Proceedings of the 10th International Chemical and
Biological Engineering Conference - CHEMPOR 2008
Braga, Portugal, September 4-6, 2008
E.C. Ferreira and M. Mota (Eds.)

A new hybrid modeling methodology based on delayed differential equations: Application to antibody expression by Pichia pastoris

M. von Stosch[1], R. Oliveira[2], J. Peres[1], S. Feyo de Azevedo[1*]
[1] LEPAE, Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
[2] REQUIMTE/CQFB, Departamento de Quimica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Keywords: Hybrid Modeling, Biosystem Dynamics, Delay Differential Equation, AR(X)
Topic: Systematic methods and tools for managing the complexity

Abstract

In this paper a novel methodology for biosystems dynamic modelling is presented in which discrete time series, namely AutoRegressive (eXogenous) models are incorporated in the traditional parametric/nonparametric hybrid modelling framework. This results in a set of Delay Differential Equations (DDE) which describe the material balances of a bioreactor system in which dynamic kinetics are mimicked by a parametric/nonparametric submodel. The idea is to display better consistency with the nature of biological systems by associating the dynamics of a cellular metabolism to a parametric/nonparametric subsystem. The proposed hybrid structure is evaluated with fed-batch experimental data taken from a process for antibody expression by recombinant *Pichia pastoris* in addition to two simulation case studies. The first of these assumes a discrete time delay and the second assumes a distributed delay between kinetics. In this paper, it is shown that the proposed hybrid model is capable of modelling discrete and distributed delays between kinetics and outperforms the standard hybrid modelling methods with static kinetic models.

1 Introduction

Time delays have been observed in many bioprocesses and, as is well known, they can be source of instabilities and oscillations. In most cases, however, only a certain time delay between the substrate uptake, biomass growth and product formation is observed such as the case of the growth phase of fed-batch *Saccharomyces cerevisiae* or the *Pichia pastoris* cultivations (Ren et al, 2003). Many phenomenological models that consider discrete delays (Wolkowicz et al, 1997), distributed delays (Daugulis et al, 1997, Wolkowicz and Xia, 1997), ordinary differential equations (ODE) of kinetic rates (Ren et al, 2003) or other time delay considering techniques have been reported. They are usually based on the general mathematical concept of Retarded Functional Differential Equations (RFDE), Bocharov and Rihan (2000). On the one hand, these models are capable of explaining stability of processes and are suitable for the estimation of process key variables. However, on the other hand, their application to other cell systems is limited and their development is cost expensive. In contrast, hybrid modelling has been reported to be a suitable, cost effective alternative capable of being applied to a number of cell systems, Oliveira (2004). Hybrid models combine mechanistic knowledge and process knowledge in form of mechanistic models and data-base nonparametric models. Mechanistic and nonparametric models can be arranged in two possible manners: parallel or serial. In the serial structure, which has been applied in this study, the process dynamics are described by time differentials of process classifying variables and the cell system is mimicked by a parametric/nonparametric submodel. However, until now, cell system dynamics have not been taken into account by these submodels, but it is known that cell systems are sources of time delays. Thus, in this paper, the dynamics of the cell system will be ascribed to the parametric/nonparametric

* Corresponding author. Tel + 351-22-508-1694. E-mail: sfeyo@fe.up.pt (S.F. de Azevedo)

submodel. Similar techniques as in the case of phenomenological models could be used, for example discrete delays or distributed delays of state variables in the kinetics or differential equations of the kinetics. The latter is not appropriate, because both the kinetic function and the kinetic values remain unknown. Therefore, a solution or estimation of the kinetics is not a straightforward process. Distributed delays are also rather unlikely to be used, due to the fact that a mathematical postulation of arbitrarily large delays for unknown weighting functions of the delayed variable would have to be assumed and this mathematical convenience is in limit biologically unrealistic (Bocharov and Rihan, 2000). Instead, the use of discrete delays in the hybrid model estimated values and in additional measurements, when available, is proposed in this paper. This is analogous to the application of discrete time series, namely AR(X) models. Whereas, in theory, an endless number of time lagged values of one variable can be used as inputs to the nonparametric function, in practice this would lead to long training times and to identification problems of the network structure and parameters (Haykin (1999)). Hence an optimal number of time lagged values exists which represent the proportion between redundancy and the additional gain of information in the inputs. This has been neglected by this paper and instead it has been shown that a limited number of time lagged variable values significantly enhances the prediction capacity of the model and that time lags and the number of time lags have been chosen by trial, which was for example also the case of Parlos et al (2000).

The remainder of the paper is organized as follows. Section 2 presents the embedding of discrete time series into the hybrid parametric/nonparametric structure proposed by Oliveira (2004) along with the changes to the sensitivity equations. The evaluation of the structure which has been presented through the analysis of two simulation cases in addition to experimental fed-batch data of a *Pichia pastoris* process is carried out in section 3. Finally, conclusions are presented in section 4.

2 Hybrid Model Structure

The serial parametric/nonparametric hybrid model structure presented in this paper is shown in figure 1. The structure is based on the originally proposed model by Oliveira (2004). However, in this paper, discrete time series, namely AR(X), are incorporated into the hybrid framework, resulting in a set of DDE for the bioreactor system. Then, the identification of the parameters of the nonparametric function is carried out.

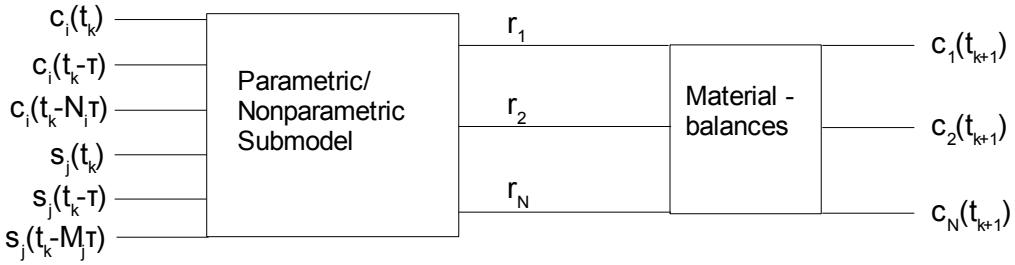

Figure 1: Structure of the proposed serial parametric/nonparametric hybrid model structure

2.1 General hybrid parametric/nonparametric structure

A bioreactor model can be expressed through a set of material balance equations, which describe the dynamics of state variables such as biomass, substrate, product, etc.,

$$\frac{dc}{dt} = r - D \cdot c + u \,. \qquad (1)$$

Here c is a vector of state variables, D is the dilution rate, u is a vector of volumetric control inputs, and r is the kinetic rate vector. The vector of kinetic rates combines, when available, first principle knowledge with nonparametric functions and according to Oliveira (2004) is:

$$r(c,w) = K(\psi_j(c) * \rho_j(X,w)) \,, \qquad (2)$$

where K is a $n \times m$ yield coefficient matrix, ψ_j are m kinetic functions from mechanistic knowledge and where ρ_j are m unknown kinetic functions which have been modelled with nonparametric techniques, X stands for the vector of inputs and w represents the vector of parameters. Nonparametric techniques have the ability of accounting for nonlinear mappings between inputs and outputs. Yet, this nonparametric modelling of the kinetic functions has barely taken the dynamics of the cell system into account, i.e. the input vector only contemplated the current concentration values $c(t)$ and/or the current exogenous inputs $s(t)$. In this study, and in analogy to the AR(X) models, discrete past values of the model outputs and the exogenous input are also contemplated as inputs to the nonparametric function, resulting in the following equation

$$X = \begin{bmatrix} c_i(t), c_i(t-\tau_i), c_i(t-2\cdot\tau_i), \ldots, c_i(t-N_i\cdot\tau_i), \\ s_j(t), s_j(t-\tau_j), s_j(t-2\cdot\tau_j), \ldots, s_j(t-M_j\cdot\tau_j) \end{bmatrix} . \quad (3)$$

In this equation c_i represents the i[th] value of vector c, τ_i is the associated time lag, N_i defines the number of time lags assumed for each value c_i of vector c, s_j is the j[th] exogenous input, τ_j the associated time lag and its lag number is defined by M_j. Note that the time lags and the numbers of time lags, τ_i, τ_i, N_i, and M_j have been chosen by trial as was previously referred to in the introduction. The nonparametric model adopted here is a three layer back propagation ANN with hyperbolic tangent activation function formulated as ρ_j:

$$\rho_j(X, w) = w_2 \cdot g(w_1 \cdot X + b_1) + b_2 , \quad (4)$$

where w is the vector form of the weights and biases, w_1, w_2, and b_1, b_2, respectively. The hyperbolic tangent activation function $g(\cdot)$ is,

$$g(x) = \frac{2}{(1+\exp(-2\cdot x))} - 1 . \quad (5)$$

By merging Equations (1) – (5) it becomes clear that these model equations which describe a bioreactor system with intracellular dynamics are DDEs as “retarded” or “lagged” phenomena are accounted by the nonparametric submodel.

2. The identification of the parameters of the nonparametric submodel

In this study, a least squares criteria of residual concentrations has been adopted to identify the nonparametric model parameters vector w through process data. This criteria is formulated by the following expression:

$$min\left\{ E = \frac{1}{P \times n} \sum_{l=1}^{P} \sum_{i=1}^{n} \frac{(c_{m,i}(t) - c_i(t, w))^2}{c_{max,i}} \right\} , \quad (6)$$

where P represents the number of measured patterns, n is the number of state variables, $c_{m,i}$ are measured state variables, $c_i(w, t)$ represent calculated state variables and $c_{max,i}$ are the scaling factors. The serial hybrid structure of ANN and material balances has been proven to be trained most effectively when using sensitivity approach along with analytical gradients, Oliveira (2004). The analytical gradients are obtained differentiating equation (1) with respect to w while taking the time lagged differential variables into consideration which reads as follows,

$$\frac{d}{dt} \cdot \frac{\partial c}{\partial w} = \sum_{k=0}^{N_i} \left\{ \frac{\partial (K \cdot \psi \cdot \rho)}{\partial c(t-k\cdot\tau)} \cdot \frac{\partial c(t-k\cdot\tau)}{\partial w} \right\} + \frac{\partial K \cdot \psi \cdot \rho}{\partial w} + D \cdot I_n . \quad (7)$$

This least square problem is solved by using the “lsqnonlin” Matlab function which uses a subspace trust region method and is based on the interior-reflective Newton method (Matlab Optimization toolbox) and which favours analytical gradients. The sensitivity equation (7) is integrated along with the delay differential model equations (1)–(5). For integration the differential equations are linearly approximated which results in a time inexpensive algorithm. Unfortunately, some error has been introduced due to this simplification, but if average kinetic rates are estimated for each time step, the error is significantly minimized. Initial

values of sensitivity equations are $(\partial c/\partial w)_{t=0}=0$, as the initial state variables are independent of w and as the gradients $(\partial c/\partial w)_{t<0}=0$ for $t<0$. The residual gradients are then obtained by using the corresponding sensitivity values. It is important to stress that the lagged values of either state variables and exogenous inputs are assumed to be equal to the initial values for all $t-\upsilon\cdot\tau<0$.

3 Results and Discussion

The evaluation of the original and the proposed (DDE) Hybrid Model has been carried out through the analysis of two fed-batch simulation cases in addition to fed-batch experimental data of a *Pichia pastoris* fermentation.

3.1 Simulation cases

In both simulation cases the bioreactor system is modelled by material balances, resulting in a set of differential equations of biomass, substrate and product concentrations and also of the reactor volume, in which the kinetics are expressed as follows. The substrate uptake rate is modelled with Monod kinetics and depends on the current substrate concentration. Specific growth is composed from some bias which accounts for maintenance and the Monod kinetics of a lag variable. Product formation depends linearly on the specific biomass growth. For the lag variable two different approaches have been adopted. The first, was inspired by Wolkowicz and Xia (1997), and it considers a discrete time lag in the substrate concentration as the lag variable. The second, which was inspired by Daugulis et al (1997), Wolkowicz et al (1997), considers distributed delays in the substrate concentration to be the lag variable. The substrate feeding in both cases is linearly controlled in regards to user desired set points of substrate concentrations.

For the evaluation of the DDE hybrid structure, three sets of data, namely training, validation and test data, have been used. This data consist of simulated fed-batches in which concentrations of biomass, substrate and product, the reactor volume and as well the feeding concentration are assumed to be the measured data.

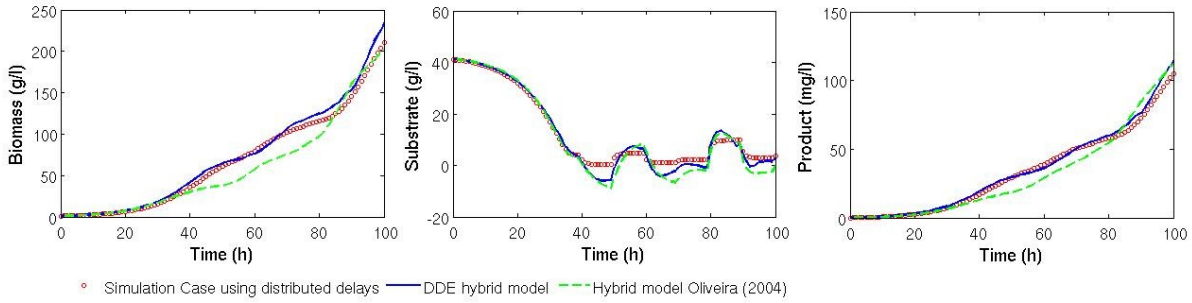

Figure 2: Plot of concentrations over time for data obtained with the DDE hybrid model (solid line, blue), the hybrid model Oliveira (2004) (dashed line, green) and the true simulation using the distributed delay model (o, no line, red)

3.1.1 Hybrid Model Structures

The standard hybrid model structure (Oliveira (2004)) and the DDE hybrid model structure describe three state variables: biomass, product and substrate concentration. The kinetics of each of them are estimated by training an Artificial Neural Network (ANN) with the training set. The estimated substrate concentration, in the standard hybrid model, is the only input to the Artificial Neural Network. In contrast, a series of time delays in the estimated substrate concentrations are considered as being inputs to the ANN for the DDE hybrid model.

The training and identification of the best network structure is carried out according to the best value, namely the greatest value, for the Bayesian Information Criteria (BIC) (Burnham and Anderson (2004)), obtained for the validation set. The test set is used to explore the model generalization capabilities.

3.1.2 Results of Simulation Cases

In Table 1 a selection of the best results obtained for both simulation cases with the standard and the DDE hybrid model are presented. The BIC values for the DDE hybrid model are found to be better than those of the standard hybrid model for both simulation cases. The consistency of all the BIC values obtained for different model structures is elevated although it is not presented here. In fact, the enhancement in model prediction has also been reflected in plots of estimated concentrations and "true" concentrations over time, as being exemplary shown for predictions of a fed-batch considering distributed delays in Fig. 2. Therefore, it is possible to conclude that the proposed DDE hybrid model is capable of accounting for all kinds of time delays observed in process by allowing series of time lagged values of state variables as inputs to the nonparametric model. Thus, a significant enhancement is to be expected for the modelling of experimental data.

3.2 Experimental data of antibody expression of recombinant *Pichia pastoris*

Pichia pastoris fermentation finds application here, as time delays between substrate uptake, biomass growth and product formation have been observed (Ren et al (2003)). Data of Temperature, pH, biomass and product concentrations and the accumulated mass of glycerol and methanol were measured as described in Cunha et al (2004). The bioreactor system is modelled by differential equations of biomass and product concentrations and of the reactor volume while other measurements are considered for the identification of kinetics.

3.2.1 Hybrid Model Structure

In total, four data sets of fed-batch fermentations have been recorded, three of which are used for the training of the hybrid model and the last which is used for its validation. Biomass and product concentrations and the reactor volume are used as state variables. Substrate concentrations, ie glycerol and methanol concentrations have unfortunately not been measured, but instead measurements of accumulated mass of feeding of glycerol and methanol were available. The estimated biomass concentration along with measurements of substrates, such as the sum of mass of glycerol and methanol, temperature and pH, at time t, are inputs for the standard hybrid model. It was assumed that the accumulated mass of feeds along with the estimated biomass concentration could compensate for the lack of substrate concentrations. However measured glycerol and methanol concentrations might have lead to an even more accurate representation of the complexity of the system under analysis. For partial compensation of time delayed substrate concentrations, a time delay in the biomass concentration was taken into account. This is due to the assumption that the delay appearing in the cell metabolism is somewhat similar to a time delay of the cell. Such, in addition to the standard hybrid model, the influences of time lags in the estimated biomass concentration and the measuring of temperature, pH, the sum of mass of glycerol and of methanol on the DDE model estimates have been studied separately and subsequently, a joint delay model for the most significant variables and time delays was studied, namely biomass concentration, temperature and the accumulated mass of glycerol and methanol. The identification of the best network structure was carried out as before.

3.2.2 Results for experimental data

A selection of results is presented in Table 1, revealing the best BIC values for the standard hybrid model and the best BIC values of the DDE hybrid model for lagged inputs of biomass concentration, pH and temperature. The influence of time lags on the estimates, when each variable was studied separately, did not result in significant enhancements of the BIC values in comparison to the standard hybrid model and have therefore not been presented here. The BIC values for the combinations of delayed variables, namely biomass concentration, pH and temperature have been found to be significantly better than the ones of the standard hybrid model, see Table 1. The enhancement of combinations of delayed variables in comparison to the separate delayed variables can be explained by the complexity of the system under study. The state of a system is generally characterized by a set of variables. The time delays in the kinetics are provoked by the cell metabolism which experiences a

change of the state of the system with time. Therefore, combinations of time delays in state characterizing variables lead to better BIC values in contrast to when they are analysed separately. In addition they are also biologically more realistic. However it has been clearly demonstrated that the use of some time lagged values of concentration estimations and exogenous measurements significantly enhances the hybrid model's process predictions.

Simulation Case	Model	NN	Time lag (h)	N_i	BIC train	BIC valid	BIC test
Discrete Delays	Standard hybrid model	2	-	-	-12332	-3256	-3399
Discrete Delays	DDE hybrid model	3	5	1	-11115	-2883	-2882
Distributed Delays	Standard hybrid model	2	-	-	-18942	-5127	-5149
Distributed Delays	DDE hybrid model	4	2	5	-17020	-4552	-4587
Experimental Data	Standard hybrid model	7	-	-	-43373	-13755	-
Experimental Data	DDE hybrid model	3	2	2	-42549	-12890	-
Experimental Data	DDE hybrid model	4	2	2	-42909	-13545	-

Table 1: Best BIC results of both simulation cases and experimental data for the standard and DDE hybrid model, where NN is the number of nodes in the hidden layer of the ANN

4 Conclusions

In order to account for dynamics in the cell metabolism, discrete time series have been incorporated into the hybrid model originally proposed by Oliveira (2004), leading to Delayed Differential Equations (DDE). More accurate prediction qualities of the DDE hybrid model than those obtained through the standard hybrid model have been achieved when applied to two simulation cases containing either discrete or distributed delays between the kinetics. Therefore, it has been concluded that the DDE hybrid model is capable of accounting for all the time delays observed in bioprocess systems. Expectations in regards to the application to the antibody expression of recombinant *Pichia pastoris* have been met, where a significant enhancement of process prediction has been achieved, in which a limited number of time lagged values of predicted concentration and exogenous measurements have been used.

References

Bocharov G.A., Rihan F.A., (2000), Numerical modelling in biosciences using delay differential equations, J. of Computational and Appl. Math., **125,** 183-199.

Burnham K.P., Anderson D.R., (2004), Multimodel inference - understanding AIC and BIC in model selection, Social Methods & Research, **33**, 261-304

Cunha A.E., Clemente J.J., Gomes R., Pinto F., Thomaz M., Miranda S., Pinto R., Moosmayer D., Donner P., Carrondo M.J.T., (2004), Methanol Induction Optimization for scFv Antibody Fragment Production in Pichia pastoris, Biotech. and Bioeng., **86**, 458-467

Daugulis A.J., McLellan P.J., Li J., (1997), Experimental Investigation and Modeling of Oscillatory Behaviour in the Continuous Culture of Zymomonas mobilis, Biotech. and Bioeng., **56**, 99-105

Haykin S. (1999), Neural Networks – A Comprehensive Foundation 2nd edition, Prentice Hall Inc., Upper Saddle River, New Jersey, America

Oliveira R., (2004), Combining first principles modelling and artificial neural networks: a general framework, Computers & Chemical Engineering, **28,** 755-766.

Parlos A.G., Rais O.T., Atiya A.F., (2000), Multi-step-ahead prediction using dynamic recurrent neural networks, Neural Networks, **13**, 765–786

Ren H.T., Yuan J.Q., Bellgardt K.H., (2003), Makrokinetic model for methylotrophic Pichia pastoris based on stoichiometric balance, J. of Biotech., **106**, 53-68

Wolkowicz G.S.K., Xia H.,(1997), Global Asympotic Behavior of Chemostat Model with Discrete Delay, J. Appl. Math., **57**,1019-1043

Wolkowicz G.S.K., Xia H., Ruan S., (1997), Competition in Chemostat: A Distributed Delay Model and its Global asymptotic Behaviour, J. Appl. Math., **57**, 1281-1310