J Ind Microbiol Biotechnol (2008) 35:1425–1433
DOI 10.1007/s10295-008-0443-5

ORIGINAL PAPER

On-line biomass estimation in biosurfactant production process by *Candida lipolytica* UCP 988

Clarissa Daisy da Costa Albuquerque · Galba Maria de Campos-Takaki · Ana Maria Frattini Fileti

Received: 24 March 2008 / Accepted: 30 July 2008 / Published online: 26 September 2008

Abstract Biomass is an important variable in biosurfactant production process. However, such bioprocess variable, usually, is collected by sampling and determined by off-line analysis, with significant time delay. Therefore, simple and reliable on-line biomass estimation procedures are highly desirable. An artificial neural network model (ANN) is presented for the on-line estimation of biomass concentration, in biosurfactant production by *Candida lipolytica* UCP 988, as a nonlinear function of pH and dissolved oxygen. Several configurations were evaluated while developing the optimal ANN model. The optimal ANN model consists of one hidden layer with four neurons. The performance of the ANN was checked using experimental data. The results obtained indicate a very good predictive capacity for the ANN-based software sensor with values of R^2 of 0.969 and RMSE of 0.021 for biomass concentration. Estimated biomass using the ANN was proved to be a simple, robust and accurate method.

Keywords Artificial neural network · Software sensor · Biomass · Biosurfactant production · *Candida lipolytica*

C. D. da Costa Albuquerque
Departamento de Estatística e Informática, Universidade Católica de Pernambuco, Rua Nunes Machado, 42, Bloco J, Térreo, Boa Vista, 50050-590 Recife, Pernambuco, Brazil
e-mail: cdaisy@unicap.br

G. M. de Campos-Takaki
Departamento de Química, Universidade Católica de Pernambuco, Rua Nunes Machado, 42, Bloco J, Térreo, Boa Vista, 50050-590 Recife, Pernambuco, Brazil

C. D. da Costa Albuquerque · G. M. de Campos-Takaki (✉)
Núcleo de Pesquisa em Ciências Ambientais, Universidade Católica de Pernambuco, Rua Nunes Machado, 42, Bloco J, Térreo, Boa Vista, 50050-590 Recife, Pernambuco, Brazil
e-mail: takaki@unicap.br
e-mail: cdaisy@unicap.br

A. M. F. Fileti
Departamento de Engenharia de Sistemas de Processos Químicos, Faculdade de Engenharia Química, Universidade Estadual de Campinas, Cidade Universitária "Zeferino Vaz", Caixa Postal 6066, 13081-970 Campinas, Sao Paulo, Brazil
e-mail: frattini@feq.unicamp.br

Introduction

The worldwide surfactant market, in the last years of twentieth century, was around $94 billion per annum, and their demand was expected to increase at a rate of 35% [11]. Even that sustainable management of natural values and tightening environmental protection laws have effectively resulted in an increasing interest in biosurfactants as possible alternatives to chemical surfactants; at present, the synthetic surfactants continue to dominate the global market because of high biosurfactant production costs. The reduction of the overall biosurfactant production costs usually depends on the strain improvements, the use of low-cost raw materials such as agricultural and industrial wastes as substrates, the use of process scale-up and the use of advanced computer-based techniques for process control and optimization. The lack of reliable, robust and low cost on-line sensors for key process variables, particularly for biological variables like biomass concentration and emulsification activity, limits the monitoring, control and cost optimization of biosurfactant production processes [1–3].

Several methods have been developed for estimating biomass. They differ in the measured phenomema or

correlating variable. An inevitable result of this is that all the different methods cannot be interrelated or applicable to all processes and organisms. Different methods emphasize different biomass properties, e.g., cell number, cell viability, metabolic state or mass total. Thus, it is also important to acknowledge the limitations of different measurement principles as well as the correlation of the method used to the variable that needs to be known. Currently, the most common method for estimating biomass is undoubtedly the absorbance measurement [19]. However, there are a wide number of techniques available to estimate biomass concentration: bioluminescence and chemiluminescence methods, epifluorescence microscopy, nephelometry, turbidimetry, electronics counting and sizing techniques (Coulter counting, flow cytometry), impedimetry, acoustics techniques, fluorescence-based methods, light scattering, microcalorimetry and others [24]. These methods are expensive and difficult to apply for on-line monitoring of biomass concentration in a biosurfactant production process. Moreover, computational methods based on conventional and/or artificial intelligence techniques may be used to predict biomass from process variables, like pH, temperature or dissolved oxygen (DO) [19].

The present work deals with the development of a prototype neural network-based software sensor for real time estimation of biomass concentration in a biosurfactant production process by *Candida lipolytica* UCP 988. Conventional off-line analyses of biomass are limited by the sampling frequency and duration, which usually take several hours. 'Software sensors' are mathematical algorithms, which provide reliable real-time estimation of unmeasured variables by using their correlation with available process data, and they also present the advantages of to-be-cheaper and faster than off-line analytical methods that require large and expensive instruments. Among the techniques used for the development of soft-sensors, artificial neural networks (ANNs) have strong potential in the on-line estimation of bioprocess [4]. Although the ability of neural networks to model non-linear and time-varying dynamics of bioprocesses—like wine making [8], penicillin fermentation [13], ethanol production [18], glucomylase production [19], lysine production [26], lipase production [22], baker's yeast fermentation [9], β-glucan extraction [12]—has been known for many years, their application as software sensor in biosurfactant production processes is recent [1, 3]. The purpose of this study is to develop an ANN-based software sensor for on-line estimation of biomass concentration, with enough complexity to capture the biosurfactant process characteristics and with sufficient simplicity to allow the model to be easily understood and implemented.

Materials and methods

Microorganism

The organism used was a strain of *C. lipolytica* UCP 988. This organism was maintained at 4 °C on yeast mold agar slants containing the following: yeast extract (0.3% w/v), malt extract (0.3% w/v), D-glucose (1% w/v), tryptone (0.5% w/v) and agar (1.5% w/v). The pH was adjusted to 5.0 with HCl.

Medium for inoculum development

Seed medium—SWDW-PASUG-2—was composed of sea water (50% v/v), distilled water (50% v/v), potassium phosphate (2.628% w/v), ammonium sulfate (2.130% w/v), urea (0.544% w/v), D-glucose (5% w/v). The initial pH of the production media was adjusted to 5.3 with 40% NaOH. The inoculum for bioemulsifier production was prepared in four Erlenmeyer flasks with capacity of 500 ml containing 100 ml of SWDW-PASUG-2 medium. Then, this suspension was incubated at 28 °C for 48 h at 150 rpm, and the culture having approximately 10^8 cells/ml was used to inoculate the bioreactor at 10% v/v.

Medium for biosurfactant production

Cultivations were conducted in a 5-l bioreactor (BioFlo2000, New Brunswick) equipped with standard probes for pH, temperature, DO and auxiliary equipment, containing 4 l of SWDW-PASUCO-2 medium: sea water (50% v/v), distilled water (50% v/v), potassium phosphate (2.628% w/v), ammonium sulfate (2.130% w/v), urea (0.544% w/v), corn oil (5% v/v). The initial pH of the production media was adjusted to 5.3 with 40% NaOH.

Full factorial design

A 2^2 full factorial design composed of a set of four experiments, with three replicates at the central point, was carried out to verify the effects and interactions of the temperature and agitation rate on the biomass concentration and emulsification activity in biosurfactant production process by *C. lipolytica* [1]. The range and levels of the factors (or independent variables) under study are given in Table 1. Statistical analysis of the factorial design was performed using Statistica® software version 6.0 (Statsoft, Inc., USA).

Biomass concentration determination

Biomass concentration was determined gravimetrically, by dry-weight measurement at 80 °C for 24 h after filtration

Table 1 Values of temperature and agitation at different levels of the 2^2 full factorial design

Independent variables	Levels		
	−1	0	1
Temperature (°C)	28	29.5	31
Agitation (rpm)	150	225	300

of the samples through 0.22 μm predried Millipore membranes.

Assay of emulsification activities

Emulsification activity was evaluated according to the method described by Cirigliano and Carman [7]. Cell-free filtrates were prepared for each culture, and the emulsification activity for water-in-hexadecane emulsions was determined [1].

Surface tension measurement

The surface tensions of the cell-free broths were measured by Du Nouy ring method [5] using a digital tensiometer model Sigma 70 (KSV Instruments Ltd, Finland) at room temperature.

Neural network software sensors development environment

Prototypes of the ANN-based software sensors were constructed using Neural Network Toolbox 4.01 [10], designed to run in the MATLAB technical computing environment (Matlab 6.1 Mathworks).

Neural network software sensor development methodology

Data acquisition

The data sets required to train, validate and test the neural software sensors were obtained from biosurfactant production experiments carried out using corn oil and sea water-based mineral medium in a 5-l bioreactor, under different temperature and agitation conditions specified in 2^2 full factorial design and at a aeration rate of 1 vvm [1]. A 2^2 full-factorial design with three replicates at the center point was carried out and five experimental data set triplicates were obtained at 28 °C and 150 rpm; 28 °C and 300 rpm; 31 °C and 150 rpm; 31 °C and 300 rpm and 29.5 °C and 225 rpm [1]. Data were recorded from the available on-line sensors (pH and DO) and from off-line sample analysis (biomass and emulsification activity) performed every 0, 4, 18 and 24 h on the first day and every 24 h for the next 6 days. The experimental data set triplicate obtained at 28 °C and 150 rpm (more economic condition: temperature and agitation were used at their lowest values), with varying pH and DO, was used to train, validate and test the neural software sensors, because in this condition, after 120 h of cultivation was obtained 5.46 unity of emulsification activity—the best emulsification activity to water-in-*n*-hexadecane emulsions of the 2^2 full-factorial design carried out—and biomass concentration of 16.103 g/l. The cell-free filtrate containing the surfactant produced by *C. lipolytica*, under the same conditions, decreased the surface tension of water from 72 to 33 mN/m [1].

Selection of input variables

Sensitivity analysis and historical knowledge about the process were used to select the most important input variables to estimate biomass concentration. To eliminate input variables generally known to have no direct or very little influence on biomass concentration, a large number of NN models were developed using different sets of input variables [1]. Although there are no systematic rules available for the secondary measurements selection, the following criteria were considered for helping to achieve the selection [14]:

(1) *Sensitivity*. The secondary variables pH and DO are relevant to biomass concentration (primary output) and also to rapidly reflect the unmeasured disturbances. In addition, the variables pH and DO embrace biomass concentration measurements that represent the entire range of variation of the phenomenon under study.
(2) *Availability.* The secondary variables pH and DO are measurable on-line without much difficulty and at a reasonable cost.
(3) *Robustness.* The biomass software sensor using the secondary variables pH and DO showed to be least sensitive to model errors.

Data preprocessing

The experimental data were divided into three sets: a training set used to adjust the weighting coefficients; a validation set used to find the optimal configuration of the software sensor and a test set was used to verify the true performance of the ANN-based software sensor chosen. The training, validation and test sets were smoothed and expanded by interpolation using a piecewise smoothing cubic spline [1, 6, 16, 20, 25]. The smoothing of relatively sparse off-line data was necessary not only to eliminate the noise related to the measurement errors but also to expand,

by interpolation, the data set used for neural model training. The measurements patterns used to train, validate and test the network were normalized between 0.1 and 0.9, because these values were found to improve convergence speed.

Training, validation and test procedures

Twenty neural network software sensor prototypes with one hidden layer and number of neurons varying from 1 to 20 were tested in the biosurfactant production process for estimation of biomass concentration. The biomass dry weight to be estimated in real-time was chosen as the output vector. The input process variables included pH and DO. The sigmoid and linear functions were used, respectively, as activation functions in hidden and output layers.

The goal of neural network training is to obtain a network, which produces small errors on the training set, but which will also respond properly to novel inputs. When a network is able to perform as well on validation set inputs as on training set inputs, the network generalizes well. The training process adjusts weights to minimize the error between the measured output and the output produced by the network. Through this adjustment, the neural network learns the input–output behaviors of the system. This procedure does not necessarily give a network with good generalization ability when the number of connection weights is relatively large. In such situation, overfitting to the training data occurs. To overcome this problem, there are several approaches such as regularization learning.

In this work, the training algorithm used was the Levenberg–Marquardt-based backpropagation algorithm [17], in conjunction with Bayesian regularization [15, 23]. One feature of this algorithm is that it provides a measure of how many network parameters (weights and biases) are being effectively used by the network. The typical performance function that is used for training feedforward neural networks is the sum of squares of the network errors (sse). It is possible to improve generalization modifying the sse performance function by adding a term, ssw, that consists of the sum of squares of the network weights and biases. The performance function resultant ssreg is defined as:

$$\text{ssreg} = \alpha\text{sse} + \beta\text{ssw} \tag{1}$$

where

$$\text{sse} = \sum_{i=1}^{N} (t_i - a_i)^2 \tag{2}$$

$$\text{ssw} = \frac{1}{n}\sum_{j=1}^{n} w_j^2 \tag{3}$$

$$\alpha = \frac{\gamma}{2\text{ssw}} \tag{4}$$

$$\beta = \frac{\text{np} - \gamma}{2\text{sse}} \tag{5}$$

and np and γ are, respectively, the total number of parameters in the network and the number of parameters effectively used by the network [10, 15]. The parameter γ is a measure of how many parameters of the network are effectively used in reducing the error function and it can vary from 0 to np. Each prototype was trained separately, ten times, with different initial weight matrices, using the same training data set with 1,000 iteration cycles and with learning rate coefficients of 0.001. After the completion of each training, the networks were validated by presenting experimental data sets, which were not used during training.

Comparison of neural network performances

Selecting appropriate criteria to differentiate between different types of models is a prerequisite for a good modeling approach. A compromise must be made between the desire to have a simple model with fewer parameters and more accurate predictions at the cost of a large number of parameters [18].

Since many of the published works on ANN application present different performance indices, it is usual authors investigated several well-known performance measures to allow comparisons with other studies (there being no universally accepted measure of neural network performance). It is important to apply multiple error measures taking into account that some measures penalize more the errors of greater magnitude (rmse), others penalize more the errors of lower magnitude (msre), others can provide useful indications of a model's overall performance (for example, coefficient of determination R^2), while others penalize models that have excessive numbers of parameters [e.g., A information criteria (AIC) and B information criteria (BIC)].

In this study, analysis of the statistical indices curves—mean squared error (mse), root mean squared error (rmse), normalized root mean squared error (nrmse), defined according to Eqs. 6, 7 and 8, respectively—were used to compare model performances and to choose the more accurate model.

$$\text{mse} = \frac{1}{N}\sum_{i=1}^{N} (t_i - a_i)^2 \tag{6}$$

$$\text{rmse} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (t_i - a_i)^2} \tag{7}$$

$$\text{nrmse} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_i - a_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_i - \bar{a})^2}} \quad (8)$$

where t_i represents the ith neural network target (real biomass); a_i represents the ith neural network response (estimated biomass); $\bar{a}$ represents the mean neural network response, which is the average of a over the N patterns (mean biomass); and N represents the total number of patterns in the corresponding set (training or validation).

The selection criteria are based on these statistical indices for the validation set instead of that for the training set to ensure model generalization. The 'goodness of fit' in both training and validation was given by the coefficient of determination, R^2, which describes the variance in the modeled variable that can be explained with the model [21, 22]. Graphical analysis of the plot of network predictions versus the experimental data was also used for corroborating the selection of the best network.

Results and discussion

Different variables affect the biomass concentration in the different phases of the biosurfactant production process by *C. lipolytica*. The complex relationship between the system conditions (temperature, pH, DO, aeration and agitation rates, media components, substrate concentrations, etc.) and the biomass concentration is not easy to describe through a simple mathematical equation. Neural network-based modeling using experimental data of the biosurfactant production process by *C. lipolytica* allowed demonstrating that, in the studied conditions, the biomass is a function of pH and DO. However, the small size of neural network developed, with only four hidden neurons, does not reduce its sufficiency, representativeness and importance. A model with a large set of input variables can cause various problems, such as overfiting, expensive and time-consuming data acquisition, without necessarily leading to an improvement in the performance of the neural network. Small-size neural networks are important for real-time process control applications. Large networks have an enormous number of connections, and therefore, the amount of data and the calculation time could be extremely high. In general, networks with fewer hidden neurons are preferable, as they usually have better generalization capabilities, fewer over-fitting problems and are more computationally efficient.

In this work, the capability of different prototypes of neural network-based software sensors with one hidden layer for estimation of biomass in biosurfactant production process by *C. lipolytica* was investigated. The ANN models that were generated were compared and the best was selected based on its global determination coefficient (R^2_g) and statistical indices (sse, mse, rmse and nrmse), and by graphical analysis of the plot of network predictions versus the experimental data sets. It was also taken into account that networks with fewer hidden neurons use smaller training data sets and therefore require less time and computational effort.

Validation data sets were assembled to assess the predictive accuracy of the trained NN. The pH and DO profiles, obtained at 25 °C, 150 rpm and 1 vvm, used in the validation set to assess the ability of generalization of the several neural network-based software sensor prototypes after the training, are shown in Fig. 1. The effect of the number of hidden neurons on the neural network training parameters epoch, np, γ and ssw is presented in first four columns of Table 2 and illustrated in Fig. 2. The effect of the number of hidden neurons on the statistical indices sse, mse, rmse, nrmse, and R^2_g—in simulations carried out using the validation set for estimation of biomass in biosurfactant production process by *C. lipolytica*—is presented in the last five columns of Table 2 and illustrated in Fig. 3. Optimal

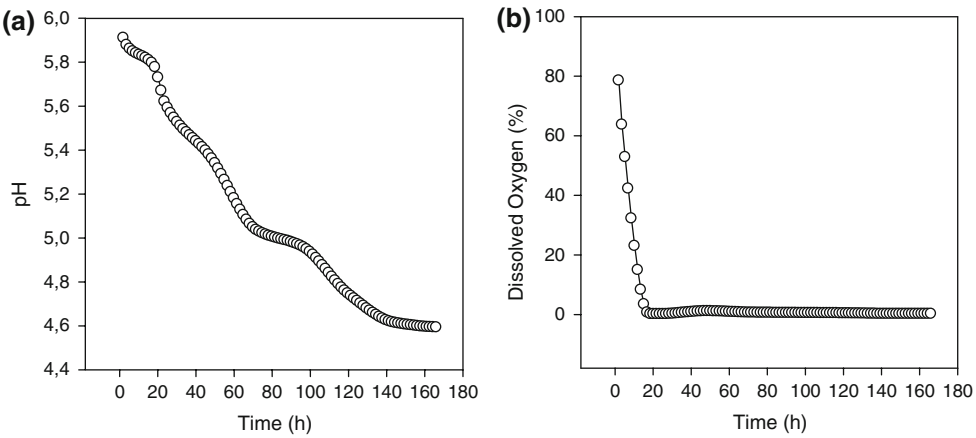

Fig. 1 Time course of (**a**) pH and (**b**) DO in biosurfactant production process by *Candida lipolytica* in a stirred tank bioreactor, with temperature, agitation and aeration controlled at 28 °C, 150 rpm and 1 vvm, respectively. Profiles used in the validation data set

Table 2 Neural network training parameters (epoch, ssw, np and γ) and statistical indices (sse, mse, rmse, nrmse and R_g^2) obtained using validation data set, as a function of the neuron number (S1) in the hidden layer

S1	epoch	np	γ	ssw	sse	mse	rmse	nrmse	R_g^2
1	240	5	3.85	9.30E+01	2.58E−01	1.29E−03	3.59E−02	3.29E−04	8.07E−01
2	93	9	8.27	5.51E+02	1.04E−01	5.22E−04	2.28E−02	2.09E−04	9.62E−01
3	513	13	10.20	7.51E+02	1.04E−01	5.20E−04	2.28E−02	2.09E−04	9.64E−01
4	234	17	14.60	1.61E+03	8.90E−02	4.45E−04	2.11E−02	1.93E−04	9.69E−01
5	104	21	15.80	7.30E+02	8.72E−02	4.36E−04	2.09E−02	1.91E−04	9.67E−01
6	463	25	16.60	9.35E+02	9.91E−02	4.95E−04	2.23E−02	2.04E−04	9.67E−01
7	144	29	17.30	1.00E+03	9.29E−02	4.65E−04	2.16E−02	1.97E−04	9.66E−01
8	79	33	17.30	1.38E+03	9.95E−02	4.98E−04	2.23E−02	2.04E−04	9.68E−01
9	68	37	19.30	1.47E+03	1.08E−01	5.39E−04	2.32E−02	2.13E−04	9.67E−01
10	124	41	17.00	6.62E+02	8.71E−02	4.36E−04	2.09E−02	1.91E−04	9.69E−01
11	66	45	20.30	3.58E+03	9.58E−02	4.79E−04	2.19E−02	2.00E−04	9.69E−01
12	63	49	20.70	4.19E+03	9.89E−02	4.94E−04	2.22E−02	2.04E−04	9.71E−01
13	20	53	21.90	2.61E+03	1.06E−01	5.28E−04	2.30E−02	2.11E−04	9.70E−01
14	59	57	23.60	5.48E+03	9.19E−02	4.59E−04	2.14E−02	1.96E−04	9.72E−01
15	28	61	23.10	6.63E+03	1.01E−01	5.07E−04	2.25E−02	2.06E−04	9.69E−01
16	7	65	23.20	8.98E+03	1.03E−01	5.16E−04	2.27E−02	2.08E−04	9.71E−01
17	6	69	26.40	1.02E+04	9.87E−02	4.94E−04	2.22E−02	2.03E−04	9.71E−01
18	5	73	25.00	8.15E+03	1.03E−01	5.13E−04	2.27E−02	2.08E−04	9.70E−01
19	4	77	24.20	1.26E+04	9.62E−02	4.81E−04	2.19E−02	2.01E−04	9.70E−01
20	8	81	27.70	1.08E+04	9.47E−02	4.74E−04	2.18E−02	1.99E−04	9.72E−01

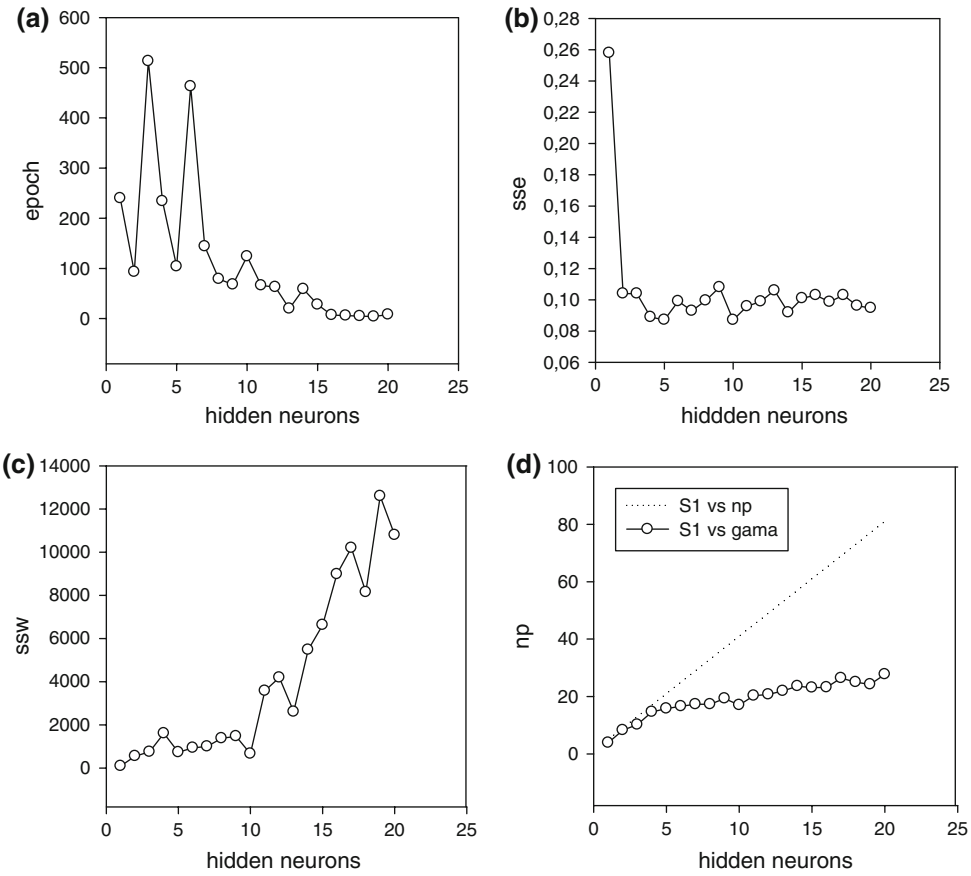

Fig. 2 **a** Epoch and neural network parameters—**b** sse, **c** ssx and **d** γ and np—as a function of the neuron number in the hidden layer

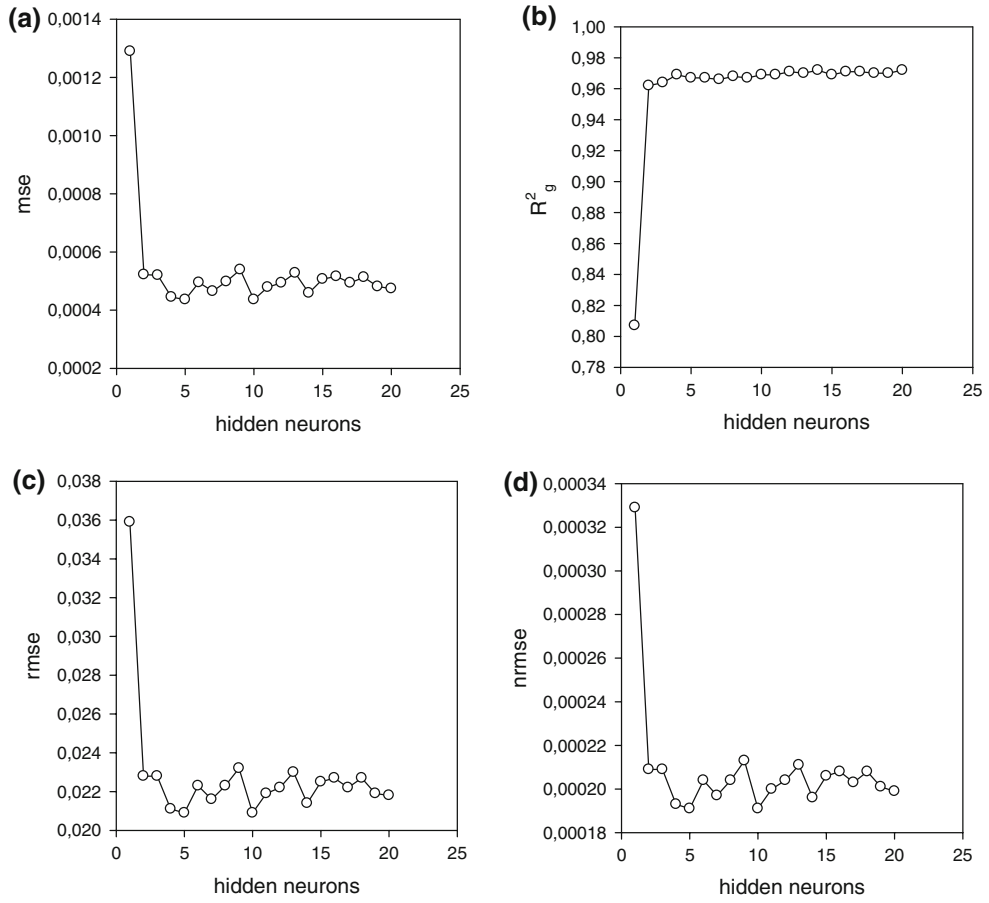

Fig. 3 Effect of the number of hidden neurons on (**a**) mse, (**b**) R^2_g, (**c**) rmse and (**d**) nrmse for estimation of biomass in biosurfactant production process by *Candida lipolytica*

numbers of nodes of input and hidden layers were determined by model selection criteria on validation data set.

Table 2 shows how the prediction accuracy is affected by the number of neurons in the hidden layer. Ideally, the values of statistical indices (sse, mse, rmse, nrmse), obtained with the validation data set, should be close to 0, indicating that the model well learned the relationship among the input and output parameters. The generalization ability and the robustness of the model are measured by the global coefficient of determination (R^2_g). If the value of global coefficient of determination R^2_g is unity, the generalization ability and the robustness of the model are maximal. The prototype that gives the best results, using the validation data set, is the one with pH and DO concentration as input pattern and with four neurons in the hidden layer. It may be observed from Table 2 that the fourth configuration (i.e., neural network with 2-4-1 topology) results in the best combination of low estimation errors, high global coefficient of determination and low number of hidden neurons. For the validation set, the model with topology 2-4-1 presented values of the statistical indices sse, mse, rmse, nrmse and R^2_g of 8.90E−02, 4.45E−04, 2.11E−02, 1.93E−04 and 9.69E−01, respectively.

Figure 3 shows that fewer than four hidden neurons caused mse, rmse and nrmse to rise sharply. More than four hidden neurons caused mse, rmse and nrmse to rise slowly. The global coefficient of determination (R^2_g) presented opposite behavior. The addition of more neurons in the hidden layer initially increased and later decreased and increased slightly the fits, but did not lead to an overfitting, reducing the neural network ability to generalize. Simulation shows that Levenberg–Marquardt-based backpropagation algorithm in conjunction with Bayesian regularization produces networks with better generalization performance and lower susceptibility to overfitting as the network size increases.

The performance of this neural network model for prediction of biomass concentration, using the training and validation data set, is illustrated in Fig. 4. The comparison of the neural network model prediction with the experimental values of biomass concentration is presented in the Fig. 4a, c using time course graph and in Fig. 4b, d using parity plots. Figure 4 illustrates the performance of the

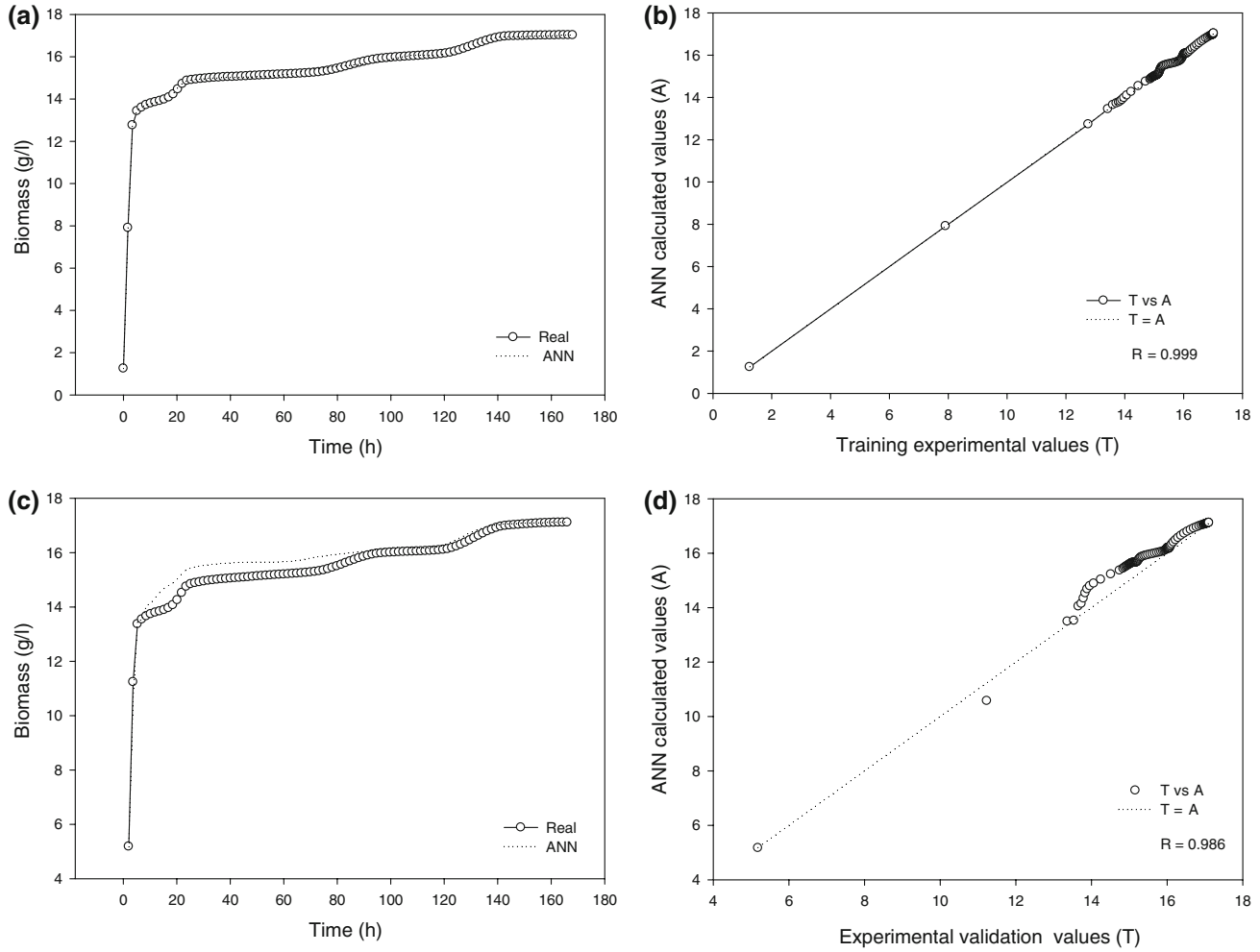

Fig. 4 Comparison between experimental biomass concentration values (*thick line*) and neural network-calculated biomass concentration values (*dotted lines*), using (**a**) training data set and (**c**) validation data set. Parity plot between ANN-predicted and measured values of biomass concentration, using (**b**) training data set and (**d**) validation data set

selected ANN software sensor with topology 2-4-1 for the prediction of biomass concentration. Figure 4a and 4c shows the predictions obtained with the ANN model for 100 data chosen randomly from the training and validation data sets, respectively. The ANN model, for both training and validation, accurately estimated the variation of real biomass concentrations values. Figure 4b and 4c show the parity plots, i.e., comparisons between the network outputs (estimations) and the corresponding targets (experimental data), for the training and validation datasets, respectively. The goodness of fit is given by the coefficient of determination R^2. If the value of R^2 is unity, the model predicts exactly every experimental point. In this case, coefficients of determination (R^2) of 0.998 and 0.972 were obtained for training and validation data sets, respectively. Coefficients of determination higher than 0.90 indicated excellent agreement of the neural network model with the experimental training and validation values, obtained for biomass concentration. As shown in Fig. 4d, the regression line is very close to the 45° line, and the dispersion of data is relatively small when measurement errors for biomass in industrial data are taken into account. This plot clearly shows that the model can satisfactorily predict biomass from data contained in the biosurfactant production process database. The slight deviation from linearity can be attributed to the noise in the experimental biomass concentration values.

Finally, empirical relations among pH, DO and biomass in biosurfactant process were investigated and established. The results obtained are very important, because they clearly reveal the sufficiency and representativeness of pH and DO as relevant input variables for on-line estimation of biomass concentration in biosurfactant production process by *C. lipolytica*. Analysis of the results demonstrated that the neural modeling approach is a useful tool for accurate and cost-effective modeling of biosurfactant production processes. Therefore, the softsensor with topology 2-4-1 developed in the present work can be used for the

supervision and understanding of the biosurfactant production process, obtaining on-line accurate measurements and replacing expensive and difficult off-line procedures.

Conclusion

Although ANN softsensors for estimation of biomass concentration are currently among the most studied bioprocess software sensor, owing to their great potential in various applications, nothing has been specifically published in the literature on their use in biosurfactant production processes by *C. lipolytica* yeast.

This paper has demonstrated how an ANN of relatively modest scale can be used to capture complex biosurfactant production process by *C. lipolytica* dynamics. The quality of results suggest that a good accurate relationship has been identified. It was found that the biomass concentration in biosurfactant production process by *C. lipolytica* can be inferred from on-line measurements of pH and DO. Simple and robust neural network-based software sensor with only four neurons in the hidden layer was very good descriptive model for *C. lipolytica* growth with pH and DO varying in batch cultures. The results showed that the selected ANN model can assuredly replace expensive instrumentation used for the estimation of biomass concentration in biosurfactant production process by *C. lipolytica*.

Although this study was restricted to a software sensor development for a biosurfactant production process by *C. lipolytica*, the methodology of model building using neural network may be applied to other chemical and biotechnological processes.

Acknowledgments This work received financial support from FINEP/CT-PETRO, CNPq and CNPq/CT-PETRO, Catholic University of Pernambuco (UNICAP) and State University of Campinas (UNICAMP). The authors are thankful to Salatiel Joaquim de Santana for his technical support and suggestions.

References

1. Albuquerque CDC (2006) Biosurfactant production process by *Candida lipolytica*: optimization, scale-up and development of artificial neural network based softsensor. PhD thesis (in portuguese), p 375
2. Albuquerque CDC, Fileti AMF, Campos-Takaki GM (2006) Optmizing the medium components in bioemulsifiers production by *Candida lipolytica* with response surface method. Can J Microbiol 52:6575–6583
3. Albuquerque CDC, Fileti AMF, Campos-Takaki GM (2006) Neural network based software sensors: application to biosurfactant production by *Candida lipolytica*. In: Mendez-Vilas A (ed) Modern multidisciplinary applied microbiology: exploiting microbes and their interactions. Wiley-VCH, Weinheim, pp 628–632
4. Assis AJ, Maciel Filho R (2000) Soft sensors for on-line bioreactor state estimation. Comp Chem Eng 24:1099–1103
5. ASTM D971 (1999) 99th Standard test method for interfacial tension of oil against water by the ring. Method American Society for Testing Materials
6. Chen L, Bernard O, Bastin G, Angelov P (2000) Hybrid modelling of biotechnological processes using neural networks. Control Eng Pract 8(7):821–827
7. Cirigliano CM, Carman GM (1984) Isolation of a bioemulsifier from *Candida lipolytica*. Appl Environ Microbiol 48:747–750
8. Cleran Y, Thilbault J, Cheruy A, Corrieu G (1991) Comparison of prediction performance between models obtained by the group method of data handling and neural networks for alcoholic fermentation rate in enology. J Ferment Bioeng 71:356–362
9. DaCosta P, Kordich C, Williams D, Gomm JB (1997) Estimation of inaccessible fermentation states with variable inoculum sizes. Artif Intell Eng 11:383–392
10. Demuth H, Beale M (2001) Neural network toolbox: for use with Matlab. User's guide. Version 4. Release 12. The Mathworks Inc., Natick
11. Desai JD, Banat IM (1997) Microbial production of surfactant and their commercial potential. Microbiol Mol Biol Rev 61:47–64
12. Desai KM, Vaidya BK, Singhal RS, Bhagwatt SS (2005) Use of an artificial neural network in modeling yeast biomass and yield of β-glucan. Proc Biochem 40:1617–1626
13. Di Massimo C, Montague GA, Willis MJ, Tham MT, Morris AJ (1992) Towards improved penicillin fermentation via artificial neural networks. Comp Chem Eng 16:283–291
14. Du Y-G, Del Villar RG, Thibault J (1997) Neural net-based softsensor for dynamic particle size estimation in grinding circuits. Int J Miner Process 52:121–135
15. Foresee FD, Hagan MT (1997) Gauss–Newton approximation to Bayesian learning. IEEE Trans Neural Netw 3:1930–1935
16. Glassey J, Montague GA, Ward AC, Kara BV (1994) Artificial neural network based experimental design procedure for enhancing fermentation development. Biotech Bioeng 44(4):397–405
17. Hagan MT, Menhaj M (1992) Training feedfoward networks with the Maquardt algorithm. IEEE Trans Neural Netw 5:989–993
18. Karim MN, Rivera L (1992) Artificial neural network in bioprocess state estimation. Adv Biochem Eng Biotechnol 46:1–33
19. Kiviharju K, Salonen K, Moilanen U, Meskanen E, Leisola M, Eerikäinen T (2007) On line biomass measurements in bioreactor cultivations: comparison of two on-line probes. J Ind Microbiol Biotechnol 34:561–566
20. Laursen SO, Webb D, Ramirez WF (2007) Dynamic hybrid neural network model of an industrial fed-batch fermentation process to produce foreign protein. Comp Chem Eng 31:163–170
21. Linko P, Zhu Y-H (1992) Neural network modelling for real-time variable estimation and prediction in the control of glucoamylase fermentation. Process Biochem 27:275–283
22. Linko S, Luopa J, Zhu Y-H (1997) Neural network as 'softsensor' in enzyme production. J Biotechnol 52:257–266
23. MacKay DJC (1994) Bayesian interpolation. Neural Comput 4:415–447
24. Pons M-N (1992) Physical and chemical sensors—actuators. In: Pons MN (ed) Bioprocess monitoring and control. Hanser series in biotecnologie. Oxford University Press, New York, pp 86–106
25. Schepers AW, Thibault J, Lacroix C (2000) Comparison of simple neural networks and nonlinear regression models for descriptive modeling of *Lactobacillus helveticus* growth in pH-controlled batch cultures. Enzyme Microb Technol 26:431–445
26. Zhu Y-H, Rajalahti T, Linko S (1996) Application of neural networks to lysine production. Chem Eng J 62:207–214