IDENTIFICATION OF REACTION SCHEMES FOR BIOPROCESSES: DETERMINATION OF AN INCOMPLETELY KNOWN YIELD MATRIX

Olivier Bernard[1], **Georges Bastin**[2]

[1]INRIA-COMORE, BP93, 06902 Sophia-Antipolis Cedex, France
olivier.bernard@inria.fr
[2]UCL-CESAME, av. G. Lemaître 4-6, 1348 Louvain-La-Neuve, Belgium
bastin@auto.ucl.ac.be

Keywords: Modeling, Nonlinear systems, Bioreactors, Validation

Abstract

In this paper we propose a methodology to determine the structure of the yield coefficient matrix K in a mass balance based model and to identify its coefficients from a set of available data. The first step consists in estimating the number of reactions that must be taken into account to represent the main mass transfer within the bioreactor. This provides the dimension of K. Then we propose a method to directly determine the structure of the matrix (*i.e.* mainly its zeros and the signs of its coefficients). These methods are illustrated with simulations of a process of lipase production from olive oil by *Candida rugosa*.

1 Introduction and motivation

The dynamical behaviour of a stirred tank bioreactor is often described by a general mass-balance model of the following form (see e.g. [1, 2]):

$$\frac{d\xi(t)}{dt} = K\, r(t) + v(t), \tag{1}$$

In this model, the vector $\xi = (\xi_1, \xi_2, \ldots, \xi_n)^T$ is made-up of the concentrations of the various species inside the liquid medium. The term $v(t)$ represents the net balance between inflows, outflows and dilution effects. The term $K\, r(t)$ represents the biological and biochemical conversions in the reactor (per unit of time) according to some underlying reaction network. The $(n \times p)$ matrix K is a constant (pseudo-)stoichiometric matrix. $r(t) = (r_1(t), r_2(t), \ldots, r_p(t))^T$ is a vector of reaction rates (or conversion rates).

The stoichiometric matrix K plays a key role in the mass balance modelling. Each column of the matrix corresponds to a chemical or biological reaction of the underlying reaction network. The coefficients $k_{ij} \quad j = 1, \ldots, p$ are associated with the j^{th} reaction. A positive $k_{ij} > 0$ means that the i^{th} species ξ_i is a product of the j^{th} reaction, while a negative $k_{ij} < 0$ means that ξ_i is a substrate of the j^{th} reaction. If $k_{ij} = 0$ the species ξ_i is not involved in the j^{th} reaction.

In this paper, we are concerned with modelling situations where the on-line concentrations ξ_i of the involved species are measured but the structure of the reaction network is a priori questionable and therefore the matrix K is partially unknown. The objective, as in [6], is to provide guidelines to the user for the identification of the structure of the reaction network and the determination of the stoichiometric matrix K from the available data. The problem is illustrated with an example.

Example: Let us consider the example of a competitive growth on two substrates [10] which could represent, for instance, the production of lipase from olive oil by *Candida rugosa*. Here the microorganism is supposed to grow on two substrates that are produced by the hydrolysis of a primary complex organic substrate.

The following 3-step reaction scheme has been assumed in the literature:

- Hydrolysis:

$$k_1 S_1 + E \longrightarrow S_2 + k_2 S_3 + E$$

- Growth on S_2:

$$k_3 S_2 + k_4 O \longrightarrow X + k_5 P$$

- Growth on S_3:

$$k_6 S_3 + k_8 O \longrightarrow X + k_7 E + k_9 P$$

where S_1 is the primary substrate (olive oil), S_2 (glycerol) and S_3 (fatty acid) are the secondary substrates. E is the enzyme (lipase), X the biomass (*Candida rugosa*), O the dissolved oxygen and P the carbon dioxide.

The associated stoichiometric matrix is:

$$K = \begin{pmatrix} -k_1 & 0 & 0 \\ 1 & -k_3 & 0 \\ k_2 & 0 & -k_6 \\ 0 & 0 & k_7 \\ 0 & 1 & 1 \\ 0 & -k_4 & -k_8 \\ 0 & k_5 & k_9 \end{pmatrix}$$

with $k_i > 0$, $i = 1, \ldots, 9$. $\triangle$

We shall assume that this reaction network is unknown to the user and has to be discovered from data of the species concentrations. Here the data will be simulated by a model but of course in practice the data are obtained from experiments.

In particular applications, the choice of a reaction network and its associated stoichiometric matrix K results in general from modelling assumptions. Sometimes however, a complete description of the reaction scheme is *a priori* not available. This can be a consequence of a lack of phenomenological knowledge on some of the involved mechanisms, letting a part of the reaction network questionable. The problem can also arise when it is desired to reduce a complicated given reaction network to a much simpler model. This situation especially occurs for models describing wastewater treatment processes which involve a very large amount of bacterial species and of different molecules to be degraded (see e.g. [8]).

We first propose a method to determine the size of the matrix K i.e. the number of independent reactions that are distinguishable from the available data. Then we show how the structure of the matrix K can be estimated, using the *a priori* available knowledge on the process. By structure we mean the sign and the location of the non-zero entries of the matrix K. In addition, the method can also provide an estimate of the parameters k_{ij} if the available knowledge is sufficient.

2 Determination of the number of reactions

2.1 Introduction

In this section, we intend to determine the minimum number of reactions which are needed in order to explain the observed behaviour of the process, without any prior knowledge on the underlying reaction network. We assume that the vectors $\xi(t)$ of species concentrations and $v(t)$ of inflow/outflow balances are measured during some time interval and exhibit significant variations with time. We assume also that the number of measured variables is larger than the number of reactions: $n > p$. The stoichiometric matrix K and the vector of reaction/conversion rates $r(t)$ are unknown.

2.2 Theoretical determination of $\dim(\mathcal{I}m(K))$

The model equation (1) can be viewed as a linear dynamical system with state ξ and inputs $r(t)$ and $v(t)$ (although we know obviously that r and v may be state dependent). If we take the Laplace transform of this equation, we get:

$$s\Xi(s) = KR(s) + V(s) \tag{2}$$

where $\Xi(s)$, $R(s)$ and $V(s)$ are the Laplace transforms of $\xi(t)$, $r(t)$ and $v(t)$ respectively. A linear filter or smoother with transfer function $G(s)$ can then be used in order to clean the data (noise reduction, decrease of autocorrelations etc ...):

$$U(s) = KW(s) \text{ with } U(s) = G(s)[s\Xi(s) - V(s)]$$

and $W(s) = G(s)R(s)$

or, in the time domain:

$$u(t) = Kw(t) \tag{3}$$

with $u(t)$ and $w(t)$ the inverse Laplace transforms of $U(s)$ and $W(s)$ respectively. They can be computed directly from the data by appropriate filtering/smoothing techniques possibly involving delay operators.

Now the question of the dimension of the matrix K can be formulated as follows: what is the dimension of the image of K, in other words, what is the dimension of the space where $u(t)$ lives. Note that we assume K to be a full rank matrix. Otherwise, it would mean that the same dynamical behaviour could be obtained with a matrix K of lower dimension, by defining other appropriate reaction rates. The determination of the dimension of the $u(t)$ space is a classical problem in statistical analysis. It corresponds to the principal component analysis (see e.g. [9]) that determines the dimension of the vectorial space spanned by the vectors k_i which are the rows of K. To reach this objective, we consider the $n \times N$ matrix U obtained from a set of N records of $u(t)$:

$$U = (u(t_1), \ldots, u(t_N))$$

We will also consider the associated matrix of reaction rates, which is unknown:

$$W = (w(t_1), \ldots, w(t_N))$$

We assume that matrix W is full rank. It means that the reactions are independent (none of the reaction rates can be written as a linear combination of the others). We assume that there are more measurement time instants than state variables: $N > n$.

Property 1 *For a matrix K of rank p, if W has full rank, then the $n \times n$ matrix $M = UU^T = KWW^TK^T$ has rank p. Since it is a symmetric matrix, it can be written:*

$$M = P^T \Sigma P$$

where P is an orthogonal matrix ($P^TP = I$) and

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & & & \ldots & & 0 \\ 0 & \sigma_2 & 0 & & & & 0 \\ \vdots & & \ddots & & & & \\ & & & \sigma_p & & & \\ & & & & 0 & & \\ & & & & & \ddots & \vdots \\ 0 & & \ldots & & & & 0 \end{pmatrix}$$

with $\sigma_{i-1} \geq \sigma_i > 0$ for $i \in \{2, ..., p\}$.

Moreover, the eigenvectors associated to the σ_i generate an orthonormal basis of $\mathcal{I}mK$.

Proof: it is a direct application of the singular decomposition theorem [7] since $\mathrm{rank}\,(M) = \mathrm{rank}\,(KW) = \mathrm{rank}\,(K) = \mathrm{rank}\,(\Sigma) = p$.

Now from a theoretical point of view, it is clear that the number of reactions can be determined by just counting the number of non zero singular values of UU^T.

2.3 Practical implementation

In practice, the ideal case presented above is perturbed for four main reasons:

- The reaction scheme that we are looking for is a first approximation of chemical or biochemical reactions which can be very complex. The "true" matrix K is probably much larger. The reactions that are fast or of low magnitude can be considered as perturbations of a dominant low dimensional reaction network that we are actually trying to estimate
- The measurements are corrupted with noise. This noise can be very important, especially for the measurement of biological quantities which suffer from a lack of reliable sensors.
- The measurements are seldom all available exactly at the same time instant t_i, and therefore they must be interpolated if we need values of $\xi(t_i)$ and $v(t_i)$ at t_i in order to build the vector U.
- In order to compute $u(t)$ we need a numerical implementation of the filter $G(s)$. This can generate additional perturbations.

2.3.1 Data normalisation

In order to avoid conditioning problems and to give the same weighting to all the variables, the data vectors $u(t_i)$ are normalised as follows:

$$\tilde{u}(t_i) = \frac{u(t_i) - e(u)}{\sqrt{N}\sigma(u)}$$

where $e(u)$ is the average value of $u(t_i)$, and $\sigma(u)$ their standard deviation.

2.3.2 Practical determination of the number of reactions

In practice, for the reasons we have mentioned above, it is well known that there are no zero eigenvalues for the matrix $M = UU^T$.

The question is then to determine the number of eigenvectors that must be taken into account in order to produce a reasonable approximation of the data $u(t)$. To answer that question, let us remark that the eigenvalues σ_i of M correspond to the variance associated with the corresponding eigenvector (inertia axis) [9].

The method then consists in selecting the p first principal axis which represent a total variance larger than a fixed confidence threshold.

For instance, in the next example, we will consider a threshold (depending on the information available on noise measurements) at 95% of the variance. This leads to the selection of 3 axis, and therefore $p = 3$.

Remark: if $\mathrm{rank}\,(M) = n$ it means that $rank(W) \geq n$. In such a case we cannot estimate p and measurements of additional variables are requested in order to apply the method presented here.

Parameter	Values	Units
c_0	0.5	g/l.day^{-1}
c_1	3	day^{-1}
c_2	1	g/l
c_3	0.2	g/l
c_4	20	g.day$^{-1}l^{-1}$
c_5	1	g/l
c_6	0.2	g/l
c_7	2	g^2/l^2
c_8	2	g/l
c_9	0.2	g/l
c_{10}	5	day^{-1}
c_{11}	15	g/l
c_{12}	0.5	day^{-1}
c_{13}	0.5	g/l

Table 1: Parameter values.

Example:

We come back to the example of a competitive growth on two substrates which has been introduced above. For the simulation purpose, we assume that the kinetics of the three reactions are given by the following expressions :

$$\phi_1(S_1, E) = c_0 \frac{S_1}{S_1+c_8} \frac{E}{E+c_9}$$

$$\phi_2(S_2, O, X) = c_1 \frac{S_2}{S_2+c_2} \frac{O}{O+c_3} X$$

$$\phi_3(S_2, O, X) = c_4 \frac{S_3}{(S_3+c_5)(S_2+c_6)} \frac{O^2}{O^2+c_7};$$

The transfer between liquid and gaseous phase is represented by the classical Henry's law:

$$qco_2 = c_{10}(P - c_{11}) \text{ and } qo_2 = c_{12}(O - c_{13})$$

The values of the coefficients c_i can be found in Table 1. The

matrix K is chosen as follows:

$$K = \begin{pmatrix} -3 & 0 & 0 \\ 1 & -5 & 0 \\ 0.3 & 0 & -0.5 \\ 0 & 0 & 0.2 \\ 0 & 1 & 1 \\ 0 & -2 & -1 \\ 0 & 0.3 & 1.5 \end{pmatrix}$$

A 30 day run of the model has been performed. The collected data have been corrupted with a white noise of high magnitude (30% of the standard deviation of each component) and sampled. Finally 380 data points are available.

The data (before sampling) are presented on Figure 1. The state variables S_2, S_3, E, X, P, O and of the gaseous flow rates q_{O_2} and q_{CO_2}have been measured. We assume here that the state variable S_1 was not recorded in order to illustrate the fact that our approach is applicable even if the full set of state variables is not available for measurement. Moreover the dilution rate and the substrate inflow rate (see Figure 2) have been selected in order to guarantee that the system is enough excited and therefore that the recorded signals will have a sufficiently informative content to expect good identification results. The

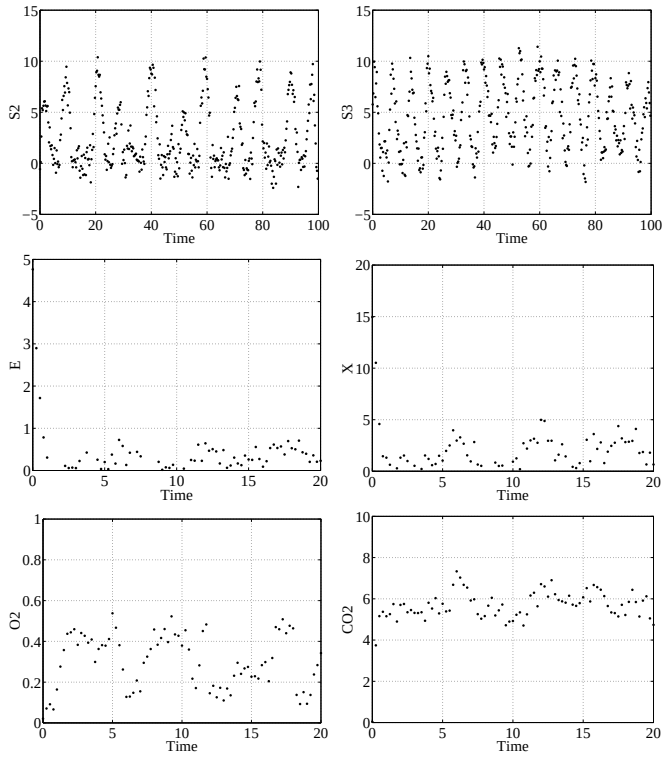

Figure 1: Experiment simulated from the kinetic modelling corrupted with an additive white noise.

vectors $u(t_i)$ are then computed from these data and subsequently normalised as explained before. Finally, the eigenvectors of UU^T are computed.

Figure 3 represents the cumulated variance associated with the

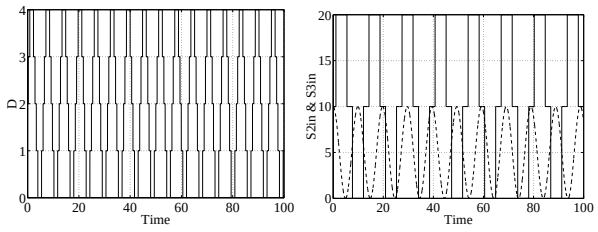

Figure 2: Dilution rate and influent concentrations S_{2in} and S_{3in} used for the simulated experiment.

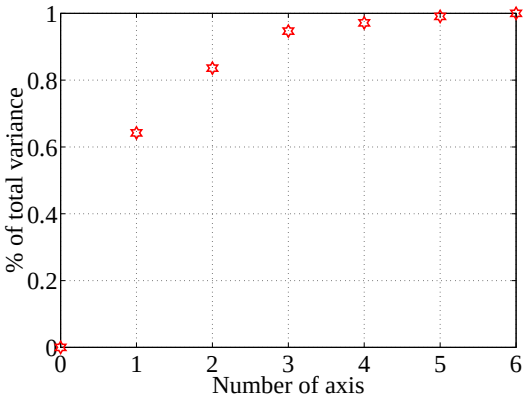

Figure 3: Total variance explained with respect to the number of reactions for the production of lipase from olive oil by *Candida rugosa*.

number of considered inertia axis. For instance, we can see that two reactions are sufficient to explain 82% of the observed variance. Since three reactions explain 95% of the total variance, it seems reasonable in this example to use 3 reactions for the model. △

The reader can refer to [4] for an application to real data, for growth and vanillin production by cultures of the fungus *Pycnoporus cinnabarinus* in bioreactors.

3 Estimation of the stoichiometric matrix K

Now that we have a value for the number of involved reactions, we are in a position to start the estimation of the (totally or partially) unknown matrix K.

3.1 Determination of $\mathcal{I}mK$

Let us use the property 1 which states that $\mathcal{I}mK$ is spanned by the eigenvectors ρ_i associated with the non zero eigenvalues of UU^T. Now, from the experimental data collected through the matrix UU^T we get p eigenvectors ρ_i that span K. It means that each column k_i of K is a linear combination of the ρ_i. In other terms, there exists a $p \times p$ matrix G such that

$$K = \rho G$$

where the columns of matrix ρ are the eigenvectors ρ_j. In other words, the family of possible stoichiometric matrices K is parameterised by G.

Remark: In general, since the reaction rates are unknown, the matrix G (and therefore the matrix K) is not identifiable: this can be easily understood on a very simple example. If $r_1(\xi)$ and $r_2(\xi)$ are two reaction rates, the term $Kr(\xi)$ can be written:

$$\begin{array}{rcl} Kr(\xi) = & k_1 r_1(\xi) + k_2 r_2(\xi) \\ = & \frac{k_1+k_2}{2}(r_1(\xi)+r_2(\xi)) + \frac{k_1-k_2}{2}(r_1(\xi)-r_2(\xi)) \end{array}$$

And therefore matrices $K = [k_1\ k_2]$ and $\bar{K} = [\frac{k_1+k_2}{2}\ \ \frac{k_1-k_2}{2}]$ can both produce the same result. The reaction rates associated with the second matrix are then: $\bar{r}_1(\xi) = r_1(\xi) + r_2(\xi)$ and $\bar{r}_2(\xi) = r_1(\xi) - r_2(\xi)$.

3.2 Additional hypotheses

In order to make the matrix G (and K) uniquely identifiable, we need to introduce additional structural constraints.

First, we shall impose (without loss of generality) that each reaction rate is normalised with respect to one species, and therefore that each column of the matrix K contains one +1 or one -1. This induces obviously additional constraints on the possible matrices G. Note that sometimes we may not know the sign of the element: the two possible cases must then be considered.

When additional constraints are still necessary, we use biological assumptions.

For instance, we can assume that a specific component is not involved in one of the p reactions (meaning that there is a zero in K). It is clear for example that the first reaction will involve only the substrates which were present at the beginning of the fermentation. We can also impose the conservation of elementary mass balances, or at least only allow for a leak of mass in the system. One can also try to find a matrix K involving the minimum number of components in each reaction (*i.e.* containing the maximum number of zeros). If these hypotheses are not sufficient, several matrices K can then be identified, parameterised by some parameter, and their biological meaning must then be assessed.

3.3 Validation

The main result provided by the previous analysis is the determination of the variables which are substrates or products in the reactions or, in other words, the obtained signs of the entries of K.

Another expected result can be the determination of the variables which are not involved in a reaction, corresponding to zero elements in the matrix K. However it is actually very unlikely that the analysis will provide estimates of the elements of K which are exactly zero. The idea consist then in replacing the very small elements by zeros, and to validate the corresponding reaction scheme using the techniques presented in [3, 5].

Example:

We shall now illustrate the proposed approach with the simulation study of lipase production from olive oil. From the previous study of the number of reactions, we know that 3 reactions should be considered.

We assume here that the first reaction is known, and therefore we only focus on the two other reactions.

K **used for simulation**	**identified matrix** K
$\begin{pmatrix} -5 & 0 \\ 0 & -0.5 \\ 0 & 0.2 \\ 1 & 1 \\ -2 & -1 \\ 0.3 & 1.5 \end{pmatrix}$	$\begin{pmatrix} -3.54 & 0 \\ 0 & -0.51 \\ 0.01 & 0.22 \\ 1 & 1 \\ -1.34 & -0.87 \\ 0.18 & 1.51 \end{pmatrix}$

Table 2: True coefficients of matrix K and identified values.

A set of noisy data of the state variables S_2, S_3, E, X, P, O and of the gaseous flow rates q_{O_2} and q_{CO_2} is produced by simulation as described in Section 2. The goal is to determine the 6×2 matrix K from this data set. More specifically, a question that we want to address is to determine, from the data, which of the two reactions produces the enzyme E.

Now we can compute the quantities U_i associated with the 6 state variables using a moving average. Next we compute the matrix $M = U^T U$. The eigenvectors ρ_i associated with the two largest eigenvalues are then the basis of $\mathcal{I}mK$. Since G is a 2×2 matrix, the columns k_1 and k_2 of K can be written:

$$k_1 = \alpha_{11}\rho_1 + \alpha_{12}\rho_2 \text{ and } k_2 = \alpha_{21}\rho_1 + \alpha_{22}\rho_2 \qquad (4)$$

Now we proceed in two successive steps:

i. Normalisation.

The stoichiometric coefficients associated to the biomass growth are normalised : $k_{14} = 1$ and $k_{24} = 1$. We get then:

$$\begin{array}{l} k_{14} = 1 = \alpha_{11}\rho_{14} + \alpha_{12}\rho_{24} \\ k_{24} = 1 = \alpha_{21}\rho_{14} + \alpha_{22}\rho_{24} \end{array} \qquad (5)$$

Using equations 4 and 5 with the obtained values of ρ_1 and ρ_2, we can now write matrix K parametrised by α_{11} and α_{22} as follows:

$$K = \begin{pmatrix} -1.42\alpha_{11} - 2.65 & -1.2\alpha_{22} + 1.12 \\ 0.2\alpha_{11} - 0.13 & 0.17\alpha_{22} - 0.67 \\ -0.08\alpha_{11} + 0.062 & -0.071\alpha_{22} + 0.28 \\ 1 & 1 \\ -0.19\alpha_{11} - 1.2 & -0.16\alpha_{22} - 0.72 \\ -0.53\alpha_{11} + 0.51 & -0.45\alpha_{22} + 1.93 \end{pmatrix}$$

ii. Additional hypotheses.

Now to determine uniquely matrix K two additional assumptions must be introduced.

Hypothesis: A reaction still takes place when only S_2 [resp. S_3] is present at the initial time, and no S_3 [resp. S_2] is produced.

In other words this means that S_2 is the only substrate of one reaction and that S_3 is the only substrate of the other one. Thus we will impose $k_{12} = 0$ and $k_{21} = 0$.

These additional constraints allows us to compute α_{11} (0.621) and α_{22} (0.93).

Finally we end up with an estimate of the matrix K (see Table 2). It is worth noting that the identified matrix K is close to the true one. The value of the (theoretically zero) coefficient k_{13} is 0.01 which can be neglected with respect to the other coefficients of K. Hence, the unknown part of the structure of the matrix K has been recognised. Moreover the estimates of the non-zero entries of the matrix K are quite accurate. $\triangle$

4 Conclusion

Modelling of bioprocesses is a difficult issue since there does not exist any laws on which the model can rely as in other fields like mechanics or electronics. Therefore it is very important to check the model adequation with the data. The proposed method should guarantee a mass balance based model whose complexity is in adequation with the data.

Acknowledgement: This work has been carrying out with the support provided by the European commission, Information Society Technologies programme, Key action I Systems & Services for the Citizen, contract TELEMAC number IST-2000-28256.

References

[1] G. Bastin and D. Dochain. *On-line estimation and adaptive control of bioreactors*. Elsevier, 1990.

[2] G. Bastin and J. VanImpe. Nonlinear and adaptive control in biotechnology: a tutorial. *European Journal of Control*, 1(1):1–37, 1995.

[3] O. Bernard and G. Bastin. Structural identifi cation of nonlinear mathematical models for bioprocesses. In *Proceedings of the Nonlinear Control Systems Symposium 98*, pages 449–454. Enschede, July 1-3, 1998, 1998.

[4] O. Bernard, G. Bastin, C. Stentelaire, L. Lesage-Meessen, and M. Asther. Mass balance modelling of vanillin production from vanillic acid by cultures of the fungus *pycnoporus cinnabarinus* in bioreactors. *Biotech. Bioeng*, pages 558–571, 1999.

[5] O. Bernard and G.Bastin. First step in mass balance modeling of bioprocesses: estimation and validation of the stoechiometric matrix . *Math. Biosciences*, submitted.

[6] P. Bogaerts and A. Vande Wouwer. Sytematic generation of identifi able macroscopic reaction schemes. In *Proceedings of the 8th IFAC Conference on Computer Applications in Biotechnology (CAB8)*. Montral, Canada, 2001, 2001.

[7] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge MA, 1993.

[8] IWA Task Group for Mathematical Modelling of Anaerobic Digestion Processes. *Anaerobic Digestion Model No. 1 (ADM1)*. IWA Publishing, London, 2002.

[9] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*. Prentice Hall, 1992.

[10] P. Serra, J. del Rio, J. Robust, M. Poch, C. Sola, and A.Cheruy. A model for lipase production by candida rugosa. *Bioprocess Engineering*, 8:145–150, 1992.