**Hybrid modeling of bioprocesses: revisiting the direct identification method**

José Pinto1, Gerald Striedner2, Rui Oliveira1\*

*1 REQUIMTE, FCT/Uni NOVA Lisboa, Portugal; 2 Dep. Biotechnology, BOKU University, Vienna, Austria*

*\*Corresponding author: rmo@fct.unl.pt*

**Highlights**

* Modification of the direct identification method.
* Much faster and stable identification of hybrid models.
* Improved scalability of hybrid models to more complex biological systems.

**1. Introduction**

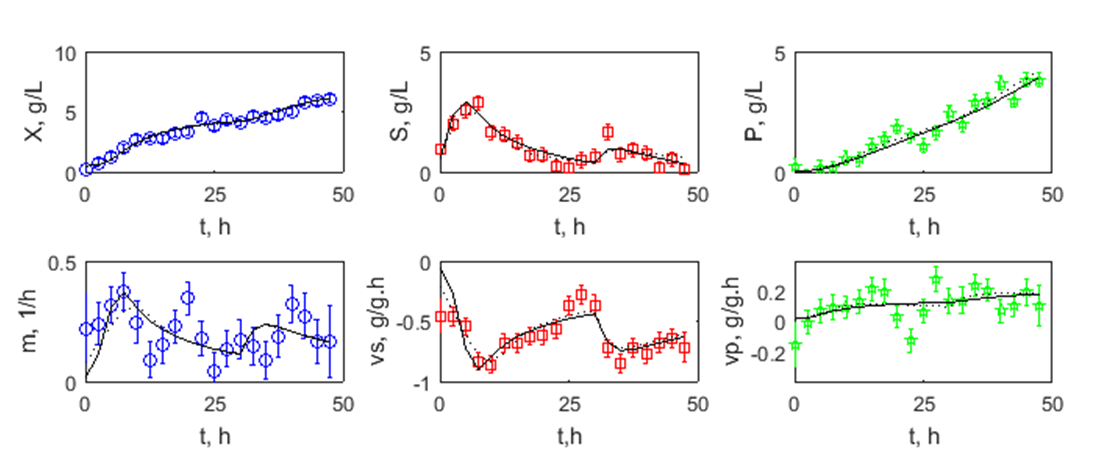
The general bioreactor hybrid model has been used many times to describe the dynamics of a bioprocess (e.g. Psichogios and Ungar, 1992, Oliveira, 2004). This model is primarily based on a set of material balance equations of the key substrates/products taking the form of a system of ODEs. The reaction kinetics such as the specific growth rate, m, substrate uptake, vs, and product synthesis rates, vp, are less defined parts of the model. They are the result of thousands of metabolic reactions regulated at different levels by genes, proteins and environmental factors. In hybrid models the rates are modeled by artificial neural networks (ANN) acknowledging the fact that they are too complex to be described mechanistically. The combination of ODEs with ANNs gives rise to so-called hybrid semiparametric models with particular identification challenges. Training the ANN linked with ODEs is not the same as a standalone ANN. There are two main methods to train the ANN: the direct (e.g. Tholudur & Ramirez, 1996) and the indirect method (e.g. Psichogios and Ungar, 1992). The indirect method is much more complex, slower and less flexible than the direct method but it typically produces more accurate and above all more stable results. For this reason it has been preferred in the literature.

**2. Modified direct identification method**

The direct identification is a two-steps method where the material balances and ANN are decoupled. In a first step the the biologic kinetic rates are calculated from the experimentally measured concentrations by solving the material balance equations (e.g. Tholudur & Ramirez, 1996). In the second step the calculated rates are used as target outputs for the training of the ANN. The training can proceed with standard techniques such as error backpropagation. In this study we propose the time integral of weighted least squares of volumetric reaction rates as cost function for ANN training in the second identification step.

**3. Results and discussion**

A synthetic data set was generated by simulation of a simple fed-batch bioprocess whereby biomass, X, grows on substrate, S, and produces product, P, over time, t. Gaussian noise was added with a standard deviation of 0.3 g/L for all variables. These data are represented by the symbols in the upper panel plots of Fig. 1. The biologic kinetic rates were calculated from the concentrations by differentiation. Smoothing splines were applied to the concentrations X, S and P over time, followed by analytical differentiation of the splines function to obtain estimates of derivatives dX/dt, dS/dt and dP/dt at given time points. A Monte Carlo approach was adopted whereby the calculation of rates is repeated 100 times by adding Gaussian noise with s.d. of 0.3 g/L to the concentrations. This procedure resulted in the “experimental” reaction rates represented in the lower panel plots of Fig. 1. These data were set as target outputs of a 3-layered ANN with 1 input (substrate concentration, S) and 3 outputs (the specific rates). The results of the modified direct identification method are shown in the lower panel for the rates (full line is true property; dashed line is ANN prediction). Finally, the integration of ODEs coupled with the trained ANN resulted in the predicted concentrations shown in the upper panel plots (full line is true property; dashed line is ANN prediction)



**Figure 1.** Hybrid modeling results using the direct identification approach. Symbols are measurements, full line is true property and dashed line is the respective hybrid model prediction. Upper panels from left to right represent concentrations of biomass (blue), substrate (red), product (green) over time. Lower panels represent from left to right the specific growth rates of biomass (blue), substrate (red) and product (green).

**4. Conclusions**

As seen in the results the modified indirect identification method enabled an almost perfect identification of the specific reaction rates (comparison between the full and dashed lines). Equivalent results are however also obtained with the classical direct approach. The most remarkable result is the fact that the integration of ODEs+ANN (i.e. the full hybrid model) resulted in almost perfect identification results in terms of concentrations. Opposed to this, the traditional direct identification method typically fails in the prediction of concentrations. Thus we conclude that the modified direct method has similar performance to the currently preferred indirect methods. This is very significant as the scale and complexity of hybrid models can be largely augmented by applying a much faster and much more flexible direct identification methods.

**References**

1. D.C. Psichogios, L.H. Ungar, AIChE. J. 38 (1992) 1499–1511.
2. R. Oliveira, Comp Chem Eng, 28 (2004) 755-766.
3. A. Tholudur, W.F. Ramirez, Biotech. Prog., 12 (1996) 302-309