16th European Symposium on Computer Aided Process Engineering
and 9th International Symposium on Process Systems Engineering
W. Marquardt, C. Pantelides (Editors)

Stochastic Grey Box Modeling of the enzymatic biochemical reaction network of *E. coli* mutants

Florin Paul Davidescu[a], Henrik Madsen[b], Michael Schümperli[c], Matthias Heinemann[c], Sven Panke[c] and Sten Bay Jørgensen[a] *.

[a]CAPEC, Department of Chemical Engineering, Technical University of Denmark, Building 227, DK 2800, Kgs, Lyngby, Denmark

[b]Department of Informatics and Mathematical Modeling, Technical University of Denmark, Building 321, DK 2800, Kgs, Lyngby, Denmark

[c]Institute of Process Engineering, Bioprocess Laboratory, ETH Zürich, Sonneggstr. 5, CH 8092, Zürich, Switzerland

This paper describes the application of a gray-box stochastic modeling framework for developing stochastic state space models for dynamic systems based on combining first principle models and experimental data. The framework is used to develop reliable predictive models for a biochemical reaction network isolated from *E. coli* mutants. The modeling purpose is to use the model to identify the bottlenecks in the reaction network to enable optimizing the production of the desired product through genetic manipulation.

1. Introduction

There is an increasing interest in producing complex fine chemicals and intermediates in the pharmaceutical industry using biochemical synthesis. Up to now, only one or a few biotransformation steps are involved in complex synthesis problems in industry, although enzymes are widely known as being specific, fast and working under mild conditions. To develop a purely enzymatic synthesis for complex molecules from completely different substrates, large reaction networks are necessary. One way to construct such a functional network is the System of Biotransformations (SBT). The SBT is based on one single organism's metabolic network containing the synthesis path including cofactor regeneration reactions in an isolated manner. Thereby, the SBT is performed as cell free extract in the production phase, combining the easy handling of a viable culture with the advantages of *in vitro* biotransformations [5]. The complexity of such large biochemical reaction networks involves a large number of reaction steps with many metabolites and enzymes, each one of them playing different roles as biocatalysts and/or as feed-forward and feed-back regulators. The general goal of this study is to identify the limitations and bottlenecks, to reduce them and to optimize the productivity of the selected reaction network. The workhorse of the de-bottlenecking and optimization process is a model describing the

*Corresponding author

biochemical reaction network with good long term prediction properties. For this particular application, the key product is Dihydroxyacetone phosphate ($DHAP$). $DHAP$ is an important precursor for the production of phosphorylated, non natural carbohydrates. Thereby, the DHAP-producing SBT contains all the reactions of the glycolysis, leading to a system of high dynamic and complexity. Therefore, it is not realistic to develop a "perfect model" from first principle engineering methods. For this reason, in this work the gray-box stochastic model development framework [1] will be used to develop a stochastic state space model. The purpose of this paper is to describe the work-flow driving application of the gray-box stochastic modeling framework for development of a kinetic model for a batch reaction network.

2. Stochastic gray-box modeling background methodology

The gray-box stochastic modeling framework, [1] was originally developed for fed-batch cultivations but it can be employed for modeling of complex nonlinear dynamic processes as well. The framework combines different mathematical and statistical tools and assists the model development in a systematic way. First, the model equations are derived from first engineering principle and then completed with diffusion terms/functions to obtain the Stochastic Differential Equations. The diffusion terms accounts for model errors and/or for the un-modeled effects. Formulating the diffusion terms by only having the diagonal terms in the square matrix of the diffusion terms ([1]) it is possible to improve the process model in a systematic way. The measurement equations include the measurements errors as well, thus in this approach it is possible to make a clear distinction between the measurement errors and process noise or model error. In the next step the set of unknown parameters together with the diffusion terms and the variances of the measurements are estimated from experimental data using a maximum likelihood or a maximum a-posteriori method. In the estimation method it will just be mentioned that the solution of the stochastic differential equation system and the innovation terms that appears in the maximum likelihood function is based on an Extended Kalman Filter. The model is un/falsified using different statistic tests. Then the model is reformulated and the iterations continued until the model is un-falsified using available data or all the information contained in the data with respect to the dynamics is exhausted. A workflow diagram of the whole modeling framework is given in figure 1.

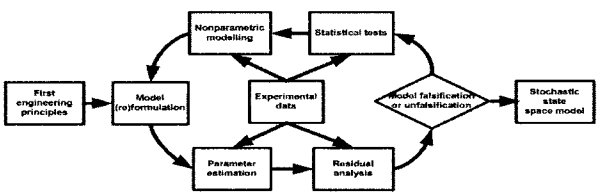

Figure 1. Stochastic gray-box modelling framework, from [1]

3. Experimental data and procedures

The experiments have been conducted by ref [5]. In phase I, fed-batch (semi-batch) fermentations of *E. coli* W3110 *tpi* are conducted until the optical density (OD600) in the bioreactor reaches a preset value. The broth is centrifuged and the cells are resuspended in SBT-buffer (100 mM HEPES, 0.84 mM KCl, 1 mM ZnSO4 and at pH = 7). The cells are disrupted by high-pressure homogenization. The remaining solids are eliminated by centrifugation/filtration and the liquid extract is recovered. The total protein concentration is determined Bradford and adjusted to the desired concentration by dilution with SBT buffer. The liquid extract contains the enzymes and compounds present in the cell at the time when the fermentation was stopped. In phase II, a volume of 5 ml of SBT extract is used for each experiment. Defined amounts of *Hexokinase*, (HK) and $Lactate - DH$ as well as ATP and NAD^+, are added. The reactions are initiated by glucose. Samples are collected according to a previously defined time plan. The experiments are terminated after 300 minutes. First, the proteins are removed by precipitation with HCl followed by centrifugation. The samples are analyzed by enzymatic assays. Glucose and glucose-6-phosphate ($G6P$) are determined together by addition of both HK and glucose-6-phosphate-dehydrogenase to form $NADPH$, which is determined spectrophotometrically. $DHAP$ is determined by addition of glycerol-3-phosphate-dehydrogenase and measuring the $NADH$ consumption spectrophotometrically. A series of four experiments has been used for the model development; three for parameter estimation and one for model validation.

4. Model development for an SBT isolated from *E. coli* mutants

In the model formulation the first measurement equation was assigned to glucose. The first step in the model development is model formulation. In order to formulate a model the existing biochemical reaction network in *E-coli* is presented with focus on the reactions around the product of interest ($DHAP$) considering the genes which are knocked out. The simplified biochemical reaction network used for model development is depicted in figure 2. The reaction between $DHAP$ and $G3P$ does not take place since the *tpi* gene i.e. responsible for the expression of the enzyme catalyzing the reaction has been knocked-out. For the current version of the model all the reactions from glucose to fructose-1,6-biphosphate, ($F16B$) were lumped into a single reaction r_1 . The second reaction considered is the reaction from $F16B$ to $G3P$ and $DHAP$, r_2 catalyzed by aldolase. The reactions consuming the $G3P$ down to pyruvate in the central carbon metabolism were all lumped into one single reaction r_3. The reaction producing lactate from pyruvate was included as reaction r_4. The reason to include these two reactions is that it is desirable to account for the consumption-production of co-factors ATP and NAD^+. The model consists of dynamic mass balances for all the species involved in the four reactions plus one for each of the two co-factors. The model equations eq. 1–10 have been completed with the diffusion terms as mentioned above.

In this first model formulation it has been considered that the reaction rates $r_1 - r_4$ are constant and then estimated together with the model parameters and with the initial values of the states

and the measurement variances $S_1 - S_2$.

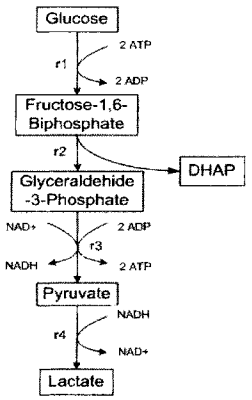

Figure 2: Reaction network used for model development

$$dc_{GL} = -r_1 + \sigma_{11} \cdot dw \quad (1)$$
$$dc_{F16B} = r_1 - r_2 + \sigma_{22} \cdot dw \quad (2)$$
$$dc_{DHAP} = r_2 + \sigma_{33} \cdot dw \quad (3)$$
$$dc_{G3P} = r_2 - r_3 + \sigma_{44} \cdot dw \quad (4)$$
$$dc_{PYR} = r_3 - r_4 + \sigma_{55} \cdot dw \quad (5)$$
$$dc_{LAC} = r_4 + \sigma_{66} \cdot dw \quad (6)$$
$$dc_{ATP} = -2 \cdot r_1 + 2 \cdot r_3 + \sigma_{77} \cdot dw \quad (7)$$
$$dc_{NAD} = -r_3 + r_4 + \sigma_{88} \cdot dw \quad (8)$$

Measurements equations:

$$y_{GLC} = c_{GL} + e, e \in N(0, S_{11}) \quad (9)$$
$$y_{DHAP} = c_{DHAP} + e, e \in N(0, S_{22}) \quad (10)$$

The parameters have been estimated using three of the available data sets and gray-box stochastic modeling software CTSM [2]. After the estimation step, significance tests have been performed for the estimates and the parameters estimate correlation matrix has been calculated as well. Several parameters were insignificant or highly correlated, therefore some parameters have been fixed and the remaining parameters reestimated. The fit for one step ahead prediction as well as pure simulation has been plotted in figure 3 and the numerical results for the parameters are given in table 1. The fit for pure simulation data, clearly indicates that the model needs improvement. Inspecting the diffusion terms in table 1 shows that the corresponding $\sigma_{11} - \sigma_{33}$ are significant, thus the drift terms of these equations are deficient. Following the methodology mentioned above r_1 and r_2 are included in the state vector one at a time. After including r_1 and r_2 consecutively, the parameters have been reestimated.

At this step in the model development it is necessary to see how we should model the reaction rate r_1. Before the nonparametric methods are applied, it is necessary to reconstruct the states of the model with r_1 as extra state. This is done by applying the extended Kalman filter (EKF) with the parameter estimates obtained after r_1 was included as a new state. The nonparametric tools e.g. additive models [1,3] are applied in order to identify the shape of the kinetic expression for the reaction rate r_1. Analyzing the reaction network in figure 2, the reaction rate may be a function of glucose, ATP, $F16B$ concentrations. The graphical results (not shown) indicate that the dependence of r_1 versus c_{GLC} appears to be the most significant. In figure 4 this dependence solely, is shown. The same steps have been applied for the second reaction rate, r_2. In this case r_2 depends of c_{G3P} and c_{DHAP} as shown in figures 5–6, while the dependence of c_{F16B} seems to be insignificant (not shown).

Considering the shape in figure 4 for the functional dependence of r_1 on glucose, then reaction rate r_1 can be modeled using Monod kinetics (eq. 11). The parameters are reestimated assuming Monod kinetics (eq. 11) for r_1. The one-step ahead prediction as well

Table 1
Estimated parameters, standard deviation, t-test, and the in/significance

Name	Estimate	Std. dev.	t-score	signif.?
c_{GLC0}	1.0072E+01	6.0307E-01	16.7004	yes
c_{DHAP0}	8.5785E-02	1.2678E-01	0.6767	no
c_{ATP0}	1.1619E+01	7.4897E+01	0.1551	no
r_1	6.5923E-02	1.1838E-02	5.5688	yes
r_2	3.9095E-02	2.5091E-03	15.5811	yes
σ_{11}	4.9197E-01	3.6057E-02	13.6442	yes
σ_{22}	1.0000E-02	3.4324E-04	29.1343	yes
σ_{33}	1.0440E-01	7.4442E-03	14.0240	yes

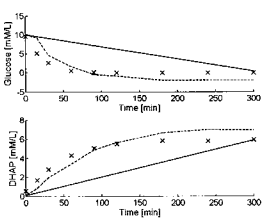

Figure 3. c_{GLC} and c_{DHAP} vs. time, exp. data: x, pure simulation: continuous line, one step ahead pred.: dashed

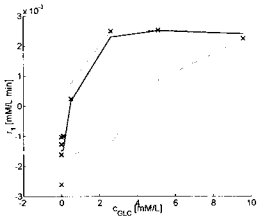

Figure 4. r_1 vs. c_{GLC}, exp. data: x, local fit: continuous line and 95% conf. intervals: dashed

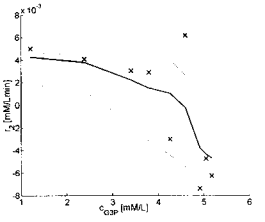

Figure 5. r_2 vs. c_{G3P}, exp. data: x, local fit: continuous line and 95% conf. intervals: dashed

as the pure simulation of the model has improved considerably for the first measurement, see figure 7.

$$r_1 = r_{1max} \cdot \frac{c_{GLC}}{K_{s1} + c_{GLC}} \tag{11}$$

After modeling reaction rate r_1 (eq. 11), reaction rate r_2 is included again as a new state and the parameters re-estimated. Using the new set of parameters, states estimation and nonparametric modeling tools are applied again. The individual dependences of r_2 on the c_{G3P}, c_{DHAP} and c_{F16B} seems to be similar with the data obtained before modeling r_1 Literature references ([4]), mentions that the reaction rate is related to the equilibrium constant, thus a dependence of the reaction rate on the difference between the forward and the backward reaction was investigated by regressing a dependence on the $c_{F16B} - c_{G3P} \cdot c_{DHAP}$ (not shown). Again, the dependence looks like a Monod term but

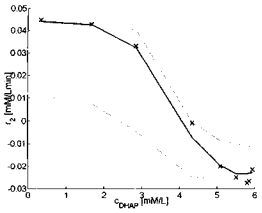

Figure 6. r_2 vs. c_{DHAP}, exp. data: x, local fit: continuous line and 95% conf. intervals: dashed

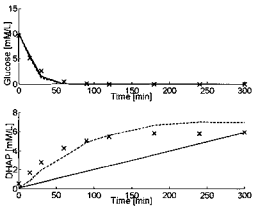

Figure 7. c_{GLC} and c_{DHAP} vs. time, exp. data: x, pure simulation: continuous line, one step ahead pred.: dashed

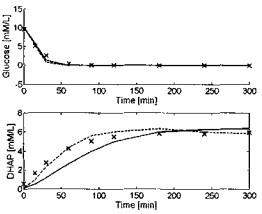

Figure 8. c_{GLC} and c_{DHAP} vs. time after modeling r_1 and r_2, exp. data: x, pure simulation: continuous line, one step ahead pred.: dashed

with this abscissa. The parameters have been reestimated, and the fit has been improved for the second measurement as can be seen in figure 8.

5. Conclusions

A gray-box stochastic model for a *E. coli* extract reaction network is under development. The model development is performed by the application of the gray-box stochastic modeling framework proposed by Kristensen [1]. The current results looks promising and now the focus is in developing and using specific experiments to provide information on different reaction kinetics in the metabolic network. Once the complete reaction network presented in figure 2 is reasonably modeled, productivity optimization will be investigated.

REFERENCES

1. Niels Rode Kristensen, Henrik Madsen and Sten Bay Jørgensen, A method for systematic improvement of stochastic gray-box models, Comp. and Chem. Eng., 28 (2004), 1431-1449.
2. Niels Rode Kristensen, Henrik Madsen and Sten Bay Jørgensen, Parameter estimation in stochastic gray-box models, Automatica, 40 (2004), 225-227 .
3. Trevor Hastie and Robert Tibshirani, Bayesian backfitting, Stat. Sci. 15 (2000) 196-213.
4. Christophe Chassagnole and others, Dynamic modeling of the central carbon metabolism of *Escherichia coli*, Biotechnology and Bioengineering, 79, No. 1 (2002) 53-73.
5. Michael Schümperli, Matthias Heinemann, Anne Kümmel, and Sven Panke, in preparation, 2005