Martin-Luther-University
Halle-Wittenberg
Centre of Bioprocess Engineering

Experiences with Monitoring and Control in Recombinant Protein Production

M. Jenzsch*, R. Simutis§, A. Lübbert*

** Institute of Bioengineering, Martin-Luther-University Halle-Wittenberg, Weinbergweg 22, D-06120 Halle (Saale), Germany*
§ Institute of Automation and Control Technologies, Kaunas University of Technology, Studentu g. 48, LT-3028 Kaunas, Lithuania

Background

The U.S. Food & Drug Administration (FDA) inspects most processes in order to assure the safety, efficacy, and security of human and veterinary drugs and biological products.
In the recently published FDA's PAT initiative, manufacturers have been encouraged to use the latest scientific advances in pharmaceutcal manufacturing and technology. Serveral possible tools are mentioned:

- Multivariate data acquisition and analytical tools
- Modern process analyzers or process analytical chemistry tools
- Process and endpoint monitoring and control
- Continuous improvement and control

Here we compare several possible indirect state estimation techniques for the concrete case of recombinant protein production with genetically modified *E.coli* bacteria.

Biomass Estimation Using Simple Regression Techniques

i) Multiple Linear Regression

A straightforward purely data driven approach by simply correlating the online measured signals of OUR, CPR and Base with the biomass concentration X as a function of time is linear regression.

$$X \quad a_0 \quad a_1 \; OUR \quad a_2 \; CPR \quad a_3 \; Base$$

Using MatLab's fitting function *'regress'* the regression parameters a_i can be determined.

Performance of estimation

$$RMSE_{X_{MLR}} \quad 1.97 \; \frac{g}{kg}$$

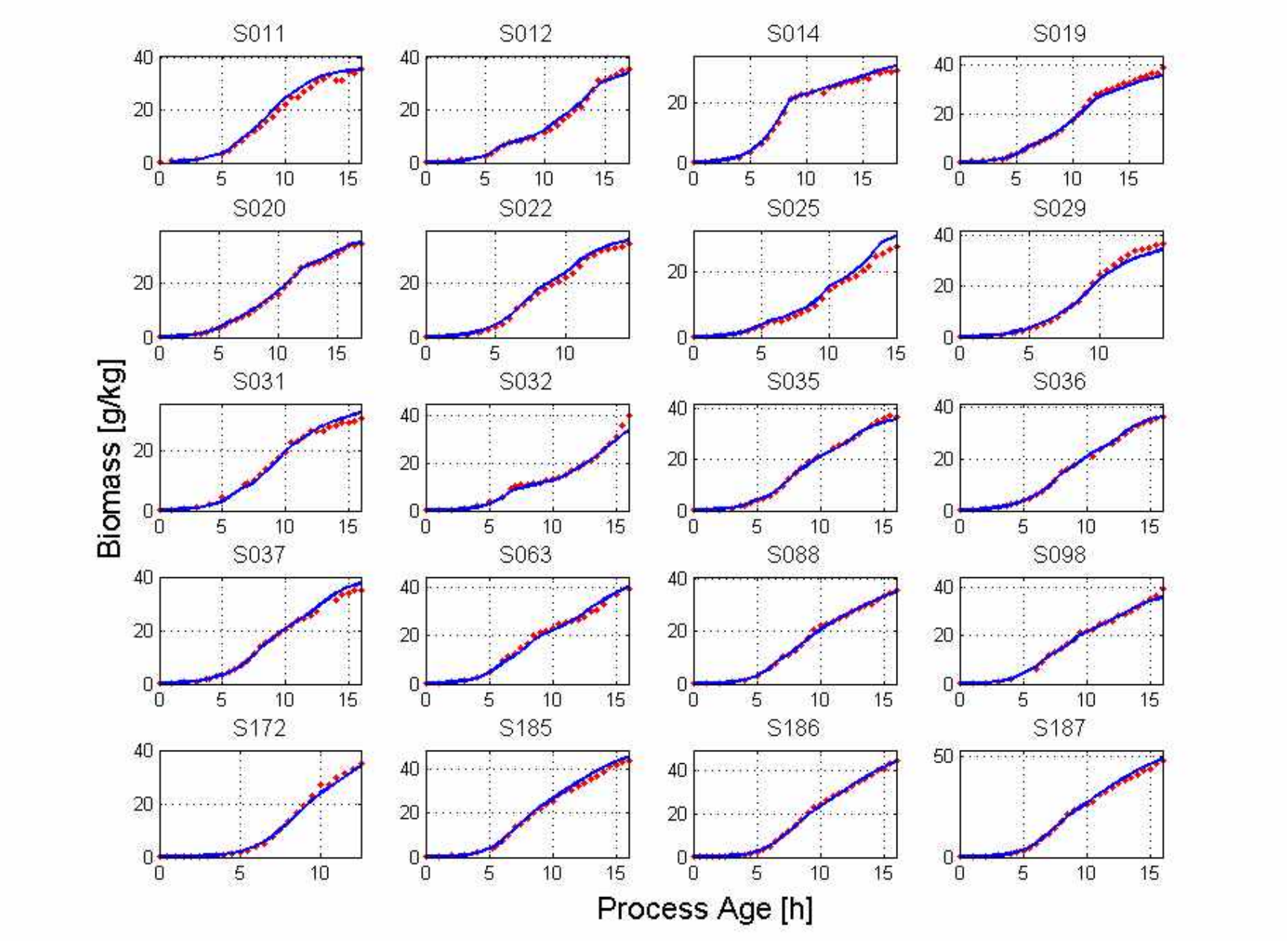
ii) Cumulative Linear Regression

The simple multiple linear regression approach can be improved using the cumulative values of OUR, CPR and Base, which show up a more direct correlation with the biomass concentration X.

$$X \quad a_0 \quad a_1 \qquad OUR \quad a_2 \qquad CPR \quad a_3 \qquad Base$$

Performance of estimation

$$RMSE_{X_{CumLin\,Reg}} \quad 1.04 \; \frac{g}{kg}$$



Biomass estimation using multiple cumulative regression depicted with a data set of 20 different *E.coli* fermentation runs. The symbols are the offline biodryweights, the lines are the online estimates of X.

iii) Cumulative Polynomial Regression

A further improvement is to be expected by a very simple nonlinear cumulative approach.
A quadratic expression can be formulated by

$$\begin{aligned} X \quad & a_0 \quad a_1 \quad {}_jOUR_j \quad a_2 \quad {}_jCPR_j \quad a_3 \quad {}_jBase_j \quad \ldots \\ & a_4 \quad {}_jOUR_j \quad {}_jCPR_j \quad a_5 \quad {}_jOUR_j \quad {}_jBase_j \quad \ldots \\ & a_6 \quad {}_jCPR_j \quad {}_jBase_j \quad a_7 \quad {}_jOUR_j \quad {}_jOUR_j \quad \ldots \\ & a_8 \quad {}_jCPR_j \quad {}_jCPR_j \quad a_9 \quad {}_jBase_j \quad {}_jBase_j \quad \ldots \\ & a_{10} \quad {}_jOUR_j \quad {}_jCPR_j \quad {}_jBase_j \end{aligned}$$

Performance of estimation

$$RMSE_{X_{CumPol\,Reg}} \quad 0.58 \; \frac{g}{kg}$$

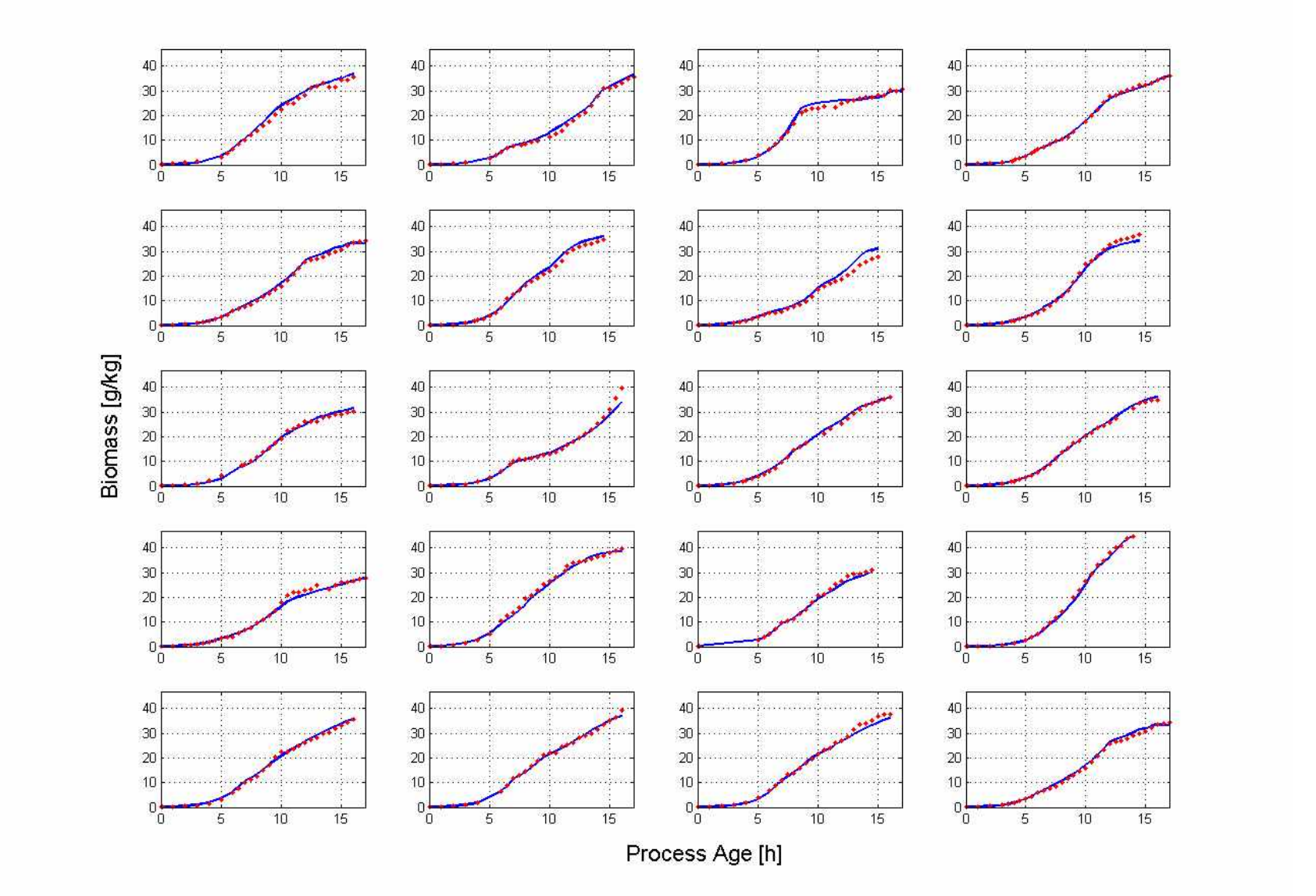
Biomass Estimation Using Artificial Neural Network

The online measured variables OUR, CPR and Base are directly mapped onto the biomass concentration X. A feedforward ANN is used with 3 input nodes (plus one bias node), and 4 hidden layer nodes, to estimate the singe output value biomass.

Performance of estimation

$$RMSE_{X_{ANN}} \quad 0.46 \; \frac{g}{kg}$$

The ANN was trained using cross validation technique: 50% of the available data were used for training and 50% data for validation.
The RMSE between the measured biomass values and the network output is 0.46 [g/kg].



Biomass estimation by means of a feedforward artificial neural network with a single hidden layer.

Summary and Discussion

Estimation Method	RMSE [g/kg]
1. Feedforward ANN	0.46
2. Cumulative polynomial regression	0.58
3. Model-based estimate on all meas. (weighted)	1.00
4. Cumulative linear regression	1.04
5. Model-based estimate on CPR meas.	1.19
6. Model-based estimate on OUR meas.	1.52
7. Model-based estimate on Base meas.	1.83
8. Multiple linear regression	1.97

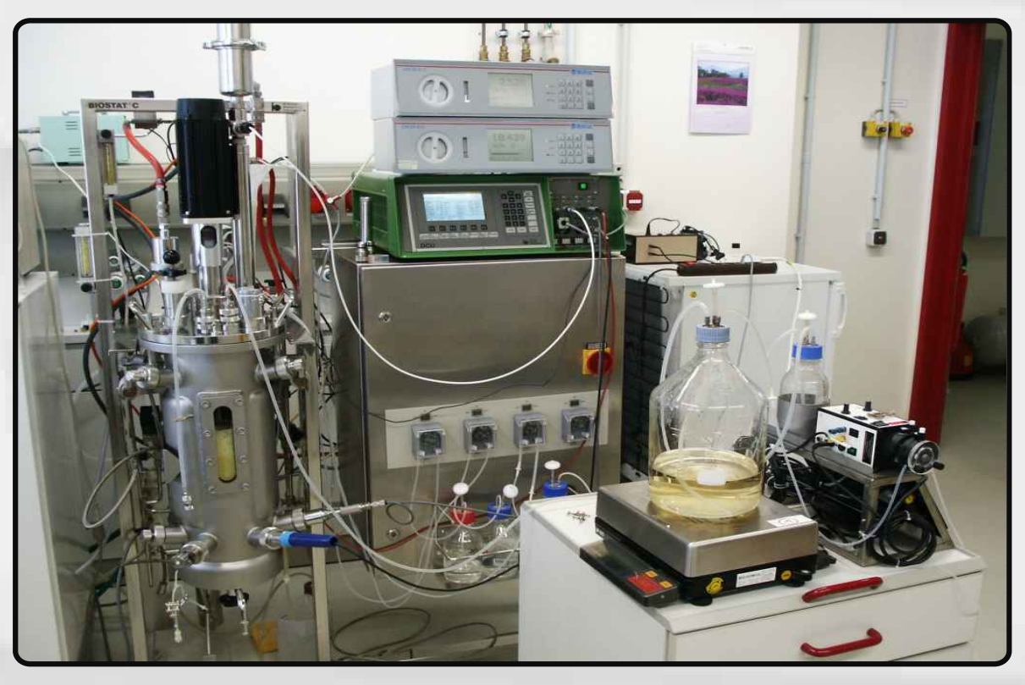
The various approaches of indirect biomass measurement in recombinant protein production fermenters show very different performance values as expressed by the root mean square deviation between the estimates and the corresponding biomass measurements compiled in the table. The main selection criteria for one or the other of these techniques are:

1. Estimation performance (accuracy, RMSE)
2. Simplicity of the computations.

At 'normal' process operation, the best estimates can be obtained with artificial neural networks.

Process Characterization

Escherichia coli BL21pET11a EGFP is used as the recombinant organism.
It is able to express the green fluorescent protein (GFP) under the control of a T7 promoter.
All experiments are performed in B.Braun's Biostat C 15-Liter laboratory fermenter.



Simple Model Based Biomass Estimation

Total biomass balance around the reactor

$$\frac{dx}{dt} \quad \frac{d\;X\;W}{dt} \qquad X\;W \quad R_X\;W$$

x ... total biomass [g]
X ... biomass concentration [g/kg]
W ... reactor weight [kg]
μ ... specific biomass growth rate [1/h]
R_x ... biomass formation rate [g/kg/h]

Relation between biomass x and oxygen uptake rate OUR is given by a Luedeking-Piret-relationship:

$$OUR \quad Y_{OX} \;\; R_X \quad m_o \; \frac{x}{W}$$

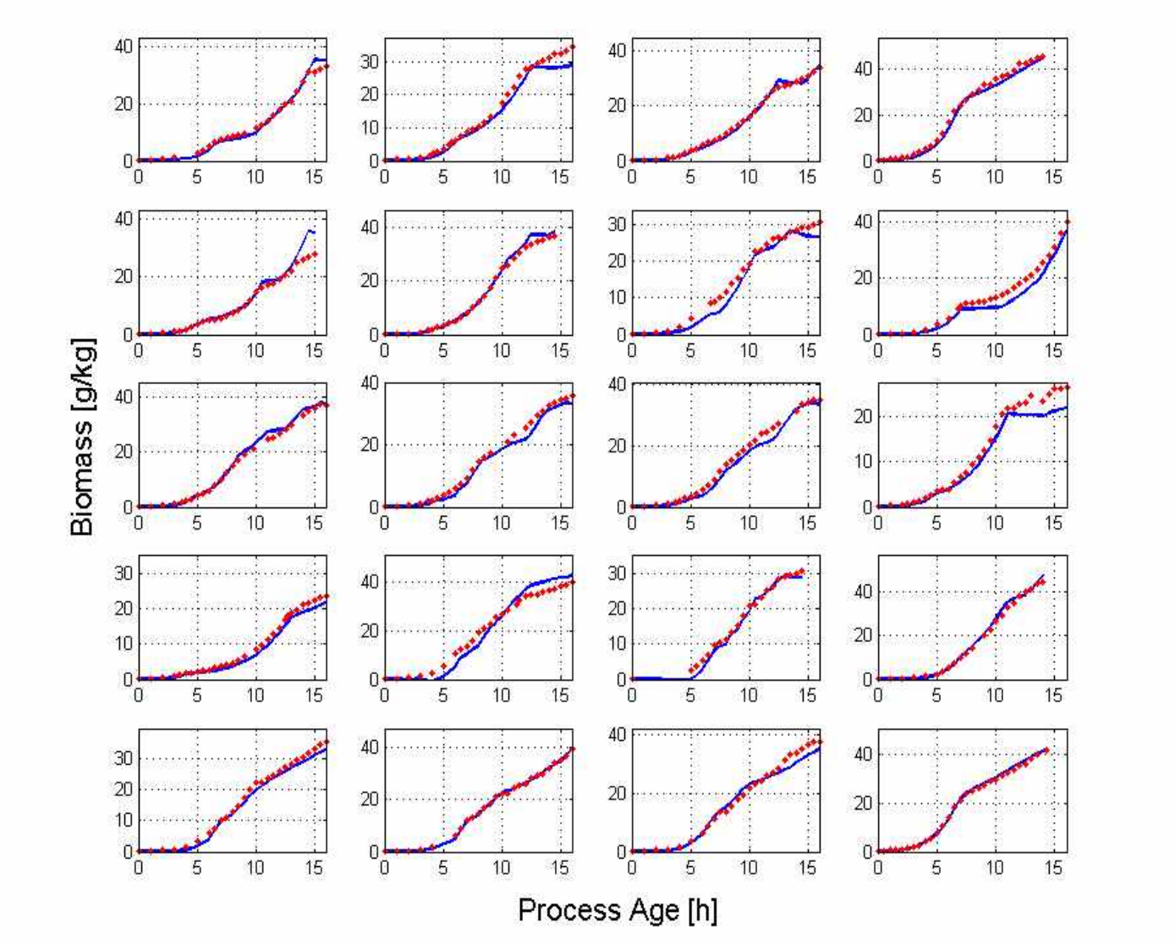
OUR ... oxygen uptake rate [g/kg/h]
Y_{OX} ... yield coefficient oxygen/biomass [g/g]
m_O ... maintenance coefficient [g/g/h]

Since W and OUR are measured, the differential equation can be solved knowing the initial biomass x_0 as well as the model parameters Y_{OX} and m_O which can be derived from experimental data.

$$x_{(t)} \quad x_{(t \quad t)} \qquad t \;\; \frac{OUR_{(t)} \;\; W_{(t)} \quad x_{(t \quad t)} \;\; m_O}{Y_{OX}}$$

Performance of estimation

$$RMSE_{X_{OUR}} \quad 1.52 \; \frac{g}{kg}$$



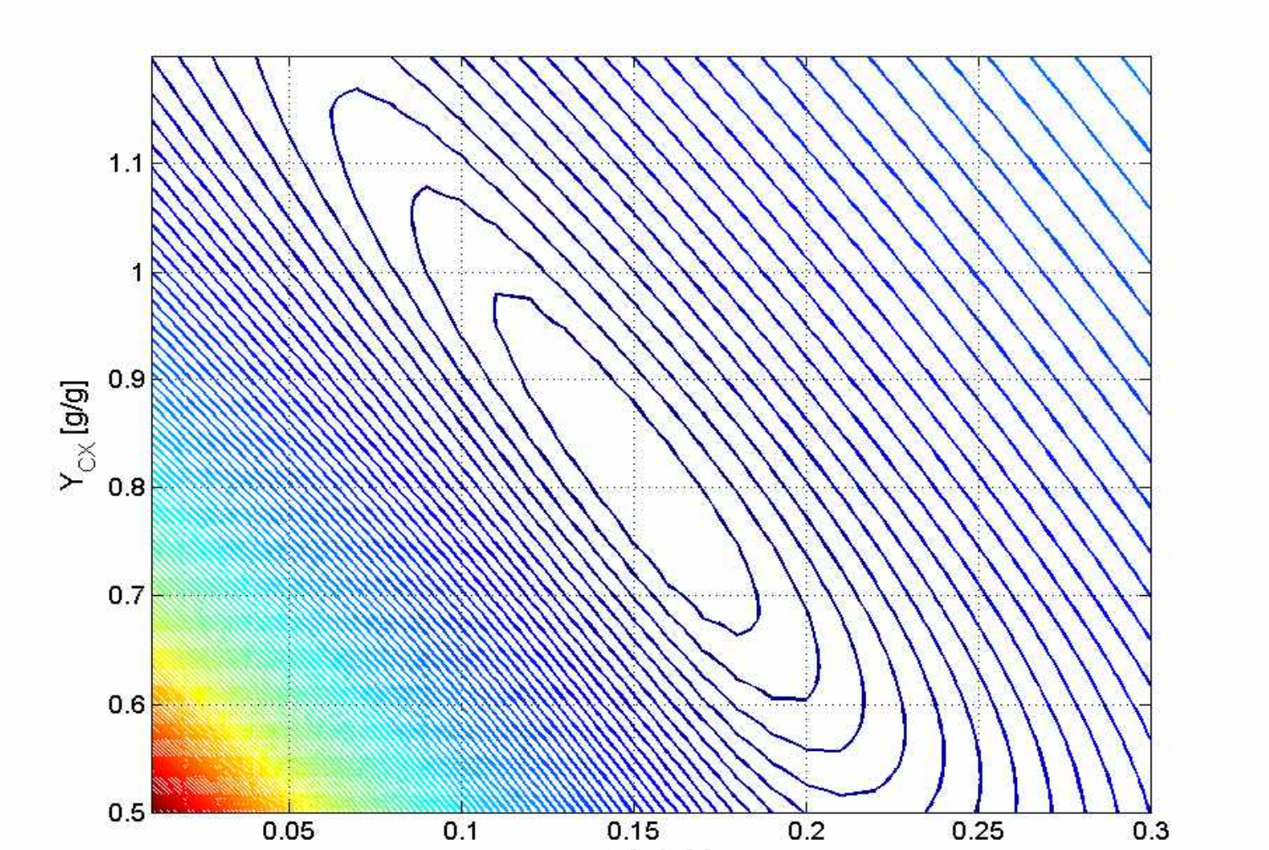
Utilizing oxygen uptake rate data for biomass estimation: Symbols represent the offline measured biomass concentrations X [g/kg], the lines are the model-based estimates.

By the same line of argumentation one can exploit other online measured data records for CPR and total ammonia consumption (Base). Equivalent Luedeking-Piret-type relationships are used:

$$CPR \quad Y_{CX} \;\; R_X \quad m_C \; \frac{x}{W}$$

$$Base \quad Y_{BX} \quad x \quad x_0$$

CPR ... carbon dioxide production rate [g/kg/h]
Y_{CX} ... yield coefficient CO_2/biomass [g/g]
m_O ... maintenance coefficient [g/g/h]
Base ... total ammonia consumption [g]
Y_{BX} ... yield coefficient ammonia/biomass [g/g]



Error surface for the identification of the model parameters Y_{CX} and m_C relating the biomass to the CPR data. The absolute minimum is difficult to obtain as different parameter combinations lead to similar identification errors.

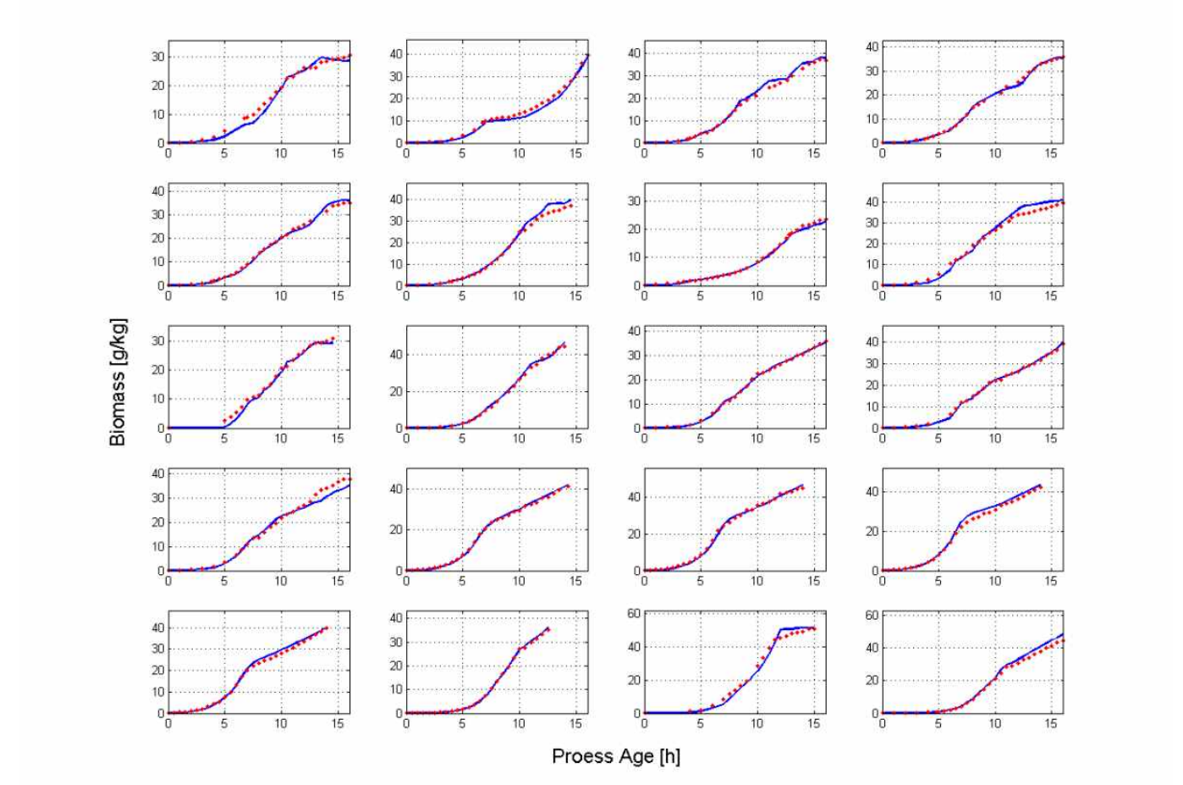
The model parameters are not independent of each other as the corresponding error surface shows.
But for the purpose of biomass estimation it is not very important to reach the exact minimum on this surface.
Thus one may use different combinations of Y_{CX} and m_C to get the practically the same results.

As usually all three measurements OUR, CPR and Base are performed simultaneously in a protein production fermenter, it makes sence to utilize them together for biomass estimation.
This can be made with a weighted average of them, where the weightings are the reciprocal values of the RMSE values of the single estimations.

$$X_{ave} \quad \frac{\dfrac{X_{OUR}}{RMSE_{OUR}} \quad \dfrac{X_{CPR}}{RMSE_{CPR}} \quad \dfrac{X_{Base}}{RMSE_{Base}}}{\dfrac{1}{RMSE_{OUR}} \quad \dfrac{1}{RMSE_{CPR}} \quad \dfrac{1}{RMSE_{Base}}}$$

Performance of estimation

$$RMSE_{X_{average}} \quad 0.996 \; \frac{g}{kg}$$



Utilizing all three online measurement variables OUR, CPR and Base for biomass estimation.

Acknowledgements

The work was performed in the project 'Technologies for the production of recombinant therapeutic proteins' by Martin-Luther-University and local protein companies. We gratefully acknowledge the financial support of this project by the Ministry of Science and Education of the German Federal state of 'Sachsen-Anhalt'.