Hybrid model based identification of biochemical processes using sensitivity equations approach

VYTAUTAS GALVANAUSKAS, RIMVYDAS SIMUTIS
Department of Process Control
Kaunas University of Technology
Studentų 50, LT-51368 Kaunas
LITHUANIA
http://www.personalas.ktu.lt/~vygal

Abstract: An approach for hybrid model parameter identification based on sensitivity equations approach is tested on a hybrid model for biosurfactant production process. The computer simulations were performed using software tools created in *Matlab (The MathWorks, Inc.)* environment. Possible difficulties and ways of practical implementation of the analyzed identification algorithms for the identification of complex hybrid models for biochemical processes are outlined. The hybrid model identification results are discussed.

Key-Words: Hybrid models, artificial neural networks, parameter identification, sensitivity equations

1 Introduction

In the context of this paper hybrid models are representations of biotechnological processes incorporating mass balance systems for key components, and mechanistic sub-models and/or artificial neuron networks (ANNs) and/or fuzzy subsystems for modeling of specific reactions rates.

Due to rather complex reaction systems in modern biotechnology and chemistry [2, 4, 5, 8, 9, 11], it is not possible to describe all important phenomena in detailed mechanistic models. Hence, alternative methods which are able to bridge some gaps in special biochemical or engineering knowledge are necessary. Usually, engineers first make use of data from production processes to cope with a particular problem; using so called engineering correlations, which are data-driven descriptions. Their aim is to formulate reproducible relations between the variables that can be manipulated in the bioreactor and the process performance criteria that determine the quality of the biochemical transformations to be carried out (product amount, its quality etc.). Additionally, they make use of more condensed representations of their data by representing experience-related knowledge in the form of rules-of-thumb. Further extensions of the data and knowledge driven approaches used in bioengineering are artificial neural networks and fuzzy systems. Hybrid modeling has been used in chemical and biochemical engineering for many years (e.g., [11]). In the recently published literature, several interesting examples demonstrate that such hybrid combinations of artificial neural networks, mechanistic kinetics and mass balance equations lead to considerable advantages [2, 3, 4, 9, 11].

In practical applications, the artificial neural network sub-models in such hybrid model require some attention concerning the appropriate training procedures: the most often used type of network in this case is the recurrent artificial neural network. This technique can also be found in commercial software packages. Its decisive characteristic is that it requires input data to be taken with a constant time increment. Where such data is available the recurrent neural network algorithms perform quite well. However, in biotechnology, where we must deal with many off-line or quasi-off-line data, such regular data records must be generated using interpolation techniques. Unfortunately, using such an interpolation we loose not only accuracy but we also may hook some artificial disturbances. Furthermore, in biotechnology we are most often interested in reaction rates, as they are known to be the key quantities allowing to evaluate the behavior of conversion processes. Rates must be determined from the measured or estimated concentrations by differentiation. In this respect noisy signals lead to even more distorted estimates of the specific rates. Therefore, many classical identification methods [10] may fail or do not lead to a solution of required precision.

Hence, one needs robust and effective process techniques for identification of hybrid models in bioprocess engineering while dealing with complex biotechnological processes that can not be modeled with the necessary precision using only simple engineering correlations and mechanistic models.

The described problems can significantly be reduced with the „sensitivity equations approach" developed for training of the hybrid models [11].

2 Identification of hybrid model parameters using sensitivity equation approach

A combination of a mathematical model represented by a set of nonlinear differential equations, mechanistic specific reaction rates expressions and an artificial neural network is shown in Fig. 1. The main problem arising with such a combination is that the usual training (parameter identification) procedures may not work or their performance is significantly reduced. Hence different training procedures must be used.

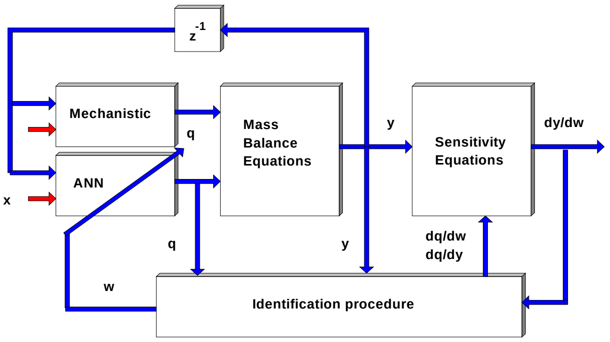

Fig. 1. Identification of ANN parameters using sensitivity equations approach

In order to show the distinct steps of the technique, the main structural blocks of the procedure are shown in Fig. 1. The differential equations system for mass balance can be written in the following generalized form:

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}\,(\mathbf{y}(t), \mathbf{x}(t), \mathbf{q}(y(t), \mathbf{w}))\,, \qquad (2.1)$$

where ***f*** is a nonlinear vector function of the system inputs ***x(t)***, outputs ***y(t)***, and biochemical reaction rates ***q(y(t),w)***, part of which we assume to be represented by means of an artificial neural network(s).

In order to train the neural network part of the hybrid process model, pairs of input/output data vectors, as measured at the real plant are to be used. The training requires that the ANN weights ***w*** should be determined in such a way that the sum of the squared deviations, *J*, between the output data $\mathbf{y}_i$ predicted by the hybrid model and the corresponding process data $\mathbf{y}_{i,exp}$ becomes minimal:

$$\frac{\partial J}{\partial \mathbf{w}_n} = \sum_i (\mathbf{y}_i - \mathbf{y}_{i,exp}) \frac{\partial \mathbf{y}_i}{\partial \mathbf{w}_n} \rightarrow min \qquad (2.2)$$

The usual way to minimize *J* is to use gradient methods for adapting of ANN weights:

$$\mathbf{w}_{n+1} = \mathbf{w}_n - g\,\frac{\partial J}{\partial \mathbf{w}_n}, \qquad (2.3)$$

where *g* is again factor ,determined by experience. By forming the derivative of *J* with respect to ***w***, we find

$$\frac{d}{dt}\frac{\partial \mathbf{y}}{\partial \mathbf{w}_n} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{w}_n} + \frac{\partial \mathbf{f}}{\mathbf{w}_n} \qquad (2.4)$$

Hence, for ANN identification, it is necessary to determine the derivatives $\partial \mathbf{y}_i/\partial \mathbf{w}_n$. In the hybrid system, however, system outputs ***y*** are different from the outputs of the neural network, therefore a training of ANN with application of conventional backpropagation technique is impossible. One of the possibilities to efficiently solve the problem is the application of the sensitivity equation approach described in [11]. In order to calculate gradients $\partial \mathbf{y}_i/\partial \mathbf{w}_n$ it is necessary to differentiate equation (2.1) with respect to weights $\mathbf{w}_n$, leading to

$$\frac{d}{dt}\frac{\partial \mathbf{y}}{\partial \mathbf{w}_n} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{w}_n} + \frac{\partial \mathbf{f}}{\mathbf{w}_n} \qquad (2.5)$$

Eq. (2.5) is a differential equation for the unknown gradients $\partial \mathbf{y}/\partial \mathbf{w}_n$ required in the ANN training algorithm. The initial condition at time $t = 0$, necessary to solve the equation (2.5), is known to be

$$\frac{\partial \mathbf{y}(t=0)}{\partial \mathbf{w}_n} = 0 \qquad (2.6)$$

When the specific rate expressions ***q*** are represented by feedforward ANN, the matrix $\partial \mathbf{f}/\partial \mathbf{w}_n$ can be readily computed by using the standard backpropagation method. This fact makes ANN attractive for mapping of specific reaction rates ***q***. In this hybrid modeling structure, the ANN is trained directly on the original concentration data (off-line measurements). It is not necessary to calculate derivatives of the smoothed concentration values as in the differential approximation technique. The training procedure for the combined system of the ANN and the mass balance equations can be regarded as an optimization procedure that minimizes the deviations of the estimated values of the process variables from the measured ones.

The general procedure of solving the ANN training problem is:

1. Set up of the mass balance eq. (2.1).
2. Choosing the structure and initial weights of ANN.
3. Establishing the sensitivity eqs. (2.5).
4. Integrating the sensitivity equations and determining $\partial y/\partial w_n$.
5. Making use of $\partial y/\partial w_n$ in the same way as with classical ANN for single example learning or batch learning.

In order to reduce training time for different applications, it can be carried out with a combination of the classical backpropagation technique and a conjugate gradient optimization method. In order to achieve better extrapolation properties of the ANN, the cross validation technique during the ANN training procedure must be used. The concept of cross validation is that after training using a given sample of data ("training set") the quality of the process representation is evaluated using a different set of data ("validation set"). The root mean square error (RMSE) between predicted and measured outputs in the "validation set" is referred to as the "validation error".

Since these modeling and identification procedures are relative simple, they additionally simplify the quick adaptation of the model to changing process conditions. This is why these techniques are so attractive for industrial applications.

3 . Hybrid model and sensitivity equations

As an accompanying example of a complex biochemical process a biosurfactant production process was taken. The biosurfactant production process in *Azotobacter vinelandii 21* fed-batch culture, its materials and methods, and the structure of the corresponding process hybrid model is given in more details elsewhere [9]. Investigations of biosurfactant production by various microorganisms indicate that biosynthesis of biosurfactant is highly influenced by the sources of carbon, nitrogen and phosphorous [6, 7]. Hence, besides the dynamical models for biomass and product concentrations, the most important process variables that influence the biosurfactant biosynthesis process are concentrations of the above components of nutrient media therefore should be modeled. Additionally, one should pay special attention to the biomass specific growth rate that proved to have significant influence on the product formation rate. Some negative inhibiting influence of the biosurfactant concentration on its specific production rate was also observed in the experimental data. Finally, the culture broth weight in bioreactor should be modeled in order to be able to account for mass flows into/out of the bioreactor during the operation in a fed-batch mode.

Based on the above considerations, the following system of mass balance equations was build:

$$\frac{dx}{dt} = \mu(x,s_1)x - (F_s + F_b)\frac{x}{V} \tag{3.1}$$

$$\frac{dp}{dt} = q_p(\mu,p,s_2,s_3)x - (F_s + F_b)\frac{p}{V} \tag{3.2}$$

$$\frac{ds_i}{dt} = -q_{s,i}(\mu,q_p)x - (F_s + F_b)\frac{s_i}{V} + F_s\frac{s_{i,f}}{V}, \tag{3.3}$$

$$i = 1,\ldots,3$$

$$\frac{dV}{dt} = F_s + F_b - F_{smp}, \tag{3.4}$$

where x, p, s_i are concentrations of biomass, biosurfactant and substrates (i is number of substrate component taken into consideration), respectively; V is volume of culture broth in bioreactor; μ, q_p, q_{si} are the specific rates of biomass growth, biosurfactant production and substrate components consumption, respectively; F_s is feed rate of substrates (manipulated variable); s_{if} are substrate component concentrations in feeding solution; F_b is base feeding rate for *pH* control purposes, F_{smp} is sampling rate; and t is time.

Identification of structure of functional relationships for the specific rates μ, q_p, q_{si} of biochemical reactions was based on analysis of experimental data [9].

The specific biomass growth rate is a function of the principal substrate glucose (component s_1) and the biomass concentrations x. The following functional relationship was found to be adequate for prediction the biomass growth dynamics [1]:

$$\mu(s_1,x) = \mu_{\max}\frac{s_1}{k_x x + s_1 + s_1^2/k_i} - k_{xx}x, \tag{3.5}$$

where μ_{max}, k_x, k_i, k_{xx} are model parameters.

The specific biosurfactant production rate q_p is related to the biomass growth dynamics (μ) and the concentrations of biosurfactant and substrate components. Due to complexity of the functional relationship $q_p(\mu, p, s2, s3)$ it is expressed by a feedforward sigmoid artificial neural network (ANN) containing 4 inputs, 2 nodes in a hidden layer, and 1 output (Fig. 2). Taking into account the bias values the ANN results in 13 free tunable model parameters.

The following substrate components are considered to be essential for biosurfactant production: ammonia nitrogen concentration (s_2) and phosphate phosphorus concentration (s_3). The Pirt type functional relationship [1] is applied for modeling the specific rate of the glucose consumption:

$$q_{s1}(\mu) = \frac{1}{Y_{xs}}\mu + m, \tag{3.6}$$

where Y_{xs}, m are the model parameters. The identified values of the parameters in the equations (3.5, 3.6) are given elsewhere [9].

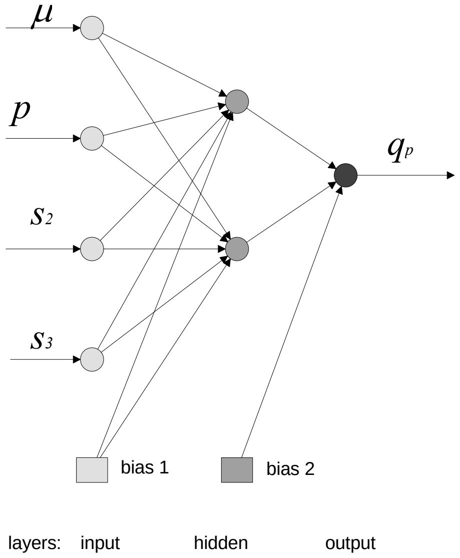

Fig. 2. Structure of the ANN for modeling of biosurfactant specific synthesis rate

Due to *a priori* unknown type of nonlinearity of functional relationships the specific consumption rates of the ammonia nitrogen (q_{s2}) and phosphate phosphorus (q_{s3}) are also modeled by means of 2 ANNs containing 2 inputs (μ and q_p), 2 nodes in hidden layer and 1 output respectively. Taking into account the bias values, each of the two ANNs results in 9 free tunable model parameters.

Additionally to the hybrid model differential equations, the corresponding sensitivity equations for the training of the ANNs were built. E. g., in order to train the ANN of the biosurfactant biosynthesis rate, the following steps were performed.

In case of specific biosurfactant synthesis rate modeling, the equation of interest is eq. (3.2), the right hand-side of which is the element of vector function $\boldsymbol{f}(\boldsymbol{y}(t), \boldsymbol{x}(t), \boldsymbol{q}(y(t), \boldsymbol{w}))$, where *p(t)* corresponds to *y(t)*, *F*$F_s(t)$ – to *x(t)*, and $q_p(t)$ – to *q(t)*. ***w*** is the vector of weights of an ANN.

With respect to eq. (5), the sensitivity equations for the ANN weights of biosurfactant specific synthesis rate can be written in the following way:

$$\frac{d}{dt}\frac{\partial p}{\partial \mathbf{w}_n} = \frac{\partial f}{\partial p}\frac{\partial p}{\partial \mathbf{w}_n} + \frac{\partial f}{\partial \mathbf{w}_n}, \tag{3.7}$$

where

$$\frac{\partial f}{\partial p} = \frac{\partial q_p(\mu, p, s_2, s_3, \mathbf{w}_n)}{\partial p}x - \frac{(F_s + F_b)}{V}, \tag{3.8}$$

and

$$\frac{\partial f}{\partial \mathbf{w}_n} = \frac{\partial q_p(\mu, p, s_2, s_3, \mathbf{w}_n)}{\partial \mathbf{w}_n}x, \tag{3.9}$$

which can be calculated in a straightforward way for the ANN of the structure, known in advance.

Finally, the sensitivity equations system is:

$$\frac{d}{dt}\frac{\partial p}{\partial \mathbf{w}_n} = \Big(\frac{\partial q_p(\mu, p, s_2, s_3, \mathbf{w}_n)}{\partial p}x - \\ - \frac{(F_s + F_b)}{V}\Big)\frac{\partial p}{\partial \mathbf{w}_n} + \frac{\partial q_p(\mu, p, s_2, s_3, \mathbf{w}_n)}{\partial \mathbf{w}_n}x \tag{3.10}$$

In the same way, the sensitivity equations for the other 2 ANNs are established.

4 Software implementation

The main aim during the creation of the software tool for model based identification and optimization of biotechnological processes was to ensure the maximal flexibility and user-friendly environment.

The process model was implemented using the elements from the standard and extensive user-defined libraries for *Matlab (The MathWorks, Inc.) Simulink* environment together with the embedded *Matlab* functions. The created user-defined libraries consist of the biotechnology and biochemistry specific expressions for modeling of specific reactions rates, mass balance equation systems and flows. By simple “dragging and dropping” one can add necessary blocks from the libraries to the model. The GUI was programmed using Matlab *guide* and allows to flexibly and efficiently change and manage model and identification algorithms specific options and settings without going deeper into the script programming details. The numerous identification methods and algorithms were implemented using standard *Matlab* functions and subroutines. Additionally, part of the software tool was programmed using *m* files and compiled *C* code.

The latter part, once created, does not need to be frequently changed by user.

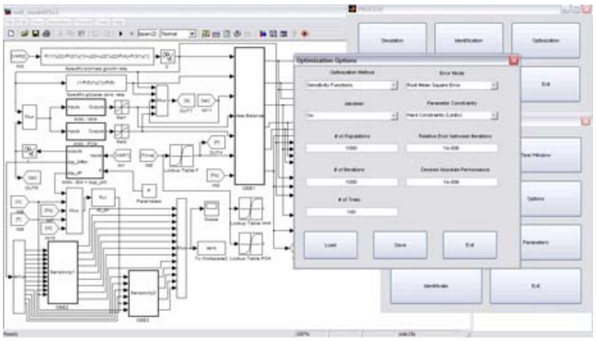

Fig. 3. Software implementation of the hybrid model and its identification procedure in *Matlab Simulink* environment

Fig. 3 depicts a typical screenshot of the software tool, consisting of the created fed-batch biosurfactant production process model and some additional windows that allow setting various model and algorithm specific options and settings, such like numeric integration method and its accuracy, identification method, number of iterations, desired absolute error and relative tolerance, etc.

The sensitivity equations approach analyzed in this paper was implemented using Matlab function LSQCURVEFIT for solving non-linear least squares problems using Jakobi matrix.

The model for fed-batch biosurfactant production process was implemented using software tool briefly described before. The mechanistic part of the model (eqs. 3.5, 3.5) was realized using the blocks from user-defined library of kinetic expressions, mass balance equations (3.1-3.4) and sensitivity equations system (3.10) – using *DEE* modules, ANNs for specific ammonia nitrogen (s_2) and phosphate phosphorus (s_3) consumptions rates – using standard ANNs of a given structure, ANN for specific biosurfactant production rate q_p (μ, p, s_2, s_3) – using embedded *Matlab* function (see Fig. 3).

5 Results and discussion

The identification using sensitivity equations approach was performed for the already described ANNs. It was made using the data from the “training" experiments and shows good modeling quality (see Fig. 4). The lowest modeling quality in both “training” and “validation” experiment sets was reached for ammonia nitrogen and phosphate phosphorus concentrations. This can be explained by the fact that not all the influencing factors were taken into account while modeling these specific reaction rates. In future investigations, it will be of advantage to improve the model with respect to these factors. It is also necessary to stress that the validation results show good modeling quality of the main state variable – biosurfactant concentration. It was comparable with the accuracy of the reference analytical measurement techniques and does not show any systematic deviations.

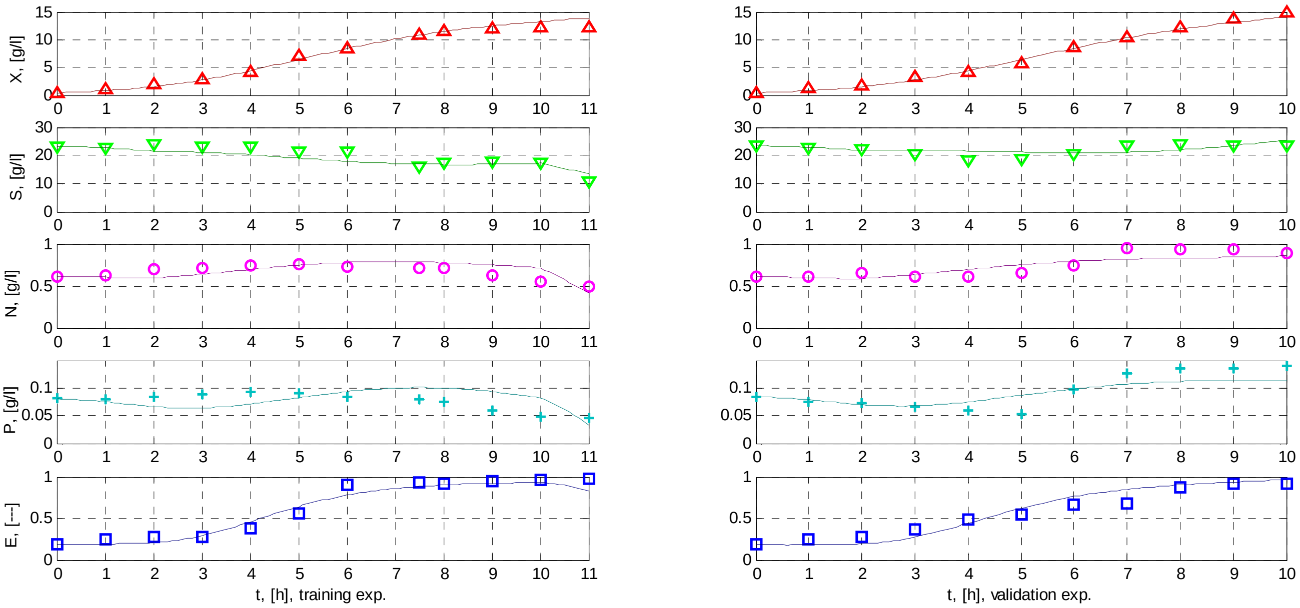

Fig. 4. Hybrid model identification results: training (left column) and validation experiments (right column)

Additionally, it is important to compare and to state that the applied identification technique was significantly faster (in an order and more, depending on the choice of the initial ANNs weights values) as compared to the evolutionary programming approach used in [3, 9] and led to the comparable or better modeling accuracy. As possible disadvantage of the sensitivity approach a necessity of additional differential sensitivity equations can be noted. I. e. having 6 (in the analyzed example) main differential equations for state variables one needs to add and simultaneously to integrate significantly higher number of sensitivity equations (in the analyzed example 13, 9 and 9 equations for each ANN respectively). Another drawback of the LSQCURVEFIT routine is the fact, that the calculation of residuals over the experiments and variables is performed within the routine and the relative over-/underweighting of particular experiment or variable is rather complicated, but possible.

6 Conclussions

The calculation results have proven that the presented approach for hybrid model identification is robust and efficient. In combination with the developed flexible and user-friendly software tool it allows to quickly and efficiently perform a parameter identification of complex hybrid models for biochemical processes.

Acknowledgements

The financial support for VG and RS in form of *NATO Reintegration Grant EAP.RIG.981479* within the *NATO Programme for Security through Science* is gratefully acknowledged.

References:

[1] Biriukov, V. V., Kantere, V. M., *Optimization of batch processes of microbiological synthesis*. Moscow, Nauka, 1985

[2] Galvanauskas V., Georgieva P., Feyo de Azevedo S. Dynamic Optimisation of Industrial Sugal Crystallization Process based on a Hybrid (Mechanistic+ANN) Model, *Proceedings of IEEE World Congress on Computational Intelligence, IJCNN 2006, July 16-21, 2006, Vancouver, Canada*, Vol. 6, p.p. 5035-5042.

[3] Galvanauskas V., Simutis R., Levišauskas D., Application of evolutionary computing for hybrid model based optimization of biochemical processes, *WSEAS Transactions on Information Science and Applications*, Issue 5, Vol. 2, May 2005, p.p. 507-511.

[4] Galvanauskas V., Simutis R., Lübbert A., Hybrid process models for process optimisation, monitoring and control, *Bioprocess and Biosystems Engineering*, 26(6), 2004, p.p. 393-400.

[5] Galvanauskas V., Volk N., Simutis R., Lübbert A., Design of Recombinant Protein production Processes, *Chemical Engineering Communications*, 191(5), 2004, p.p. 732-748.

[6] Horan N. J., Jarman T. R., Dawes E. A., Effects of carbon source and inorganic phosphate concentration on the production of alginic acid by a mutant of *Azotobacter vinelandii* and on the enzymes involved in its biosynthesis, *J. General Microbiol.*, Vol. 27, 1981, p.p. 185-191.

[7] Kosaric N., *Biosurfactants: production, properties, applications*, NewYork, Marcel Dekker, 1993

[8] Levišauskas D., Galvanauskas V., Henrich S., Wilhelm K., Volk N., Lübbert A., Model-Based Optimization of Viral Capsid Protein Production in Fed-Batch Culture of Recombinant *Escherichia coli*, *Bioprocess and Biosystems Engineering*, 25(4), 2004, p.p. 255-262.

[9] Levišauskas D., Galvanauskas V., Žunda G., Grigiškis S., Model-based optimisation of biosurfactant production in fed-batch culture *Azotobacter vinelandii*. *Biotechnology Letters*, 26(14), 2004, p.p. 1141-1146.

[10] Ljung L., *System identification: Theory for the users*, Englewood Cliffs, Prentice Hall, 1991

[11] Schubert J., Simutis R., Dors M., Havlik I., Lübbert A., Bioprocess optimization and control: Application of hybrid modelling, *J. Bio-techn.*, 35, 1994, p.p. 51-68.