td>
<td>mM</td>
<td>[0.01 – 3]</td>
<td>0.050</td>
<td>0.077</td>
<td>35%</td>
<td>1.856</td>
<td>0.015</td>
<td>0.026</td>
<td>42%</td>
<td>0.153</td>
<td>0.028</td>
<td>0.069</td>
<td>18%</td>
<td>0.195</td>
</tr>
<tr>
<td>$\mu_{max,3}$</td>
<td>h⁻¹</td>
<td>[0.01 – 0.2]</td>
<td>0.010</td>
<td>0.032</td>
<td>51%</td>
<td>0.160</td>
<td>0.014</td>
<td>0.044</td>
<td>41%</td>
<td>0.197</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$K_{Gl/c,3}$</td>
<td>mM</td>
<td>[0.1 – 10]</td>
<td>0.092</td>
<td>0.230</td>
<td>30%</td>
<td>9.615</td>
<td>0.000</td>
<td>5.289</td>
<td>81%</td>
<td>9.631</td>
<td>0.157</td>
<td>1.000</td>
<td>63%</td>
<td>9.990</td>
</tr>
<tr>
<td>$K_{Gl/n,3}$</td>
<td>mM</td>
<td>[0.01 – 3]</td>
<td>0.047</td>
<td>0.063</td>
<td>46%</td>
<td>1.930</td>
<td>0.029</td>
<td>0.070</td>
<td>59%</td>
<td>0.870</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$K_{N,3}$</td>
<td>mM</td>
<td>[1 – 10]</td>
<td>0.174</td>
<td>1.832</td>
<td>74%</td>
<td>9.31</td>
<td>0.120</td>
<td>3.153</td>
<td>98%</td>
<td>9.347</td>
<td>1.084</td>
<td>8.233</td>
<td>84%</td>
<td>9.990</td>
</tr>
<tr>
<td>$\mu_{max,5}$</td>
<td>h⁻¹</td>
<td>[0.01 – 0.2]</td>
<td>0.010</td>
<td>0.051</td>
<td>47%</td>
<td>0.099</td>
<td>0.014</td>
<td>0.043</td>
<td>39%</td>
<td>0.069</td>
<td>0.023</td>
<td>0.063</td>
<td>48%</td>
<td>0.129</td>
</tr>
<tr>
<td>$K_{Gl/c,5}$</td>
<td>mM</td>
<td>[0.1 – 10]</td>
<td>0.092</td>
<td>0.230</td>
<td>30%</td>
<td>9.615</td>
<td>0.000</td>
<td>5.289</td>
<td>81%</td>
<td>9.631</td>
<td>0.157</td>
<td>1.000</td>
<td>63%</td>
<td>9.990</td>
</tr>
</tbody>
</table>

Table 7: Identifications results for model 1, 2, 3 and 4. Admissible intervals (A.I.) are set to focus research within the solution space area where the parameters are likely to be found. Results presents the minimum and maximum parameter values of the identifications with residual costs below the defined cut-off threshold as well as the estimated optimal solution $\theta^*$ and the related coefficient of variation C.V.