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

PRODUCT FORMATION KINETICS IN A RECOMBINANT PROTEIN PRODUCTION PROCESS

Stefan Gnoth[1], Marco Jenzsch[1], Rimvydas Simutis[2], Andreas Lübbert[1,*]

[1]Center of Bioprocess Engineering, Martin-Luther-University Halle-Wittenberg, D-06120 Halle/Saale, Germany
[2]Institute of Automation and Control Systems, Kaunas University of Technology, LT-3028 Kaunas, Lithuania
** Author for correspondence (Tel: +49-345-5525-942; Fax: +49-345-5527-260;*
E-mail: andreas.luebbert@biochemtech.uni-halle.de

Abstract: Protein formation in recombinant protein production cannot yet be modeled in a way sufficiently accurate for process supervision and control. Here we propose using a new hybrid approach based on mass balances for the state variables involved, where the kinetics are represented by artificial neural networks (ANN). We first demonstrate by means of simulations that this method works well even when the networks are trained on noisy process data. Then, secondly, we show that the method is applicable to real fermentation data. As an accompanying example we use an E.coli culture that produces a recombinant protein, namely the green fluorescent protein GFP, which remains dissolved within the cytoplasm. For this case the ANN resulted in a concrete relationship between the specific product formation rate π, the specific growth rate μ and the specific product concentration p/x. The $\pi(\mu)$-part of the relationship confirms what was obtained with a conventional approach and the additional information about the influence of the specific product concentration characterizes the metabolic load of the cell.

Keywords: Fermentation processes, Process models, Neural networks, Identification Soft sensing

1. INTRODUCTION

Bacterial fermentation is a major workhorse for producing recombinant therapeutic proteins; hence, it is very desirable to derive optimal fermentation control strategies.

As the mass of product that can finally be purified from the culture depends of the amount of biomass x employed and their performance, represented by their specific product formation rate π, one is interested in high cell density cultivations with well performing cells (Lee 1996, Riesenberg and Guthke 1999). In most industrial production systems, both factors determining product mass are primarily dependent on the specific biomass growth rate μ. This may be trivial for x, but is in most cases also valid for π: The growth rate that a particular fermentation medium supports, determines the physiological state of the cells and particularly the cell's protein-synthesizing machinery, and in most industrially relevant cases, recombinant protein production is under growth rate control (Neidhardt et al. 1990). Consequently, much work has been devoted to controlling the specific biomass growth rate in fermentation processes (Shioya 1992, Yoon et al. 1994, Levisauskas et al. 1996, Kim et al. 2004, Picó-Marco et al. 2005, Jenzsch et al. 2005, and 2006a, Soons et al. 2006). Numerical exploitable models of fermentation processes for recombinant protein manufacturing thus need a sufficiently accurate submodel relating the specific growth rate μ to the specific product formation rate π, the so-called π-μ-relationship (Pirt 1993).

Traditionally, optimal process trajectories have been obtained from mechanistic models of the processes under consideration (e.g. Levisauskas et al. 2003).

The latter can be derived step-by-step where the actual model version is used to compute the optimal process procedure, e.g. in terms of the productivity with respect to the product, and improvements of a model are deduced from the deviations between the predicted values and those measured in a validation experiment (Galvanauskas et al. 1997, 2004).

Here we propose a new alternative to this basic approach which is purely data-driven. It has the disadvantage of needing much data, but the decisive advantage of not being restricted by unproven model assumptions. At running production plants the supply of many data records is not a problem at all, hence, in these cases, the advantages clearly prevail.

The method proposed is based on artificial neural networks describing the more or less insufficiently known process kinetics within a well known set of basic mass balance equations. Since such hybrid modeling usually suffers from the fact that there are no directly measurable data for the key variables, e.g. μ and π, we must train the artificial neural networks depicting the really interesting kinetic relationships indirectly, extending the work of Simutis and Lübbert (1997). We solved this problem by a stepwise training of neural networks using online measured variables and, additionally, corresponding off-line values for the amount of biomass x, and total product mass p. The result of this training procedure is a $\pi(\mu)$-profile which can be used for process simulation, and finally in process supervision and control.

Validation of the model was performed at the example of E.coli fermentations, where the soluble GFP, the green fluorescence protein was produced in its active form within the cells' cytoplasm.

2. STRUCTURE OF THE DATA-DRIVEN MODEL

2.1 General idea behind the model.

The backbone of the process model is a classical system of mass balance equations for all species, the masses of which are changing significantly during the cultivation process. The components considered here are total biomass x, and total product mass p.

The first step in modeling the kinetics is representing the specific growth rate μ. It can be determined using nonlinear relationships in form of an ANN with important process variables such as carbon dioxide production rate (CPR), total biomass x, time after induction t_{ai} , etc.. Also, other online variables can be used to strengthen this relation, e.g., the oxygen uptake rate, as well as the base fed into the reactor during pH control.

This specific growth rate representation can directly be used within balance equations determining the amount of biomass. In the upper part of Figure 1 this procedure is schematically shown. It can be interpreted as an ANN-aided software sensor estimating the total biomass x. Once this ANN is trained, it can supply μ(t)-values for training the a second artificial neural network computing π. This procedure is shown in the lower part of Figure 1.

2.2 Training of the artificial neural network system

Simple feedforward networks are used that map the input variables across a hidden layer of 5 nodes (hyperbolic tangent) onto a single output variables μ or π respectively. As already mentioned, online measurements data (CPR, t_{ai}, ...) are used as inputs together with biomass x and product p, estimated in the time step before. For network training we used offline measurement data for biomass x as well as total product mass p from previously performed experiments.

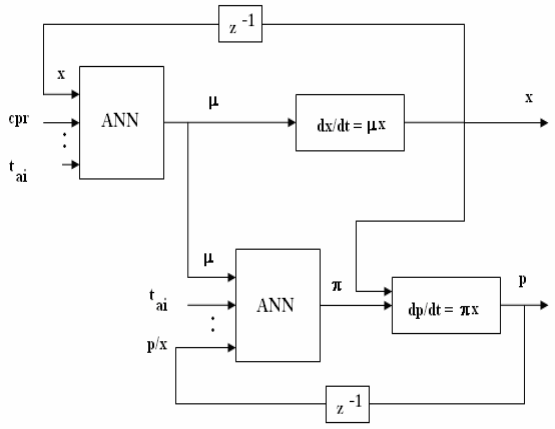

Figure 1: Scheme of the proposed procedure for identification of the $\pi(\mu)$-relationship. The artificial neural networks (ANN) are feedforward networks with a single hidden layer.

Network training was based on the sensitivity equations approach (e.g., Schubert et al., 1994). This can be applied to train neural networks, which are incorporated into differential equation systems of the form

$$\frac{dy}{dt} = f(y(t), W) \qquad (1)$$

where W are weights of the neural network, and y are process state variables. The training is essentially a fit of this equation to experimental offline measurement data for biomass x and product p. Its efficiency can be improved if the gradients $\partial y / \partial W$ can be exploited. These gradients satisfy an ordinary differential equation that can easily be derived from equation (1) by partial derivatives with respect to the weights W.

$$\frac{d}{dt}\frac{\partial y}{\partial W} = \frac{\partial f}{\partial y}\cdot\frac{\partial y}{\partial W} + \frac{\partial f}{\partial W}, \; with \; \left.\frac{\partial y}{\partial W_p}\right|_{t=0} = 0 \qquad (2)$$

Equation (2) is referred to as the sensitivity equation. With the solutions, the $\partial y / \partial W$ values, the well known neural networks training procedures (backpropagation, gradient methods, cf. e.g., Rumelhard and McClelland 1986) can be applied to train the neural network.

The sensitivity equation approach for specific growth rate estimation appears when y is replaced by x, the biomass and equation (1) is specified by the equation defining the specific growth rate μ

$$\frac{dx}{dt} = \mu \cdot x = \mu(x, cpr, t_{ai}, \ldots W)x \qquad (3)$$

In order to determine the sensitivity of the rate of change of x with respect to changes in the network weights W, equation (3) is partially be differentiated with respect to W

$$\frac{d}{dt}\frac{\partial x}{\partial W} = \frac{\partial f}{\partial x}\cdot\frac{\partial x}{\partial W} + \frac{\partial f}{\partial W} = (\mu + x\cdot\frac{\partial \mu}{\partial x})\cdot\frac{\partial x}{\partial W} + x\cdot\frac{\partial \mu}{\partial W} \quad (4)$$

Eq. (4) can be identified to be an ordinary differential equation in $\partial x/\partial W$. This can be solved with the initial condition $\partial x(t=0)/\partial W=0$, if the sensitivities $\partial\mu/\partial x$ and $\partial\mu/\partial W$ can be supplied. These can easily be determined if the structure of the artificial neural network is known (Rumelhard and McClelland 1986).

Once the μ profile is computed with the first part of the model as depicted in Figure 1, it can be used to train the network estimating the specific product formation rate π and the product mass p in the second part of the model. The approach is essentially the same as in the first part. Because values for π are also not known beforehand, we once again use the sensitivity approach.

The sensitivity equations for the estimation of π from μ are derived from the basic equation

$$\frac{dp}{dt} = \pi \cdot x = \pi(\mu, t_{ai}, \frac{p}{x}, \ldots W_p)\cdot x \qquad (5)$$

Again the sensitivities are obtained by partial derivative of both sides of the equation with respect to the network weights W_p:

$$\frac{d}{dt}\frac{\partial p}{\partial W_p} = \frac{\partial f}{\partial p}\cdot\frac{\partial p}{\partial W_p} + \frac{\partial f}{\partial W_p} = (x\cdot\frac{\partial \pi}{\partial p})\cdot\frac{\partial p}{\partial W_p} + x\cdot\frac{\partial \pi}{\partial W_p} \qquad (6)$$

Once again the sensitivity equation (6) can be interpreted as an ordinary differential equation that can be solved with the initial condition

$$\left.\frac{\partial p}{\partial W_p}\right|_{t=0} = 0 \qquad (7)$$

Using values $\partial x/\partial W$, $\partial p/\partial W_p$ and off line measurements of x and p, the well known back-propagation and cross-validation procedures can be applied to train the network (Leonard and Kramer 1990, Haykin 1999).

3. TEST SIMULATIONS

In order to test the proposed procedure it is straightforward to first examine it at well defined conditions. These can be provided by means of numerical simulations using a model (see Appendix), where a concrete π(μ)-relationship is used resembling realistic process conditions. Typical results for the π(μ)-relationship obtained from the simulated fermentation data are shown in Figure 2.

The performance criterion was the standard deviation σ of the π(μ)-relationship assumed in the model and that estimated from the simulated data.

$$J = \sigma(\pi_{model} - \pi_{estimated}) \qquad (6)$$

The value J of this criterion depends on the quality of the data obtained from the process, i.e. on the accuracy or noise which corrupts the process and the measurement devices. Hence, the data from the model were distorted by adding a zero-mean-noise component on x, p, and CPR.

The noise levels were chosen to be 3% of the actual values of *cpr* and *x*. For the product mass the uncertainty was assumed to be somewhat higher. The noise for p ranged from 3-20% of the actual value. 10 individual simulations were made for each noise levels and the corresponding standard deviations are recorded. The results are compiled in Table 1.

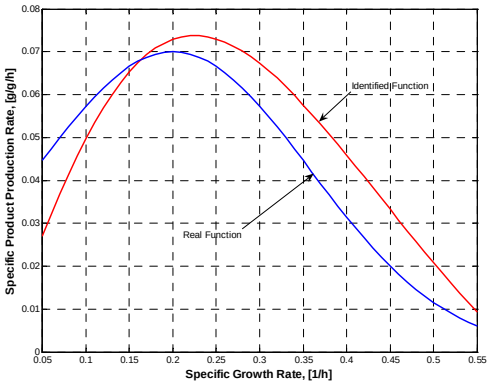

Figure 2: Estimation of the π(μ) relationship from data simulated using a classical model expression (Appendix). The noise levels of cpr and x were assumed to be 3% of the actual values, 5 % measurement noise was assumed for p. Off line measurement values were computed from the model with time increments of 1 [h], The standard deviation for the estimate of π was determined to be σ=0.008 [g/g/h].

Table 1: Standard deviation as a function of the noise level on the product mass p

Relative Noise on p	σ
5 %	0.008
7 %	0.0083
10 %	0.0085
15 %	0.009
20 %	0.012

We can conclude from this investigation that up to 15% noise in the protein data do not severely influence the accuracy of the π(μ)-estimate. It should be noted that we are speaking about randomly appearing measurement errors in the product mass p. A notable influence begins above 15% error. However, as 15% is a good estimate for the accuracy of protein measurements in our laboratory, the method should work in practice.

Also, the influence of the sampling interval for off-line measurements on identification quality was tested. The result was that the usually taken intervals of 0.5 h or 1 h do not lead to different results within the accuracy that can be obtained with this method.

4. EXPERIMENTAL

Experimental data were taken from cultivation processes with a genetically modified E.coli strain that produces the well known green fluorescent protein (Jenzsch et al. 2006). All experiments were performed with *E.coli* BL21(DE3) as the host cell. The recombinant target protein was coded on the plasmid pET 11a and expressed under control of the T7 promoter after induction with isopropyl-thiogalactopyranosid (IPTG). The strain was resistant against ampicillin. The product appears in its active (fluorescing) form within the cells' cytoplasm.

All the experiments were performed within BBI Sartorius System's BIOSTAT® C 15-L- bioreactor. The fermenter was equipped with 3 standard 6-blade Rushton turbines that could be operated at up to 1400 [rpm]. The aeration rate could be increased up to 24 [sLpm]. Aeration rate and then stirrer speed were increased one after the other in order to keep the dissolved oxygen concentration at 25 [%] saturation.

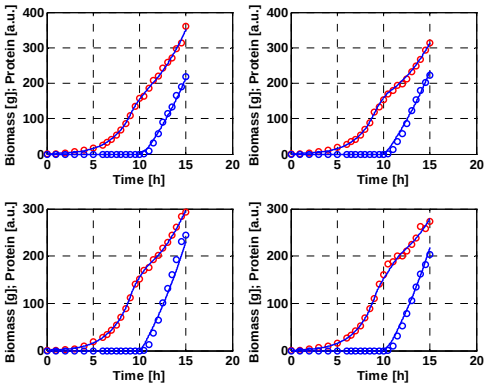

Figure 3: Typical examples of the experimental data: The full symbols are the offline measurement values. The lines are the online estimates obtained from the hybrid model.

The fermentations were operated at pH 7 and a temperature of 35 [°C] in the fed-batch mode, starting with a volume of 5 [L]. The main C- and energy source, glucose, was fed at a concentration of 400 [g/kg]. For more details the reader is conferred to Jenzsch et al. (2006a).

CO_2 in the vent line was measured with MAIHAK's Unor 610®, O_2 with MAIHAK's Oxor 610®. The total ammonia consumption during pH control was recorded with a balance beneath the base reservoir. All these quantities were measured online.
Biomass concentrations were measured offline via optical density at 600 [nm] with a Shimadzu® photospectrometer (UV-2102PC). Glucose was determined enzymatically with a YSI 2700 Select Bioanalyzer. The product was measured with a spectro-fluorimeter (Hitachi F-2500).

Cultivations were started as fed-batch processes, where the substrate was added with an exponential feed rate F(t) computed for a fixed set-point μ_{set} of the specific growth rate. During the biomass formation phase, the specific growth rate μ_{set} was kept at 0.5 [1/h].

3. MODEL IDENTIFICATION

Data records from 29 cultivation runs were used to identify the process model described above. The biomass estimates, which can directly be compared with experimental data, perfectly agree with the measurements (Figure 3). The root-mean-square deviation was 4 [g] for biomass.

For the $\pi(\mu,p/x)$-relationship the result depicted in Figure 4 was obtained. It clearly shows that the preferred specific growth rate for GFP-production is 0.14 [1/h]. However, the more product becomes accumulated within the cytoplasm, the lower is the specific product formation rate π.

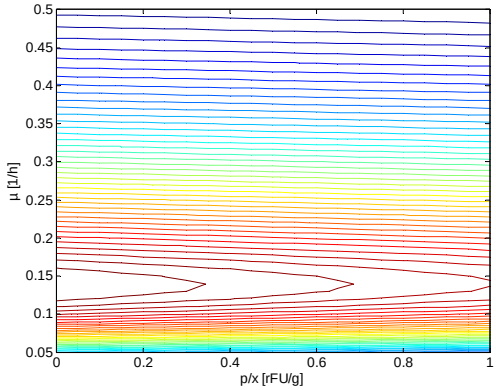

Figure 4: $\pi(\mu,p/x)$ relationship derived from 29 experiments performed under various culture conditions

4. RESULTS AND DISCUSSION

Previous work on optimizing the operational procedure for recombinant protein manufacturing processes was based on rather simple assumptions about the specific protein formation kinetics. In most cases simple formal approaches, e.g., a Luedeking-Piret-type model assumption was made.

As heterologous product formation in microbial systems is an extremely complicated process, generally accepted mechanistic models that can be used for process supervision and control are not yet available. Hence it is straightforward to remove any unproven model assumption from determining this kinetics and to perform a pure data-driven analysis.

The $\pi(\mu)$-relationship resulting from the data analysis worked out in this paper has a form that is immediately convincing. An optimum appears at a relatively low specific growth rate μ. This gives the protein molecules time to fold correctly. Further, the specific product formation rate π drops while foreign product is accumulating within the cell's cytoplasm. This seems reasonable in the light of a burden on the cells by the product when the latter becomes accumulated within the cells.

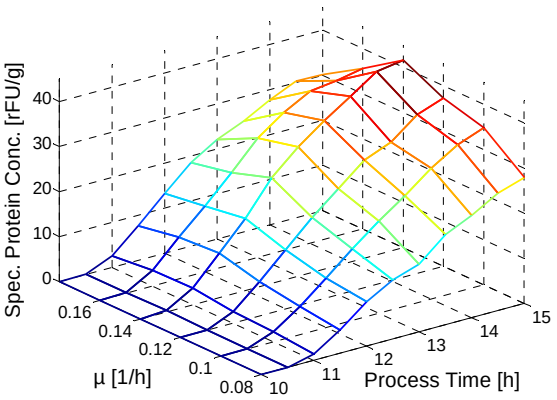

Figure 5: Specific protein concentration profiles from 6 fed batch fermentations performed with different specific growth rates after induction at 10 [h].

The optimal specific biomass growth rate of 0.14 [1/h] exactly matches with data taken from a set of experiments with the same biological system, where μ was controlled to a fixed value during the entire production phase (Jenzsch 2006). Results from 6 experiments performed with different μ_{set}-values are depicted in Figure 5. Maximal specific protein concentrations appear at the same setpoint for μ. Thus, these results confirm the optimal specific growth rate appearing in Figure 4.

The procedure proposed to determine the π(μ)-kinetics in recombinant protein manufacturing processes is quite easy to apply. Having the sensitivity equations (4,6) the training could be performed with a relatively simple MATLAB program taking advantage of its curve fitting library routine "`lsqcurvefit`".

As can be seen in the scheme depicted in Figure 1, once trained, the neuronal networks involved, only require online available input signals. They do not only supply the π and μ estimates, they also allow to estimate the current biomass x and the specific growth rate μ as well as the specific product formation rate π and the total product mass p. Hence, during subsequent experiments, this network can be applied as a software sensor for these quantities.

What can be done with the result? The resulting model with the biomass and product formation kinetics can be used in numerical optimization procedures. These will lead to optimal μ(t) control functions for these processes. Such profiles could be used directly in μ-controlled fermentation runs or, in industrial manufacturing processes, where the batch-to-batch reproducibility is an issue, in x- or cpr-controlled fermentations.

5. ACKNOWLEDGEMENTS

This work has been supported by the Ministry of Cultural Affairs of the state "Sachsen-Anhalt", Germany. We gratefully acknowledge this support.

6. REFERENCES

Galvanauskas, V., Simutis, R., Lübbert, A. (1997), Model-Based Design of Biochemical Processes: Simulation Studies and Experimental Tests, *Biotechnology Letters*, **19**, 1043-1047

Galvanauskas, V., Volk, N., Simutis, R., Lübbert, A. (2004), Design of recombinant protein production processes, *Chem.Eng.Commun.*, **191**, 732-748

Haykin, S. (1999), *Neural Networks: A Comprehensive Foundation*, 2nd ed., Upper Saddle River, NJ: Prentice Hall,

Jenzsch, M. (2006), *Advanced Monitoring & control in Microbial Cultivation Processes for Recombinant Protein Production*, Doctoral Dissertation, Martin-Luther-University Halle-Wittenberg, Germany

Jenzsch, M., Gnoth, S., Beck, M., Kleinschmidt, M., Simutis, R., Lübbert, A. (2006c), Open loop control of the biomass concentration within the growth phase of recombinant protein production processes, *J. Biotechnol.*, **127**, 84-94

Jenzsch, M., Simutis, R., Eisbrenner, G., Stückrath, I., Lübbert, A. (2006b), Estimation of biomass concentrations in fermentation processes for recombinant protein production, *Bioproc.Biosyst. Eng.*, **29**, 19-27

Jenzsch, M., Simutis, R., Lübbert, A. (2005), Application of Model Predictive Control to Cultivation Processes for Protein Production with Genetically Modified Bacteria, 511-516 in: *Computer Application in Biotechnology 2004 (CAB9),* Pons MN, van Impe JFM, eds., IFAC/Elsevier, ISBN 0 08 044251 X

Jenzsch, M., Simutis, R., Lübbert, A. (2006a), Generic model control of the specific growth rate in recombinant Escherichia coli cultivations, *J. Biotechnol.*, **122(4)**, 483-493

Kim, B.S., Lee, S.C., Lee, S.Y., Chang, Y.K., Chang, H.N. (2004), High cell density fed-batch cultivation of *Escherichia coli* using exponential feeding combined with pH-stat, *Bioproc.Biosyst.Eng*, **26**:147–150

Lee, J., Lee, S.Y., Park, S., Middelberg, A.P.J. (1999), Control of fed-batch fermentations, *Biotechnol.Adv.*, **17**, 29–48

Lee, S.Y. (1996), High cell-density culture of Escherichia coli, *Trends Biotechnol.*, **14**, 98–105

Leonard, J., Kramer, M.A. (1990), Improvement of the backpropagation algorithm for training neural networks, *Comput.Chem.Eng.*, **14**, 337-341

Levisauskas, D., Galvanauskas, V., Henrich, S., Wilhelm, K., Volk, N., Lübbert, A. (2003), Model-Based Optimization of Viral Capsid Protein Production in Fed-Batch Culture of recombinant Escherichia coli, *Bioprocess and Biosystems Engineering*, **25**, 255-262

Levisauskas, D., Simutis, R., Borvitz, D., Lübbert, A. (1996), Automatic control of the specific growth rate in fed-batch cultivations processes based on exhaust gas analysis, *Bioproc. Eng.*, **15**, 145-150

Neidhardt, F.C., Ingraham, J.L., Schaechter, M. (1990), *Physiology of the bacterial cell, a molecular approach*, Sinauer, Sunderland, MA

Picó-Marco, E., Picó, J., De Battista, H. (2005), Sliding mode scheme for adaptive specific growth rate control in biotechnological fed-batch proc-

esses, *International Journal of Control*, **78 (2)**, pp. 128-141.
Pirt, S.J. (1994), Product formation in cultures of microbes, *Pirtferm Papers*, Pirtferm Ltd., London
Riesenberg, D., Guthke, R. (1999), High-cell-density cultivation of microorganisms, *Appl.Microbiol.Biotechnol.*, 51, 422– 430
Rumelhard D. E., McClelland, J. L. (1986), *Parallel Distributed Processing*, Vol. 1, MIT Press, Cambridge, MA
Schubert, J., Simutis, R., Dors, M., Havlik, I., Lübbert, A. (1994), Bioprocess optimization and control: Application of hybrid modelling, *J. Biotechn.*, **35**, 51-68
Shioya, S. (1992), Optimization and control in fed-batch bioreactors, *Adv. Biochem. Eng. Biotechnol.*, **46**, 1
Simutis, R., Lübbert, A. (1997), Exploratory analysis of bioprocesses using artificial neural network-based methods, *Biotechnology Progress*, **13(4)**, 479-487
Soons, Z.I.T.A., Voogt, J.A., van Straten, G., van Boxtel, A.J.B. (2006), Constant specific growth rate in fed-batch cultivation of *Bordetella pertussis* using adaptive control, *Journal of Biotechnology*, published online 2006-07-10
Yoon, S.K., Kang, W.K., Park, T.H. (1994), Fed-batch operation of recombinant *Escherichia coli* containing Trp promoter with controlled specific growth rate. *Biotechnol.Bioeng.*, **43**, 995–999

7. APPENDIX

Model used for process simulation

In order to be able to test the proposed estimation procedure one needs process data from which the π(μ)-relationship is know beforehand. Only then one can determine the accuracy which can be obtained with this estimation technique.

The model used for the simulation study was kept very simple. It basically contains only 2 state variables, the biomass x and the product mass p.

$$\frac{dx}{dt} = \mu x$$
$$\frac{dp}{dt} = \pi x$$

μ is the specific growth rate, and π the specific product formation rate.

It assumes that μ is controlled by adjusting the feed rate F of the substrate to the culture accordingly. This guarantees that μ is constant for some cultivation time interval. After a sufficient amount or mass p of product is accumulated within the cells, the metabolic burden to the cell will reduce the growth rate below the μ_{set}. Hence we assume:

$$\mu = \mu_{set} \quad if \quad \mu_{set} < \mu_{max} \cdot \frac{1}{1 + \frac{p/x}{K_p}}$$

$$\mu = \mu_{max} \cdot \frac{1}{1 + \frac{p/x}{K_p}}, \quad otherwise$$

For the π(μ)-relationship we assume a fixed parametric form which is essentially a bell-shaped function. The entire model is formulated by the following equation.

$$\pi = \pi_{max} e^{-\frac{(\mu-\mu_o)^2}{K\mu}}$$

Additionally we assume that CPR, the usual carbon dioxide production rate can be measured online. Together with the measured culture weight w(t), the total CPR-mass cpr=CPR*w can be determined online. This is connected to the state variable x and μ by means of a Luedeking-Piret-type relationship.

$$cpr = Y_{cx} \cdot \mu \cdot x + m_c x$$

The following model parameters were assumed in the simulations

μ_{max} = 0.55 [1/h]; μ_o = 0.2 [1/h] ; K_p =0.05 [g/g];

K_μ = 0.05 [1/h^2]; Y_{cx}=0.6 [g/g],; m_c=0.1 [g/g/h];

π_{max} = 0.07 [g/g/h]; μ_{set} = 0.05-0.45 [1/h] ;

Trajectories of the simulation of a typical protein formation experiment are depicted in Figure 6. The simulation was performed under the following conditions:
1.The initial conditions for solving the balance equations are $x(0)$= 0.5; $p(0)$=0.
2.The integration time was taken from 0 to 9 h.
3.Induction was assumed to be at time t_{ind}=6 h.

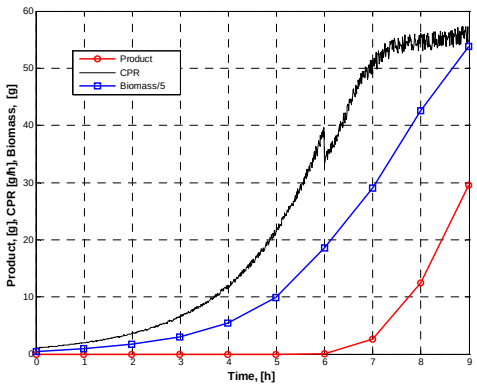

Figure 6: Biomass, protein mass and total CPR simulated by the model as a function of the cultivation time t.

As the modeling results will be used for testing the proposed procedure, the values of the state variables are computed for the time instants only where measurement values are usually taken (sampling times). For both, x and p sampling intervals of 1 h or 0.5 h were assumed. As cpr is an online measurement, much shorter sampling time intervals of 0.01 h were assumed.