Can we assess the model complexity for a bioprocess ? Theory and example of the anaerobic digestion process

O. Bernard*, B. Chachuat**, A. Hélias* and J. Rodriguez***

* INRIA Comore, 2004 Route des Lucioles, B.P. 93, 06902 Sophia-Antipolis Cedex, France.
(e-mail: Olivier.Bernard@inria.sophia.fr)

** MIT, Department of Chemical Engineering, Cambridge, MA 02139-4307, USA.
(e-mail: bchachua@mit.edu)

*** Dep. of Chemical Engineering, University of Santiago de Compostela,
rúa Lope Gómez de Marzoa s/n ES-15782 Santiago de Compostela, Spain.
(e-mail: jorger@usc.es)

Abstract
In this paper we propose a methodology to determine the structure of the pseudo-stoichiometric coefficient matrix K in a mass balance based model, *i.e.* the maximal number of biomasses that must be taken into account to reproduce an available data set. It 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. The method is applied to data from an anaerobic digestion process and shows that even a model including a single biomass is sufficient. Then we apply the same method to the "synthetic data" issued from the complex ADM1 model, showing that the main model features can be obtained with 2 biomasses.

Keywords
Anaerobic digestion; Bioreactors; Modelling; Nonlinear systems

INTRODUCTION AND MOTIVATION

Modelling of biological processes is a long and tedious task for which theoretical tools are lacking. The difficulty is even exacerbated for wastewater treatment processes which include a broad range of substrate and wide variety of biomasses. Even worse, the influent substrate, with varying concentration and composition is rarely completely known. As a result the bacterial ecosystem degrading the influent often consists in a complex consortium of bacteria, with possible species successions. On top of this there is a lack of sensors to monitor the evolution of all the process variables, and one has most of the time to deal with aggregated variables, such as chemical oxygen demand (COD), volatile suspended solids (VSS), etc. that are very raw indicators of the state process.
In these conditions the question of how to design a model is crucial. Especially the trade-off between model complexity – allowing to represent most of the known phenomena – and adequation with the available experimental information is capital. In this paper, we want to address this problem, and propose a method to assess the model complexity (in a sense that will be defined latter on) with respect to a given data set.
To achieve this goal, we assume that the process can be represented by a general mass balance model often used to represent the dynamical behaviour of a stirred tank bioreactor (see e.g. (Bastin and Dochain, 1990; Bastin and van Impe, 1995)):

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

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). It is supposed to depend on the state ξ and on external environmental factors such as temperature, light or pressure, etc.
The pseudo-stoichiometric (PS) 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} means that the i^{th} species ξ_i is a product of the j^{th} reaction, while a negative k_{ij} 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 unknown. The objective, is to provide guidelines to the user to determine the size of reaction network from the available data.
The usual approach dedicated to the determination of reaction networks relies on the linearisation of the dynamics around a reference solution (Eiswirth *et al.*, 1991; Chevalier *et al.*, 1993) and identification of the local Jacobian matrix. Here, in the spirit of (Chen and Bastin, 1996; Bernard and Bastin, 2005), we exploit the structure of the bioprocess (equation (1)) and our arguments do not rely on any linearisation.
Generally, the choice of a reaction network and its associated PS matrix K results from modelling assumptions. Sometimes however, several choices are possible between reaction networks of various complexities. The problem can also arise when it is desired to reduce a complicated given reaction network to a much simpler model in order to achieve a better adequation between model and available information quality.
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 apply this method on data issued from an anaerobic digestion process and compare two potential models. Finally we analyse the "synthetic data" simulated with the complex ADM1 model (Batstone *et al.*, 2002) including 7 biomasses, and show that its main behaviour can be roughly simplified using only 2 biomasses.

DETERMINATION OF THE NUMBER OF REACTIONS

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 PS matrix K and the vector of reaction/conversion rates $r(t)$ are unknown.

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. The vector $u(t)$ can be computed directly from the data by appropriate filtering/smoothing techniques possibly involving delay operators.
For example, the moving average is a very simple filter that can be applied to (1), and provides an expression of the form (3) with (T denotes the considered moving average window):

$$u(t) = \frac{1}{T}\left[\xi(t) - \xi(t-T) \; - \int_{t-T}^{t} v(\tau)d\tau\right] \text{ and } w(t) = \frac{1}{T}\left[\int_{t-T}^{t} r(\tau)d\tau\right] \tag{4}$$

This moving average was used in the considered example.
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. (Johnson and Wichern, 1992)) that determines the dimension of the vector 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 estimates 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 consider more time instants t_i 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 $i \in \{2, ..., p\}$.

This property is a direct application of the singular decomposition theorem (Horn and Johnson, 1993) since $\text{rank}\ (M) = \text{rank}\ (KW) = \text{rank}\ (K) = \text{rank}\ (\Sigma) = p$.
Now from a theoretical point of view, it is clear that the number of reactions can be determined by counting the number of non zero singular values of UU^T.

Practical implementation

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

- The reaction network 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 by noise. This noise can be very important, especially for the measurement of biological quantities for which reliable sensors are not available.
- In order to compute $u(t)$ we need a numerical implementation of the filter $G(s)$. Moreover an interpolation is often required to estimate the values of $\xi(t_i)$ and $v(t_i)$ at the same time instants t_i. These processes generate additional perturbations.

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:

$$\tilde{u}_i(t_j) = \frac{u_i(t_j) - a(u_i)}{\sqrt{N} s(u_i)}$$

where $a(u_i)$ is the average value of the $u_i(t_k)$ for $k \in \{1..N\}$, and $s(u_i)$ their standard deviation.

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) (Johnson and Wichern, 1992).
The method then consists in selecting the p first principal axis which represent a total variance larger than a fixed confidence threshold.
Remark: if $\text{rank}\ (M) = n$ 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.

APPLICATION TO REAL DATA FROM AN ANAEROBIC DIGESTER

Process presentation

In this section we consider the data set considered in (Bernard *et al.*, 2001) which has been acquired on a a fully instrumented fixed bed anaerobic digester (Steyer *et al.*, 2002), located in Narbonne (France), at the "Laboratoire de Biotechnologie de l'Environnement" (LBE) of INRA. Raw industrial

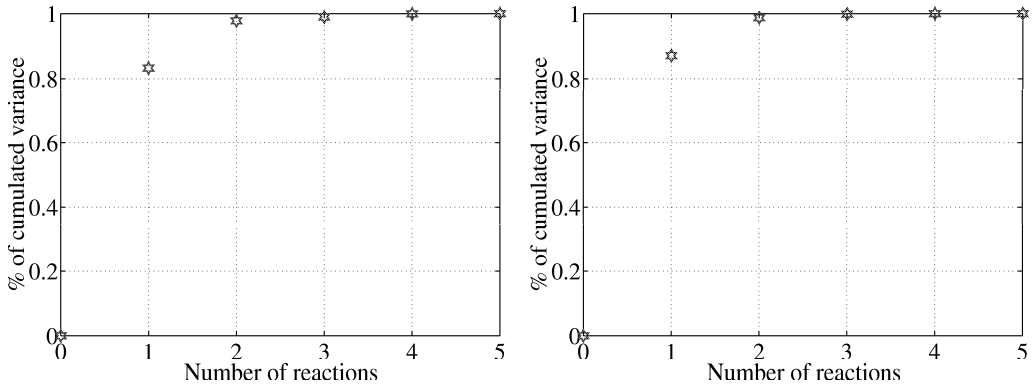

Figure 1: Cumulated variance with respect to the number of chosen axis for 70 days of experiments (see (Bernard *et al.*, 2001)). Left: real data. Right: virtual plant (ADM1 model).

distillery wastewaters obtained from local wineries in the area of Narbonne, France, were used. They have changing characteristics according to the wineries where the wastewater is taken from. The process is a pilot-scale up-flow anaerobic fixed bed reactor and has a circular column of 3.5m height, 0.6m diameter and a useful volume of about 1m^3 (Steyer *et al.*, 2002). This process has a classical on-line instrumentation gathering measurements every 3 minutes of liquid flow rates, temperature and pH in the reactor and biogas flow rate and composition (*i.e.*, CO_2, CH_4 and H_2 content in the biogas (Steyer *et al.*, 2002)). Manual sampling were carried out once a day to measure soluble chemical oxygen demand (COD) in the liquid phase, the volatile fatty acids (VFA) and volatile suspended solids (VSS). The data set consist then in a series of measurements of CH_4 and CO_2 flow rates, total alkalinity, total inorganic carbon, COD,VFA and VSS.

Results

The proposed method was applied to the available data set and the obtained variance distribution is represented in Figure 1. It is worth noting that a reaction network involving only 1 biomass (and thus one reaction) represents 83.2% of the variability. With 2 biomasses, 97.8% of the variability are represented, which justified the choice of the model presented in (Bernard *et al.*, 2001).
This analysis proves that even a very simple model, consisting in a single biomass would already be potentially able to reproduce the observed data. We thus considered a simple modified Haldane model (Andrews, 1968), where the reaction scheme consists then simply in one reaction:

$$k_T S_T \xrightarrow{r_T} X_T + k_{6T}\, CH_4 + k_{4T}\mathrm{CO}_2 \tag{5}$$

We consider that a proportion α of the biomass is in the liquid phase and is therefore affected by the dilution. We obtain the following model:

$$(\mathrm{AMH1}) \left\{ \begin{array}{lcl} \dot{X}_T = & = & (\mu_T(S_T) - \alpha D)X_T \\ \dot{S}_T & = & D(S_{Tin} - S_T) - k_T \mu_T(S_T) X_T \end{array} \right. \tag{6}$$

With Haldane bacterial kinetics: $\mu_T(S_T) = \bar{\mu}_{Tmax} \frac{S_T}{S_T + K_{ST} + \frac{S_T^2}{K_{IT}}}$
The methane flow rate can then be computed: $q_M(S_T, X_T) = k_{6T}\, \mu_T(S_T) X_T$
This simple model can of course not predict the concentration of VFA, the TIC or the gaseous flow rate of CO_2.

	AMH1	AM2	ADM1
State variables	2	6	26
Biomasses	1	2	7
Number of reactions	1	2	19
Parameters	5	13	86
Outputs	3	8	32

Table 1: Complexity of the 3 considered models. Outputs are defined as quantities that can be compared to possible measurements (e.g., VFA, pH, VSS, etc.).

After a phase of model calibration (see (Chachuat *et al.*, 2004) for more details) we were then able to compare model and data. The results are presented on Figure 2.

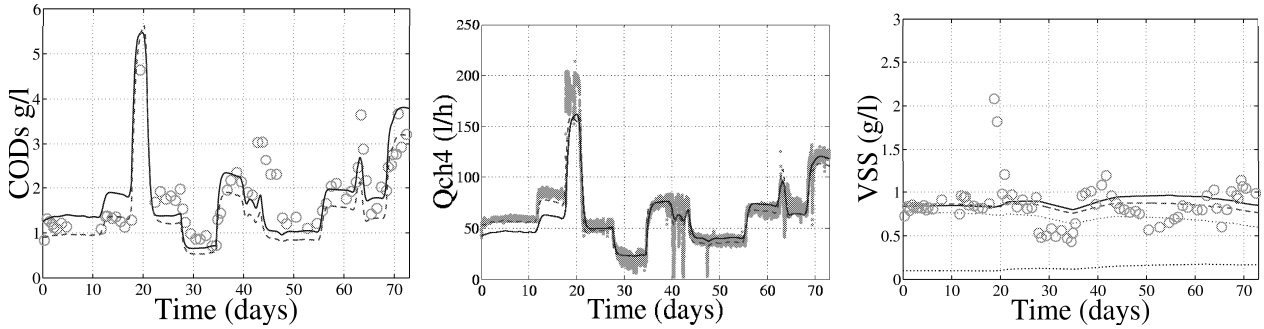

Figure 2: Comparison between simulation results and measurements (o) for COD, methane flow rate and VSS. AMH1 model (—) and model AM2 (- -) as presented in (Bernard *et al.*, 2001) .

It demonstrates that this very simple model is able to properly describe the behaviour of the soluble COD and of the methane flow rate.
Of course this simplistic model will not be able to predict the VFA concentration or the TIC concentration, unless a fixed ratio *e.g.* with COD is assumed.
As a consequence, this simple 1-biomass AMH1 model is suitable to base a strategy for COD regulation (Mailleret *et al.*, 2004), provided that the system does not reach an overload situation.

APPLICATION TO SYNTHETIC DATA ISSUED FROM THE ADM1 MODEL

Introduction

In this section we will consider the data produced by a “virtual plant” made of the model ADM1 (Batstone *et al.*, 2002) which was implemented using Matlab Simulink. This model describes with much more details the various pathways involved in anaerobic digestion. As a result, the complexity of this 7 biomass model is much higher than for the previous described models 1. The model ADM1 has been roughly calibrated by mainly modifying the solid retention time in order to be qualitatively agreement with the data presented in the previous section. However it is clear that a calibration procedure -which turn out to be a very tedious task for this complex model- would probably lead to a very good fit with the data.

Results

The synthetic data provided by this virtual process were then sampled at the same frequency than the real plant and analysed using the same procedure. The result is presented on Figure 1 and shows that despite the model complexity, the main features of the generated can *a priori* be reproduced by a 1-Biomass model (87.1 % of variance) or by a 2 Biomasses model (98.7 % of variance).
Finally, it appears on Figure 3 that both model AM2 (Bernard *et al.*, 2001) and ADM1 (Batstone *et al.*, 2002) are able to reproduce the limited set of data. Of course, model ADM1 is able to predict far more variables (and especially the various volatile fatty acids), and can for example forecast a propionate accumulation. However model AM2 is suitable to base a controller whose objective would be the regulation of VFA, the regulation of the ratio of methane flow rate over CO_2 flow rate, or any other objective involving one of the 6 model variables.

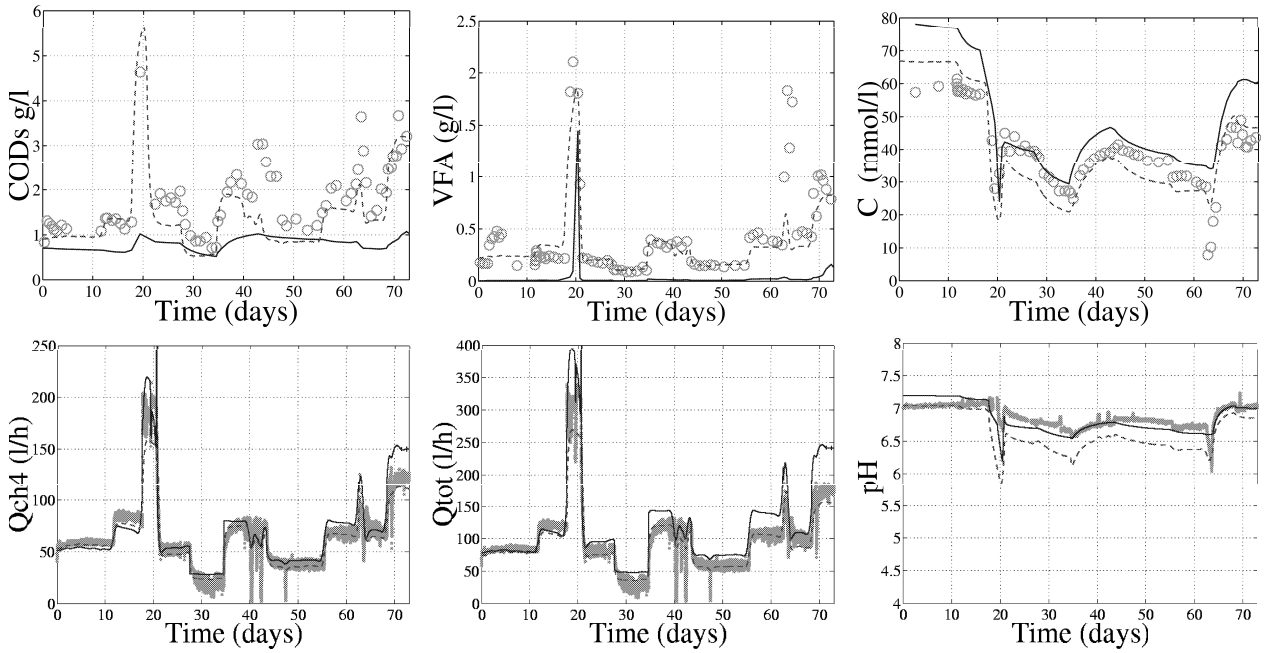

Figure 3: Comparison between real data (o), model AM2 with 2 biomasses (Bernard *et al.*, 2001) (- -) and model ADM1 with 7 biomasses (—) (Batstone *et al.*, 2002).

CONCLUSION

Determining a reaction network for a bioprocess is a difficult issue mainly because of the complexity inherent to biological systems. We show in this paper how to identify the space generated by the columns of K in order to determine the minimum number of reactions (or biomasses) requested to reproduce the data.
The method allows to show that surprisingly, even very simple models can accurately reproduce some considered variables. These minimal models will be specifically useful for developing advanced controllers which generally cannot deal with complex models leading to mathematical intractability.
The second point that was shown is that a complex model can have a behaviour reducible to a much simpler model (at least in some working domain). Once again this justifies the idea of using simplified models to base automatic algorithms (controllers, software sensors, fault detection, etc.). The more complicated models including most of the available phenomenological knowledge on the process can

then be used as a virtual plant to test the ability of the advanced algorithms to reach their objectives in more realistic conditions limiting then the number of long and expensive experiments to be carried out.

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

References

Andrews, J.F. (1968). A mathematical model for the continuous culture of microorganisms utilizing inhibitory substrates. *Biotechnology & Bioengineering* **10**, 707–723.

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

Bastin, G. and J. F. van Impe (1995). Nonlinear and adaptive control in biotechnology: a tutorial. *European Journal of Control* **1**(1), 1–37.

Batstone, D., J. Keller, R. I. Angelidaki, S. V. Kalyuzhnyi, S. G. Pavlostathis, A. Rozzi, W. T. M. Sanders, H. Siegrist and V. A. Vavilin (2002). *Anaerobic Digestion Model No.1 (ADM1)*. IWA Publishing. London.

Bernard, 0., Z. Hadj-Sadok, D. Dochain, A. Genovesi and J.-P. Steyer (2001). Dynamical model development and parameter identification for an anaerobic wastewater treatment process. *Biotechnology & Bioengineering* **75**, 424–438.

Bernard, O. and G. Bastin (2005). On the estimation of the pseudo-stoichiometric matrix for macroscopic mass balance modelling of biotechnological processes. *Mathematical Biosciences* **193**(1), 51–77.

Chachuat, B., O. Bernard and J.-P. Steyer (2004). A two-step procedure for estimating the parameters in mass-balance based bioprocess models. In: *Proc. Watermatex 2004*. Beijing, China.

Chen, L. and G. Bastin (1996). Structural identifiability of the yield coefficients in bioprocess models when the reaction rates are unknown. *Math. Biosciences* **132**, 35–67.

Chevalier, T., I. Schreiber and J. Ross (1993). Toward a systematic determination of complex reaction mechanisms. *J. Phys. Chem* **97**, 6776 – 6787.

Eiswirth, M., A. Freund and J. Ross (1991). *Mechanistic classification of chemical oscillators and the role of species*. Chap. 1, pp. 127–199. Vol. 80 of *Advances in Chemical Physics*. Wiley. New-York.

Delbès, C., R. Moletta and J. J. Godon (2001). Bacterial and archaeal 16S rDNA and 16S rRNA dynamics during an acetate crisis in an anaerobic digestor ecosystem. *FEMS Microbiology Ecology* **35**, 19–26.

Horn, R. A. and C. R. Johnson (1993). *Matrix analysis*. Cambridge University Press, Cambridge MA.

Johnson, R. A. and D. W. Wichern (1992). *Applied multivariate statistical analysis*. Prentice Hall.

Mailleret, L., O. Bernard and J.-P. Steyer (2004). Robust nonlinear adaptive control for bioreactors with unknown kinetics. *Automatica* **40**(8), 365–383.

Steyer, J. P., J. C. Bouvier, T. Conte, P. Gras and P. Sousbie (2002). Evaluation of a four year experience with a fully instrumented anaerobic digestion process. *Water Science and Technology* **45**, 495–502.