10th International IFAC Symposium on
Computer Applications in Biotechnology
Preprints Vol.1, June 4-6, 2007, Cancún, Mexico

KNOWLEDGE BASED DISCOVERY IN FED-BATCH BIOPROCESS

Andrei Doncescu * **Sebastien Regis** **

* *LAAS CNRS, Avenue du Colonel Roche,31007 Toulouse France*
** *GRIMAAG Group, French West Indies University, 97159 Pointe-à-Pitre Guadeloupe, France*

Abstract: We present a data mining approach based on a clustering method to detect and characterize states in a fed-batch processes. This method is based on the detection of singularities in biochemical signals and on the correlation between these signals. A segmentation based on maxima of wavelets transform is used to make an adaptive and dynamical correlation of the signals. The segmentation enables the detection of the borders of states whereas the correlation enables to characterize the physiological states. The method is applied successfully on a fed-batch process and particular states (difficult to detect with classical methods of classification) are detected and characterized.

Keywords: Bioreactor, Fed-batch fermentation, Classification, Wavelet, Correlation.

1. INTRODUCTION

Yeasts are a very well-studied micro-organisms and today, such micro-organism like *Saccharomyces Cerevisiae* are largely used in various sectors of the biomedical and biotechnology industrial bioproces. So, this is a critical point to control such processes. Model-based methods are the most used tool for the bioprocesses because of the mathematical modelisation of the phenomena (see (Roels, 1983)), but these methods using simulation techniques can lead to wrong conclusions because of lack of description parameters or during an unexpected situation. Nowadays non-model-based methods have an increasing success in bioprocesses. The non-model-based methods are mainly based on the analysis of biochemical signals (also called biochemical parameters). Two directions have been explored:

(1) the "manual" *on-line* analysis : it does not allow to identify in an instantaneous manner and with certainty the physiological state of the yeast.
(2) the "manual" *off-line* analysis : it allows to soundly characterize the current state, but generally too late to take into account this information and to adjust the process on the fly by actions of regulators allowing to adjust some critical parameters such that pH, temperature (addition of basis, heat, cooling).

To remedy these drawbacks, computer scientists in collaboration with micro-biologists develop tools for supervised control of the bioprocess. They use the totality of informations provided by the sensors during a set of sample processes to infer some general rules to which the biological process obeys (see for example in (Aguilar-Martin *et al.*, 1999)). These rules (Steyer *et al.*, 1991) can be used to control the next processes. Classification, supervised methods, learning and more generally data mining are also used for these

bioprocess. For application like batch processes (where all physiological states are well known), all these different methods give good results but for processes as fed-batch, where all the states are not known very well, it is more difficult to apply supervised methods. In this paper we present an unsupervised method whose aim is to detect and characterize the states of fed-batch process. It is based on adaptive segmentation and correlation. A similar approach (Régis *et al.*, 2004) was used to detect the differences between physiological states and the command action, but it was not used to detect all the states in a process. The paper is organized as follow. In the section 2, we describe the problem of bioprocess using yeasts. In the section 3, we present the related work which have motivated this approach. In the section 4, the method is presented and first results are presented in section 5. A conclusion is made in section 6.

2. YEAST PRODUCTION APPLICATION: EXAMPLE

The methodology has been applied to a biotechnological process. *Saccharomyces Cerevisiae* is studied under oxidative regime (i.e. no ethanol production) to produce yeast under a laboratory environment in a bioreactor. Two different procedures are applied: a batch procedure that is followed by a continuous procedure. The batch or fed batch procedure is composed by a sequence of biological stages. This phase can be thought as a start-up procedure. Biotechnologists state that the behaviour in the batch procedure influences later in induced phenomena in the further phase. So complete knowledge of the discontinuous phase is of great importance for the biotechnologist. The traditional way to get acquainted of such knowledge is at present carried out through offline measurements and analysis which most of the time produce results when the batch procedure has ended, thus lacking of real time performance. Instead, the proposed methodology allows for real time implementation. This example deals with the batch procedure. Among the set of available online signals the expert chooses the subset of signals which, according to the expert knowledge contain the most relevant information to determine the physiological state:

(1) DOT : partial oxygen pressure in the medium
(2) O2 : oxygen percent in the output gas
(3) CO2 : carbon dioxide percent in the output gas
(4) pH
(5) OH- ion consumption : derived from control action of the pH regulator and the index of reflectivity

The consumption of negative OH ions is evaluated from the control signal of the pH regulator. The actuator is a pump, switched by an hysteresis relay, that inoculates a basic solution (NaOH). The reflectivity, which is measured by the luminance, seems to follow the biomass density. Nevertheless its calibration is not constant and depends on the run.

Our application focus on the evolutive behavior of a *bio-reactor* (namely yeast fermentation) that is to say an evolutive biological system whose interaction with physical world, described with pH, pressure, temperature, mixing antifoam addition, etc..., generates an observable reaction. This reaction is studied by the way of a set of sensors providing a large amount of (generally) numerical data, but, thanks to the logical framework, symbolic data could also be integrated in the future. For an approach based upon classification and fuzzy logic, one can see (Aguilar-Martin *et al.*, 1999) : this work is devoted to discover the different states of the bio-reactor but not to predict its behavior.

In a yeast culture, measures result of biology phenomena and physical mechanisms. That is why to bring the culture, it is always decisional between biology and physico-chemical. The biological reaction is function of the environment and an environmental modification will improve two types of biological responses. The first one is a quasi steady-state response, the micro-organism is in equilibrium with the environment. The biological translation of this state is kinetics of consummation, production and this phenomenon is immediate. The second biological response is a metabolic one, which can be an oxidative or fermentative mode, or a secondary metabolism. The characteristic of this response is that the time constants are relatively long. For cultures, in term of production, the essential parameters are metabolism control and performance (productivity and substrate conversion in biomass yield). With this goal, the process must be conducted by a permanent intervention in order to bring the culture to an initial point to a final point. This control can be done from acquired measures on process, which are generally gases. Indirect measures show the environmental dynamic, which is shown by gas balance, with respiratory quotient (RQ) and pH corrector liquid (see figure 1).

Then, there are physical phenomenon, which are associated to real reactors. These mechanisms can be decomposed in many categories : transfer phenomenon (mass, thermal and movement quantity), regulation (realised by an operator), introduction of products, and mixing. These mechanisms interfere with biology and it is significant to notice that relaxation times of these phenomena are of size order of response time of biolog-

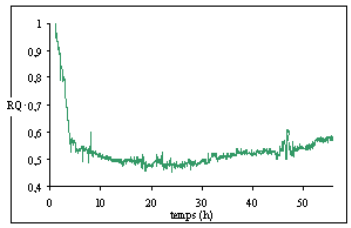

Fig. 1. An example of respiratory quotient evolution during a culture. x-axis is the time of the experience, y-axis is the amplitude of the signal

ical response. With all these phenomena, a variable can be described by the following equation (see (Roels, 1983)) :

$$\frac{dV}{dt} = \Delta.\left(\frac{V_{equilibrium} - V(t)}{\tau_{physical}}\right) + r_{V(t)} + \Phi_{V(t)} \quad (1)$$

where:

- $\frac{dV}{dt}$ corresponds to the dynamic of the system.
- $\Delta.\left(\frac{V_{equilibrium}-V(t)}{\tau_{physical}}\right)$ is variable variation between biological and physical parameters.$\tau_{physical}$ is the time constant of physical phenomena; this constant can not be characterised because it depends on reaction progress.
- $r_{V(t)}$ is the volumic density of reaction of the variable V, it is a biological term.
- $\Phi_{V(t)}$ corresponds to an external intervention which results of a voluntary action.

Moreover, it is essential to observe that there is a regulation loop between biology and physic (see figure 2). The problematic is, from measures, to isolate or eliminate perturbations. These responses depend on physical phenomena or human interventions (process regulation). It is to quantify biological kinetics and by this way to optimise biological kinetics and control that is to say identify modifications of the biological behaviour. For example, in the case of yeast production, it is important to maintain an oxidative metabolism by the control of glucose residual concentration, fermentative metabolism is prejudicial to the yield. The aim is to maintain an optimal production to avoid the diminution of substrate conversion yield, that is to say to remark the biological change between oxidative and fermentative metabolism. We propose to use an unsupervised method using singularities and correlation in order to detect and characterize all the phenomena (biological, regulation, etc.) occurring in a bioprocess.

3. RELATED WORK AND MOTIVATION

Several works using various approaches lead independently to each others, to the conclusion that

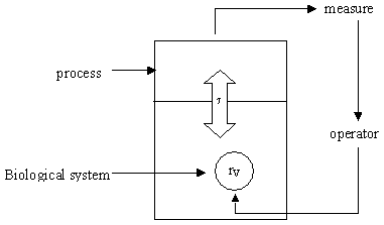

Fig. 2. Interactions between the biological system, the process and the operator.

the limits of a states are linked to the singularities of biochemical signals: for instance, Steyer et al. (Steyer *et al.*, 1991) (using expert system and fuzzy logic), Bakshi and Stephanopoulos (Bakshi and Stephanopoulos, 1994) (using expert system and wavelets) and Doncescu et al. (Doncescu *et al.*, 2002) (using inductive logic) show that the beginning and the end of a state correspond to singularities of the biochemical signals measured during the process.
In a fed-batch bioprocess, a physiological state can occur several times during the experience. After the detection of states, it is then necessary to characterize these states. The caracterisation is often based on the statistical properties of the biochemical signals. Classification Methods based on Principal Components Analysis (PCA) (Ruiz *et al.*, 2004), adaptive PCA (Lennox and Rosen, 2002), and kernel PCA (Lee *et al.*, 2004) enables to distinguish and characterize the different states.
For the boundaries of the states, we propose to use the Maximum of Modulus of Wavelets Transform (Mallat and Zhong, 1992)(Mallat and Hwang, 1992) to detect the singularities of the signals. The singularities are selected according to their Hölder exponent evaluation. The characterisation of the states is based on the correlation product between the signals on intervalls whose boundaries are the selected singularities.

4. THE UNSUPERVISED CLUSTERING METHOD IN DETAILS

4.1 Detection and Selection of singularities by wavelets and Hölder exponent

The singularities of the biochemical signal correspond to the boundaries of the states. These signals are non-stationnary and non-symmetrical signals; they are not chirps and have no infinite oscillations (see figure **??**). Several authors have proposed to use wavelets to detect the singularities of the signals for the detection of states: Bakshi and Stephanopoulos (Bakshi and Stephanopoulos, 1994) and more recently Jiang et al. (Jiang *et al.*, 2003). Besides singularities correspond to maxima of modulus of wavelets coefficients. The

wavelets are a powerful mathematical tool of non stationarity signal analysis. Wavelets are very used in images analysis and compression, but they know an increasing success in all data processing. Wavelets Transformation (WT) is a rather simple mechanism to decompose a function into a set of coefficients depending on scale and location. The definition of the wavelets transform is:

$$W_{s,u}f(x) = (f \star \psi_{s,u})(x) = \int f(x)\psi(\frac{x-u}{s})dx \quad (2)$$

where ψ is the wavelet, f, is the signal, $s \in R^{+*}$ is the scale (or resolution) parameter and $u \in R$ is the translation parameter. The scale plays the role of frequency. The choice of the ψ wavelet is a complicate task.

A wavelet is a function $\psi(t)$ with a zero average:

$$\int \psi(t)dt = 0 \quad (3)$$

The wavelet is translated and dilated

$$\psi_{u,s} = \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s}) \quad (4)$$

allowing it to convolve the analyzed signal which different size of "window" wavelet function.For the detection of the singularities and inflexion points of the biochemical signal, we use the Maxima of Modulus of Wavelets Transform (Mallat and Zhong, 1992)(Mallat and Hwang, 1992). The idea is to follow the local maxima at different scales and the most important will be propagated from low frequencies to high frequencies. These maxima correspond to singularities, particularly when the walevet is a derivative of a smooth function:

$$\psi(x) = \frac{d\theta(x)}{dx}$$

$$W_{s,u}f(x) = f * \psi_{s,u} = f(x) * \frac{d\theta(x/s)}{dx}$$

Yuille and Poggio (Yuille and Poggio, 1986) have shown that if the wavelet is a derivative of gaussian, then the maxima belong to connexe curves which are neverbroken from a scale to other. The detection of the singularities of the signal is thus possible by using the wavelets (see for example figure 3).

Jiang et al. (Jiang *et al.*, 2003) have proposed to select the maxima by using thresholding. Besides, all the singularities are not relevant et only somes singularities are meaningful. However the thresholds proposed by Jiang et al. are chosen empirically. To select the meaningful singularities, we proposed to use the Hölder exponent. The Hölder exponent is a mathematical value which enables to characterize the singularities. The fractal dimension enables also to characterize singularities but only the Hölder exponent can characterize

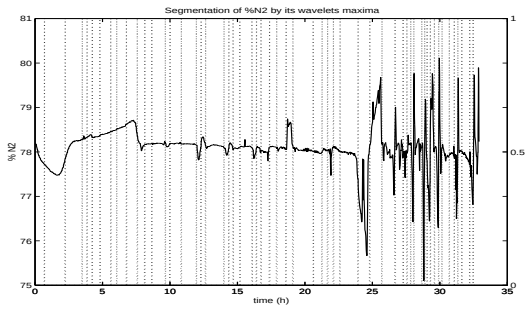

Fig. 3. Segmentation of N2 (nitrogen). Each vertical dotted line correspond to a singularity of the signal detected by wavelets. The wavelet is a DOG (first derivative of gaussian) and the scales go from 2^0 to 2^3.x-axis is the time of the experience, y-axis is the amplitude of the signal.

locally each singularity. A singularity in a point x_0 is characterized by the Hölder exponent (also called Hölder coefficient or Lipschitz exponent). This exponent is defined like the most important exponent α allowing to verify the next inequality:

$$|f(x) - P_n(x - x_0)| \leq C|x - x_0|^\alpha \quad (5)$$

We must remark that $P_n(x - x_0)$ is the Taylor Development and basically $n \leq \alpha(x_0) < n + 1$. The Hölder exponent could be extended to the distribution. For example the Hölder exponent of a Dirac is -1. A fast computing leads to a very interesting result of the Wavelets Transform (Jaffard, 1997):

$$|W_{s,u}f(x)| \simeq s^{\alpha(x_0)} \quad (6)$$

This relation is remarkable because it allows to measure the Hölder exponent using the behavior of the Wavelets Transform. Therefore, at a given scale $a = 2^N$ the $W_{a,b}f(x)$ will be maximum in the neighborhood of the signal singularities. The detection of the Hölder is linked to the vanishing moment of the wavelet: if n is the vanishing moment of the wavelet, then it can detect Hölder coefficient $\leq n$ (Mallat and Hwang, 1992). We use a wavelet (DOG: first derivative of gaussian) with a vanishing moment equal to 1; consequently we can only detect Hölder coefficient smaller than 1. This is not a real problem because we are interesting by the singularities as step or dirac and the Hölder coefficient of these singularities are smaller than 1. Besides, the meaningful singularities of the fed-batch bioprocess have Hölder exponent smaller than 1 which correspond to sharp singularities. This is this kind of sharp variation which is meaningful for the fed-batch bioprocess fermentation because of many external regulation of the process. Moreover for Hölder coefficient greater than 1 particularly for integer values, there are difficulties to interpret the Hölder coefficient (see (Meyer, 1990) cited in (Mallat and

Zhong, 1992)). To evaluate the Hölder coefficient from the wavelets, there is two main ways:

(1) the graphical method. It consists in finding the maximum line i.e. the maximum which propagates through the scales, and compute the slopes of this maximum line (often with a log-log representation). The computed slope corresponds to the Hölder coefficient (Mallat and Hwang, 1992).
(2) the minimisation method. It consists in minimizing a function whose one of the parameter to evaluate is the Hölder coefficient (Mallat and Zhong, 1992). The function is the following:

$$\sum_j \left(ln_2(|s_j|) - ln_2(C) - j - \frac{\alpha(x_0)-1}{2} ln_2(\sigma^2 + 2^{2j}) \right)^2 \quad (7)$$

where s_j represents the maximum at scale j, C is a constant depending on the singularity localised in x_0, σ is the standard deviation of an approximate gaussian of the singularity (see (Mallat and Zhong, 1992)), and $\alpha(x_0)$ the Hölder exponent.

The graphical method is the most fast and the most used method, but the evaluation of the Hölder coefficient is sometimes imprecise as noted in (Struzik, 1999)(Nugraha and Langi, 2002). For the second method, *a priori*, all methods of minimisation can be used for the evaluation. In (Mallat and Zhong, 1992), a gradient descent algorithm is proposed to resolved the minimisation, but this technique is very sensitive to local minima. More recently, a minimisation using Genetical Algorithms has been proposed (Manyri *et al.*, 2003) and used in bioprocess (Régis *et al.*, 2004). More precisely it uses Differential Evolutionary (DE) algorithms. The DE algorithms was introduced by Rainer Storn and Kenneth Price (Storn and Price, 1996). We use this method of evaluation by DE to compute the Hölder exponents for the selection of singularities of the biochemical signals.

4.2 Characterisation by correlation product and classification

Once the states are bounded by the detected and selected singularities using the wavelets, they are characterized by the analysis of the correlations between the biochemical signals. Traditionnally, the characterization of the states can be made by the calculus of the distance between the different values of the measured biochemical parameters and the prototypes of the different classes in supervised cases, or by if-then rules. The if-then rules describe the relations between the biochemical parameters with the point of view of an expert (Steyer *et al.*, 1991)(Steyer, 1991). We assume that if-then rules can be implicitly replaced by correlations between the biochemical parameters which represent the relations between the biochemical parameters with the point of view of statistical context. On each intervall defined by the singularities, a product of correlation is computed bewteen the signals two by two. The correlation coefficient (also called Bravais-Pearson coefficient, see (Saporta, 1990)) is given by the equation:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sigma_x \sigma_y} \quad (8)$$

where x_i represent the values of one parameter (in a given intervall), y_i the values of the second parameter (in the same intervall), n the number of elements, $\overline{x}$ the average of the elements x (of the first biochemical signal), $\overline{y}$ the average of the elements y (of the second biochemical signal), et σ_x et σ_y the standard deviation of each of the two signals.
The correlation coefficient is equivalent to the cosinus of the scalar product between two biochemical signals projected in the correlation circle of a PCA realized between the two biochemical signals. On each intervall, the sign of each correlation coefficient between two signals is kept. Each intervall is thus characterized by a set of positive of negativ signs. The intervalls with the same set of signs are put in the same class as illustred in the figure 5. Ruiz et al. (Ruiz *et al.*, 2004) pro-

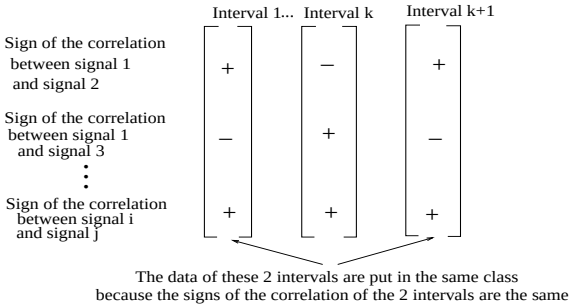

Fig. 4. Principle of the classification method based on wavelets, Hölder exponent and correlation coefficient

pose a classification method based on PCA for a neighbouring application (wastewater treatment): the data are projected in the space generated by the two first principal components. The method enables to reduce the size of the space of data and to take account of the correlation of the signals. However the PCA doesn't take account of the time: the temporal evolution of the process is not taken into account. Ruiz et al. propose to use a window of time analysis of fixed size. But as the window has a fixed size, it doesn't really take account of the changs occuring during the bioprocess. So the method proposed in this article

seems to be more adapted if it is necassary to take account of the variation of the process.

5. EXPERIMENTAL RESULTS

Tests have been made on a fed-batch fermentation bioprocess. This bioprocess is a biotechnological process using yeats called *Saccharomyces Cerevisiae* during about 34 hours. 11 biochemical signals have been measued during the bioprocess. The maximum scale is chosen empirically. Each signals has 2448 samples. Mallat and Zhong (Mallat and Zhong, 1992) propose to use as maximal scale $log_2(N)+1$ where N is the number of measured samples of the signals. However if we use this maximal scale, several singularities would be removed. The maximal scale is then chosen with an expert in microbiology.
The classification providen by the method gives interesting results. Particularly, the most interesting result concerns the detection and the caracterisation of a state resulting of an extern action. Besides, the class number 8 corresponds to the addition of an acid [1] in the bioprocess. All the apparition of class 8 correspond exactly to a the addition of acid. These results were confirmed and valited. As far as we know, it is the first time that this kind of non-model-based can find characterize automatically the addition of acid in a fed-bacth process. The results are promising and further analysis of the classification is necessary.

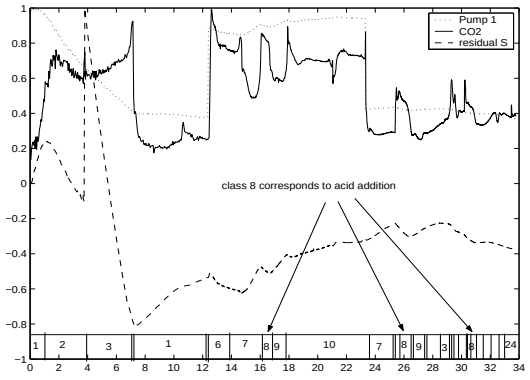

Fig. 5. Classification providen by the method. The wavelet is a DOG and the scales go from 2^0 to 2^{10}.x-axis is the time of the experience, y-axis is the amplitude of the signal. Above the x-axis, some of the most important classes are given, particularly the class number 8.

6. CONCLUSION AND FURTHER WORK

In this paper, we have presented a method of classification based on wavelets, Hölder exponent and coefficient correlation for the detection and the characterisation of states in bioprocess. The states detection is based on the detection and the selection of singularities of the biochemical signals using the Maximum of Modulus of Wavelets Transform and the evaluation of Hölder exponent. The states characterisation is based on coefficient correlation between signals. Further work include tests with other kinds of wavelets. The use of the values of the coefficient correlation instead of the sign for the characterisation is a way to explore. The next step is an on-line (real time) classification of the bioprocess. This approach is generic and sould also be used for all others data mining application using multiple time series: medical data, genetic data, econometry, etc.

[1] because of industrial confidentiality, we are not allowed to give more information

ACKNOWLEDGEMENTS

We want to thank the LBB of INSA-Toulouse for their help and collaboration. Special thanks to M. Jacky Desachy for his help on the field of clustering. This work has been partially supported by the research office of the Region Guadeloupe, French West Indies.

REFERENCES

Aguilar-Martin, J., J. Waissman-Vilanova, R. Sarrate-Estruch and B. Dahou (1999). Knowledge based measurement fusion in bio-reactors. In: *IEEE EMTECH*.

Bakshi, B.R. and G. Stephanopoulos (1994). Representation of process trends-III. multiscale extraction of trends from process data. *Computer and Chemical Engineering* **18**(4), 267–302.

Doncescu, A., J. Waissman, G. Richard and G. Roux (2002). Characterization of biochemical signals by inductive logic programming. *Knowledge-Based Systems* **15**(1-2), 129–137.

Jaffard, S. (1997). Multifractal formalism for functions part 1 and 2. *SIAM J. of Math. Analysis* **28**(4), 944–998.

Jiang, T., B. Chen, X. He and P. Stuart (2003). Application of steady-state detection method based on wavelet transform. *Computer and Chemical Engineering* **27**(4), 569–578.

Lee, J.-M., C. Yoo, I.-B. Lee and P. Vanrolleghem (2004). Multivariate statistical monitoring of nonlinear biological processes using kernel PCA. In: *IFAC CAB'9*. Nancy, France.

Lennox, J. and C. Rosen (2002). Adaptative multiscale principal components analysis for online monitoring of wastewater treatment. *Water Science and Technology* **45**(4-5), 227–235.

Mallat, S. and S. Zhong (1992). Characterization of signals from multiscale edges. *IEEE Trans. on PAMI* **14**(7), 710–732.

Mallat, S. and W.-L. Hwang (1992). Singularity detection and processing with wavelets. *IEEE Trans. on Information Theory* **38**(2), 617–643.

Manyri, L., S. Regis, A. Doncescu, J. Desachy and JL Urribelarea (2003). Holder coefficient estimation by differential evolutionary algorithms for saccharomyces cerivisiae physiological states characterisation. In: *ICPP-HPSECA*. Kaohsiung, Taiwan.

Meyer, Y. (1990). *Ondelettes et Opérateurs*. Vol. I. Hermann.

Nugraha, H. B. and A. Z. R. Langi (2002). A wavelet-based measurement of fractal dimensions of a 1-d signal. In: *IEEE APCCAS*. Bali, Indonesia.

Régis, S., L. Faure, A. Doncescu, J.-L. Uribelarrea, L. Manyri and J. Aguilar-Martin (2004). Adaptive physiological states classification in fed-batch fermentation process. In: *IFAC CAB'9*. Nancy, France.

Roels, J.A. (1983). *Energetics and kinetics in biotechnology*. Elsevier Biomedical Press.

Ruiz, G., M. Castellano, W. Gonzàlez, E. Roca and J.M. Lema (2004). Algorithm for steady states detection of multivariate process: application to wastewater anaerobic digestion process. In: *AutMoNet 2004*. pp. 181–188.

Saporta, G. (1990). *Probabilités, et Analyse des données et Statistique*. Technip.

Steyer, J.P. (1991). Sur une approche qualitative des systèmes physiques : aide en temps rèel à la conduite des procédés fermentaires. Thse de Doctorat. Université Paul Sabatier. Toulouse France.

Steyer, J.P., J.B. Pourciel, D. Simoes and J.L. Uribelarrea (1991). Qualitative knowledge modeling used in a real time expert system for biotechnological process control. In: *IMACS International Workshop "Decision Support Systems and Qualitative Reasoning"*.

Storn, R. and K. Price (1996). Minimizing the real functions of the icec'96 contest by differential evolution. In: *Proc. of the 1996 IEEE International Conference on Evolutionary Computation*.

Struzik, Z. R. (1999). *Fractals: Theory and Application in Engineering*. pp. 93–112. Springer Verlag.

Yuille, A. and T. Poggio (1986). Scaling theorems for zero-crossing. *IEEE Transaction for zero-crossing* **8**(1), 15–25.