# IDENTIFICATION OF YIELD COEFFICIENTS IN AN *E. COLI* MODEL - AN OPTIMAL EXPERIMENTAL DESIGN USING GENETIC ALGORITHMS

**Ana C. A. Veloso[1,2], I. Rocha[1] and E. C. Ferreira[1]**

[1]*Centro de Engenharia Biológica, Universidade do Minho*
*4710-057 Braga, PORTUGAL*
[2]*Escola Superior Agrária de Bragança, Campus de Santa Apolónia, Apartado 172*
*5301-855 Bragança, PORTUGAL*

*e-mail: {anaveloso, irocha, ecferreira}@deb.uminho.pt*

Abstract: An optimal experimental design for yield coefficients estimation in an unstructured growth model of fed-batch fermentation of *E. coli* is presented. The feed profile is designed by optimisation of a scalar function based on the *Fischer Information Matrix*. A genetic algorithm is proposed as the optimisation method due to its efficiency and independence on the initial values. *Copyright 2004 IFAC*

Key-words: Optimisation, genetic algorithms, Fisher Information Matrix, fed-batch fermentation, *E. coli*, yield coefficients.

## 1. INTRODUCTION

The simulation of bioprocesses has always been of major academic and industrial interest, being an effective tool for the design and development of robust (model based) algorithms for optimisation, monitoring, control, and characterization of many industrial processes. The development of mathematical models that are able to describe these processes has become an essential task since it is usually much faster and less expensive to model a system and to simulate its operation than to perform laboratory experiments (Versyck and Van Impe, 1999).

The development of a model is usually an iterative two-steps process: designing the model structure and evaluating the model parameters for the proposed structure from simulated and experimental data. The description of bioprocesses usually requires the use of differential and algebraic equations involving frequently highly non-linear and stiff models (Banga *et al.*, 2002).

In the experiment design, a set of experimental conditions, namely the measurement ports, the sampling times, the filters used before sampling, and the input signal, must be selected in order to maximize the information regarding the properties of the system that are pertinent to a particular application. In fact, each of those conditions has a bearing on the information obtained from the experiment (Goodwin, 1987).

Therefore, the use of model-based experimental design may give suitable suggestions for efficient and informative experiments. Hence, it is necessary to establish a mathematical function that allows the calculation of the efficiency of an experiment with regard to the experimental aim. Throughout the optimisation procedure, several experimental conditions are evaluated and the experimental set-up leading to the maximum (or minimum) value of the selected objective function represents the optimal experimental design.

Among the objective functions used in literature, several functions based on the *Fisher Information*

*Matrix* (*FIM*) have been proposed to evaluate the parameter estimation accuracy. In fact, *FIM* contains information concerning parameter sensitivities and measurement errors and, thus, allows the quantification of the parameter estimation quality.

In this work, a fed-batch fermentation process of *Escherichia coli* will be studied from a modelling point of view. *E. coli* is usually grown under that mode of operation due to the well-known negative effect of acetate, which is produced when the substrate, glucose, is present above certain concentration (Versyck and Van Impe, 1999; Rocha and Ferreira, 2002a,b).

So, the aim of the present study is to calculate an optimal feeding rate that maximizes the determinant of *FIM* (*D-criterion*) and conducts to the optimal experimental design for the identification of yield coefficients of a fed-batch fermentation of *E. coli*. Results of various optimisation runs using genetic algorithms are presented and discussed.

## 2. PROCESS MODELLING

Process simulation was conducted with a developed model derived from the general state space dynamical model described by Bastin and Dochain (1990). Accordingly, the dynamics of a reaction network in a stirred tank bioreactor can be described by the following mass balance equations written in matrix form as:

$$\frac{d\xi}{dt} = Kr(\xi,t) - D\xi + F - Q \qquad (1)$$

in which $\xi$ is a vector representing the $n$ state components concentrations ($\xi \in \Re^n$), $r$ is the growth rate vector corresponding to $m$ reactions ($r \in \Re^m$), $K$ is the matrix of yield coefficients ($K \in \Re^{n \times m}$), $F$ is the vector of feed rates and $Q$ is the vector of gaseous outflow rates ($F$, $Q \in \Re^n$), $D$ is the dilution rate (being $D^{-1}$ the residence time).

The *E. coli* fermentation process is accepted as occurring in two possible metabolic regimens (Rocha and Ferreira, 2002a): (i) a respirative-fermentative regime (*RF*), corresponding to an acetate production state, (ii) a respirative regime (*R*), corresponding to a simultaneously acetate and glucose consumption pathway.

The associated dynamical model can be represented as follows, where *S*, *O*, *X*, *C*, and *Ac* represent sugar (glucose), dissolved oxygen, biomass, dissolved carbon dioxide, and acetate concentrations, respectively; $\mu_1$, $\mu_2$, and $\mu_3$ are the specific growth rates; $k_i$ are the yield (stoichiometric) coefficients; $F_{in}$ and $S_{in}$ are the substrate feed rate and the influent glucose concentration, respectively; $W$ is the culture medium weight. *CTR* is the carbon dioxide transfer rate from liquid to gas phase, and *OTR* is the oxygen transfer rate from gas to liquid phase.

$$\frac{d}{dt}\begin{bmatrix} X \\ S \\ Ac \\ O \\ C \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ -k_1 & -k_2 & 0 \\ 0 & k_3 & -k_4 \\ -k_5 & -k_6 & -k_7 \\ k_8 & k_9 & k_{10} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} X - D \begin{bmatrix} X \\ S \\ Ac \\ O \\ C \end{bmatrix} + \begin{bmatrix} 0 \\ \left(\frac{F_{in}}{W}\right) S_{in} \\ 0 \\ OTR \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ CTR \end{bmatrix} \qquad (2)$$

An additional equation is used to calculate the variation of the culture medium weight with the time:

$$\frac{dW}{dt} = F \qquad (3)$$

## 3. OPTIMAL EXPERIMENTAL DESIGN

The successful application of optimal experimental design in fed-batch bioreactors has been reported by several authors (Munack, 1989; Ejiofor *et al.,* 1994; Versyck and Van Impe, 1999).

Parameter estimation can be formulated as the minimization of the following *identification function* by optimal selection of the parameter vector $k$:

$$J_I(k) \equiv \int_0^{t_f} (y(k) - y_m)^T P (y(k) - y_m) dt \qquad (4)$$

in which $y_m$ is the vector of measured outputs, *y(k)* is the vector of model predictions by using the parameter vector $k$, $P$ is a user-supplied square weighting matrix and $t_f$ is the final experimental time. In order to analyse the information content of the state trajectories obtained in a certain experiment, *FIM* can be used (Munack, 1989; Versyck and Van Impe, 1999):

$$FIM \equiv \frac{1}{N} \int_0^{t_f} \left(\frac{\partial y}{\partial k}(t)\right)^T P \left(\frac{\partial y}{\partial k}(t)\right) dt \qquad (5)$$

where $N$ is the experimental number of data used. Weighting matrix $P$ is usually chosen as the inverse of the measurement error covariance matrix, implying that the more a measurement is corrupted by noise, the less it will count in the information criterion. The *FIM* is the inverse of the error covariance matrix of the Best Linear Unbiased Estimator (BLUE). It contains information about the measurement errors and parameter sensitivities and, thus, allows the quantification of the quality of the parameter estimation. Depending on the requirements imposed by the application, a specific scalar of this *FIM* is used as the performance index for optimal experimental design to increase the parameter identifiability. Several optimal design criteria are

discussed in literature (Walter and Pronzato, 1990, Versyck and Van Impe, 1999). In this work the so-called *D-criterion* for optimal experimental design was adopted:

$$J_{II}(t_f) = \det(FIM) \tag{6}$$

where det(*FIM*) is the determinant of the *Fisher Information Matrix*. In order to optimize the global accuracy, the determinant of the *FIM* must be maximized. This is equivalent to minimizing the geometric mean of estimation error and the volume of the uncertainty ellipsoids.

For chemical and biological systems, it is known that model identification is not an easy task. There are two major difficulties: (i) the estimation of the yield coefficients values of matrix *K*; and (ii) the determination of a suitable structure for the reaction rate model $r(\xi)$ and the estimation of the respective kinetic coefficients (Chen, 1992).

In this work, the approach proposed by Chen (1992) is chosen. This approach is based on the fact that the identification of the yield coefficients can be decoupled from that of the reaction kinetics. In fact, it is possible to identify the yield coefficients in a first step without modelling the reaction rates, followed by the modelling of the reaction rates and the identification of the related kinetic coefficients in a second step, using known yield coefficients. It is interesting to notice that these two sets of parameters are involved in the model in different ways: the model is linearly parameterized by the yield coefficients while it is in general nonlinearly parameterized by the kinetic coefficients. The method of Chen (1992) makes use of the state transformations based on the general structure of the model and it has already been used by Rocha and Ferreira (1996) to identify yield coefficients in the production of baker's yeast.

The methodology proposed by Chen (1992) uses the same model structure represented by eq. (1) modified by introducing a *U* vector that corresponds to the *F-Q* term:

$$\frac{d\xi}{dt} = Kr(\xi,t) - D\xi + U \tag{7}$$

Defining a non-singular partition:

$$\begin{bmatrix} K_a \\ K_b \end{bmatrix} = LK \tag{8}$$

where *L* is a square matrix ($L \in \Re^n$) and $K_a$ has full rank. Being *p* the rank of the yield coefficient matrix *K*, then $K_a \in \Re^{p\times n}$, $K_b \in \Re^{(n-p)\times m}$. The induced partition of $\xi$ and *U are:*

$$\begin{bmatrix} \xi_a \\ \xi_b \end{bmatrix} = L\xi, \quad \begin{bmatrix} U_a \\ U_b \end{bmatrix} = LU \tag{9}$$

with $\xi_a$, $U_a \in \Re^p$ and $\xi_b$, $U_b \in \Re^{n-p}$.

The following state transformation is introduced:

$$Z \equiv A\xi_a + \xi_b \tag{10}$$

where the $(n\text{-}p)\times p$ matrix *A* is the unique solution of the following equation:

$$AK_a + K_b = 0 \tag{11}$$

i.e. $A = -K_b K_a^+$ (12)

where $K_a^+$ represents a generalised or pseudo-inverse of $K_a$. Using eqs. (10) to (12), the eq. (7) yields the following:

$$\frac{d\xi_a}{dt} = K_a r(\xi_a, Z - A\xi_a) - D\xi_a + U_a \tag{13}$$

$$\frac{dZ}{dt} = -DZ + AU_a + U_b \tag{14}$$

As it can be seen, eq. (14) does not involve the reaction rate *r* explicitly, and so, it can be used to estimate the yield coefficients via the identification of the matrix *A*, independently of the structure of the reaction rates, as suggested by Chen (1992). However, in eq. (14) there is a dependence of *Z* with respect to matrix *A* (and therefore depending on the yield coefficients). If $U_a$ is a zero vector, the integration of the above-mentioned equation may be used for the estimation of the yield coefficients. Alternatively *Z* can be decomposed as follows:

$$Z \equiv AZ_a + Z_b \tag{15}$$

where the new variables $Z_a \in \Re^p$ and $Z_b \in \Re^{n-p}$ are governed by the following dynamics, named by Chen (1992) as the *auxiliary model*:

$$\frac{dZ_a}{dt} = -DZ_a + U_a \tag{16}$$

$$\frac{dZ_b}{dt} = -DZ_b + U_b \tag{17}$$

$$\xi_b = Z_b - A(\xi_a - Z_a) \tag{18}$$

The equation for $\xi_b$, which may be understood as an output vector, is obtained substituting eq. (10) in eq. (15). Defining the output vector *y* and the regressor vector $\phi$ as follows:

$$y \equiv Z_b - \xi_b \tag{19}$$

$$\phi = \xi_a - Z_a \tag{20}$$

eq. (18) takes the following standard linear regression form:

$$y(t) = A\phi(t) \tag{21}$$

with $t \in [0, t_f]$, where $t_f$ is the total time of the experiment. This auxiliary model, containing only the transport dynamics of the system, can be considered as a linear time varying model with state $Z$, input $U_a$, $U_b$ and $\xi_a$, and output $\xi_b$. The model is nonlinear relatively to the yield coefficients but linear in order to the elements of matrix $A$. Furthermore, the unknown parameters are only involved in the output eq. (18). The regression equation (eq. 21) is the basis for the estimation of the yield coefficients.

Difficulties related to the identifiability of the yield coefficients were extensively studied by Chen (1992). This author established that the properties of identifiability of the yield coefficients, for a given system characterised by matrix $K$, are independent of the partition chosen and demonstrated the necessary condition for the identification of those coefficients for a given reaction (elements of a column of $K$) right through matrix $A$: *the yield coefficients of reaction j are identifiable through A only if* $n_j\text{-}1 \leq n\text{-}p$ *being* $n_j$ *the number of components involved in reaction j, p the rank of matrix K, n the number of state variables*. So, $n\text{-}p$ coefficients can be identified for each reaction without any knowledge concerning the reaction rate $r(\xi, t)$.

With the methodology proposed by Chen (1992), the accuracy of the yield coefficients estimation is obtained using the regressor analysis, resulting in following equivalent *FIM*:

$$FIM \equiv \frac{1}{N}\int_0^{t_f} \phi(t) P \phi^T(t) dt \tag{22}$$

$$\text{where } \left(\frac{\partial y}{\partial k}\right) = \phi^T(t) \tag{23}$$

Taking into account the above-mentioned formulation it is necessary to determine the regressor trajectories, $\phi(t, \theta, \xi, U)$, which are functions of the parameter vector to be estimated for the calculation of the yield coefficients.

In this work, for each regimen the following state partition was chosen: $\xi_a^T = [S \quad O]$ and $\xi_b^T = [X \quad Ac \quad C]$. In this case, the induced partitions for $U_a$ and $U_b$ are the same for both regimens:

$U_a^T = \left[\left(\frac{F_{in}}{W}\right) S_{in} \quad OTR\right]$ $\quad U_b^T = [0 \quad 0 \quad -CTR]$.

For $K_a$ and $K_b$ the partitions are given in Table 1.

Table 1 Induced partitions for $K$

| Regimen | $K_a$ | $K_b$ |
|---|---|---|
| *RF* | $\begin{bmatrix} -k_1 & -k_2 \\ -k_5 & -k_6 \end{bmatrix}$ | $\begin{bmatrix} 1 & 1 \\ 0 & k_3 \\ k_8 & k_9 \end{bmatrix}$ |
| *R* | $\begin{bmatrix} -k_1 & 0 \\ -k_5 & -k_7 \end{bmatrix}$ | $\begin{bmatrix} 1 & 1 \\ 0 & -k_4 \\ k_8 & k_{10} \end{bmatrix}$ |

## 4. OPTIMISATION USING GENETIC ALGORITHMS

Genetic algorithms (GAs) are stochastic algorithms that can be an alternative to traditional optimisation methods, being suitable for complex non-linear models where finding the global optimum is a difficult task. They are based on the mechanisms of natural selection and genetics followed by biological evolving species. First, an initial population, containing a predefined number of individuals (or solutions), is created randomly. The potential solution is coded as a vector called chromosome representing a possible solution in the multidimensional search space. Goodness of each solution in the population is evaluated by using a predefined fitness criterion. Upon fitness assessment of all chromosomes in the population, a new generation of individuals is created from the actual population, by using three genetic operators: reproduction, crossover and mutation (Roubos *et al.*, 1999).

In recent years, GAs have been applied to fed-batch fermentation optimisations (Roubos *et al.*, 1999; Nguang *et al.*, 2001; Na *et al.*, 2002; Rocha and Ferreira, 2002b). In the present case, GAs were applied to maximize the global accuracy of the parameter estimation considering two different optimisation objectives: (i) estimate the optimal substrate feed rate trajectory ($F_{in}$); and (ii) optimise $F_{in}$ and the influent glucose concentration ($S_{in}$). Therefore, the following objective function was used:

$$J(t_f) = \det(FIM_{RF}) \times \det(FIM_R) \tag{24}$$

Eq. (24) considers the two regimens that may occur during the *E. coli* fed-batch fermentation, allowing maximizing the informative richness for each of the regimens and so, to determine with a higher accuracy the yield coefficients related to each regimen. For this purpose, the *Genetic and Evolutionary Algorithm Toolbox* (*GEAtbx 3.3*) for MATLAB developed by Pohleim (2003) was used. It works with several genetic operators and supports binary, integer and real-valued representations, being the last one chosen in this work due to the well-known advantages (Roubos *et al.*, 1999). A main function is called from

the user interface. This function calls all necessary evolutionary operators and the objective function. Additionally, the main function performs nearly all the data management and result collection (Pohleim, 2003). In this case, the number of individuals evaluated in each iteration was 100 corresponding to one population. The objective function was evaluated in a script '*.m*' file, a routine that calculated the value of eq. (24) in an iterative way using the MATLAB version 6.5 subroutine ODE23s to solve the differential equations of the model represented in eq. (2). A penalty function was used to implement the maximal culture medium weight constraint (if $W(t) >$ 5 kg, $J(t_f) = 0$). Regarding the feed rate profile, a purely numerical approach was used, being the feed rate divided in 11 nodes corresponding to 10 piecewise linear polynomials with constant time intervals. Moreover, the feed rate is limited by the pump capacity being 0,4 kg/h the maximum feed rate. Initial values for *S*, *X*, *Ac*, *C* and *W* were 0 g/kg, 5 g/kg, 0 g/kg, 0.3 g/kg and 3 kg, respectively.

Two optimisation strategies were studied: *SGA1* searches the optimal substrate feed rate $(F_{in})$ and *SGA2* searches the optimal $F_{in}$ and the optimal influent glucose concentration $(S_{in})$. Table 2 presents the constraints taken into account in each case.

Table 2 Constraints used in the optimisation

| | $t_f$ (h) | $S_{in}$ (g/kg) |
|---|---|---|
| *SGA1* | 25 | 250 |
| *SGA2* | 25 | [200, 300] |

Figure 1 shows the convergence of the objective function towards the optimal solution with the number of iterations performed for each strategy studied.

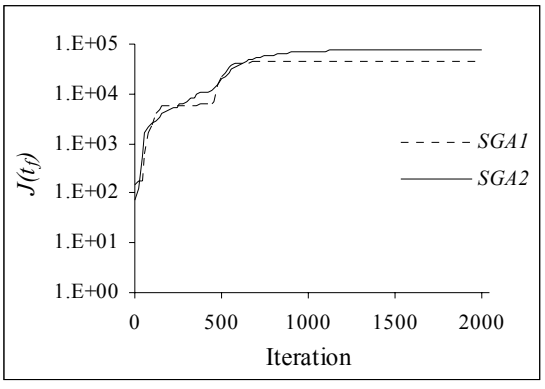


Fig. 1. Evaluation of the objective function with the number of iterations performed.

The results obtained show that the information content increases with the optimisation of $S_{in}$. Therefore, it seems that the optimisation of other conditions, namely the final experimental time and the initial biomass content, may allow the improvement of the information content. The optimal feed rates profiles obtained are shown in Figure 2. It should be remarked that for *SGA2* the optimal influent glucose concentration obtained was 204 g/kg.

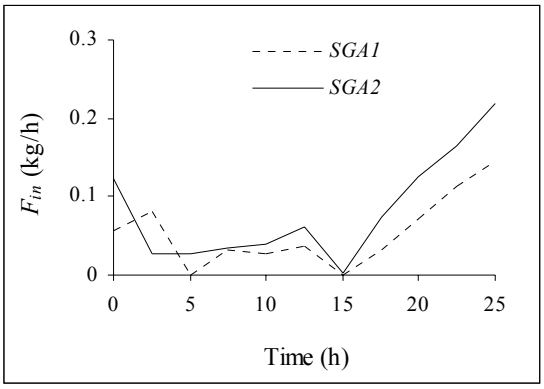


Fig. 2. Optimised feed rate profiles.

The optimised values obtained for each strategy were used to validate the estimation of the yield coefficients. The validation is accomplished by comparing the *X*, *Ac* and *C* estimated values calculated using eq. (21) with the values obtained from the simulation using eq. (2). The results obtained for *SGA2* (Fig. 3) show that both regimens occur with similar duration time periods allowing a satisfactory calculation of the regressors as illustrated by the good accordance between simulated and estimated values. Similar results were obtained for *SGA1*.

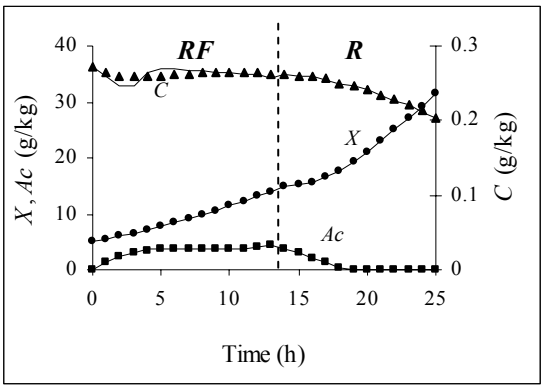


Fig. 3. Validation of the yield coefficients estimation for *SGA2*. Symbols represent simulated values and lines correspond to the estimated values.

## 5. CONCLUSIONS

The optimal experimental design of yield coefficient of an *E. coli* fed-batch fermentation was accomplished by the quantification of the richness given by the *Fisher Information Matrix*. The results

obtained show that the identification problem is overcome by introducing a model transformation that allows a linear formulation without knowing the kinetic parameters (Chen, 1992). Moreover, the genetic algorithms showed to be an efficient tool for the optimisation of highly non-linear models. The advantage of the identification technique presented is the possibility of obtaining the feed rates profiles before making any experimental work. The experimental validation of this work is under investigation.

## ACKNOWLEDGEMENTS

A.C.A. Veloso and I. Rocha are most grateful for the financial support provided by PRODEP and Fundação para a Ciência e a Tecnologia (PRAXIS XXI/BD/16961/98), respectively.

## REFERENCES

Banga, J.R., Versyck, K.J. and Van Impe, J.F. (2002). Computation of Optimal Identification Experiments for Nonlinear Dynamic Process Models: a Stochastic Global Optimization Approach. *Ind. Eng. Chem. Res*., **41**, 2425-2430.

Bastin, G., and D. Dochain (1990). *On-line Estimation and Adaptive Control of Bioreactors*. Elsevier Science Publishers, Amsterdam.

Chen, L. (1992) *Modelling, Identifiability and Control of Complex Biotechnological Systems*, Ph.D. thesis, Université Catholique de Louvain, Belgium.

Ejiofor, A.O., Posten, C.H., Solomon, B.O. and Deckwer, W.-D. (1994). A robust fed-batch feeding strategy for optimal parameter estimation for baker's yeast production. *Bioproc. Eng.*, **11**, 135-144.

Goodwin, G.C. (1987). Identification: Experimental Design. *Systems & Control Encyclopedia. Theory, Technology, Applications*, (M.G. Singh, Ed.), **4**, Pergamon Press, Oxford, 2257-2264.

Munack, A. (1989). Optimal Feeding Strategy for Identification of Monod-Type Models by Fed-batch Experiments. *Computer Applications in Fermentation Technology: Modelling and Control of Biotechnological Processes* (4th Int. Conf., Cambridge, UK, 1988, N.M. Fish, R.I. Fox, N.F. Thornhill, Eds.), Elsevier Applied Science, London, 195-204.

Na, J.-G., Chang, Y.K., Chung, B.H. and Lim, H.C. (2002). Adaptive optimization of fed-batch culture of yeast by using genetic algorithms. *Bioproc. Biosyst. Eng.*, **24**, 299-308.

Nguang, S.K., Chen, L.Z. and Chen, X.D. (2001). Optimisation of fed-batch culture of hybridoma cells using genetic algorithms. *ISA Transations*, **40**, 381-389.

Pohlheim, H. (2003). Documentation 3.3c for Genetic and Evolutionary Algorithm Toolbox for use with Matlab: toolbox 3.3.

Rocha, C. and Ferreira, E.C. (1996). Design of Optimal Experiments for Identification of Yield Coefficients in a Baker's Yeast Model. 1st European Symposium on Biochemical Engineering Science (Glennon, B., Kieran, P.M., Luyben, K.Ch.A.M., Eds.), 99-100, Dublin.

Rocha, I. and Ferreira, E.C. (2002a). On-line simultaneous monitoring of glucose and acetate with flow-injection analysis during high-cell-density fermentation of recombinant *Escherichia coli*. *Anal. Chim. Acta*, **462**, 293-304.

Rocha, I. and Ferreira, E.C. (2002b). Optimisation methods for improving fed-batch cultivation of *E. coli* producing recombinant proteins. Proceedings of the *10th Mediterranean Conference on Control and Automation – MED2002*, Lisbon, Portugal.

Roubos, J.A., Straten, G. van and Boxtel, A.J.B. van (1999). An evolutionary strategy for fed-batch bioreactor optimization; concepts and performance. *J. Biotechnol.*, **67**, 173-187.

Versyck, K.J. and Van Impe, J.F. (1999). Feed rate optimization for fed-batch bioreactors: from optimal process performance to optimal parameter estimation. *Chem. Eng. Comm*., **172**, 107-124.

Walter, E. and Pronzato, L. (1990). Qualitative and Quantitative Experimental Design for Phenomenological Models – A Survey. *Automatica*, **26**, 195-213.