J Ind Microbiol Biotechnol (2009) 36:1139–1148
DOI 10.1007/s10295-009-0595-y

ORIGINAL PAPER

Modeling of polygalacturonase enzyme activity and biomass production by *Aspergillus sojae* ATCC 20235

Figen Tokatli · Canan Tari · S. Mehmet Unluturk · Nihan Gogus Baysal

Received: 13 March 2009 / Accepted: 11 May 2009 / Published online: 29 May 2009

Abstract *Aspergillus sojae*, which is used in the making of koji, a characteristic Japanese food, is a potential candidate for the production of polygalacturonase (PG) enzyme, which of a major industrial significance. In this study, fermentation data of an *A. sojae* system were modeled by multiple linear regression (MLR) and artificial neural network (ANN) approaches to estimate PG activity and biomass. Nutrient concentrations, agitation speed, inoculum ratio and final pH of the fermentation medium were used as the inputs of the system. In addition to nutrient conditions, the final pH of the fermentation medium was also shown to be an effective parameter in the estimation of biomass concentration. The ANN parameters, such as number of hidden neurons, epochs and learning rate, were determined using a statistical approach. In the determination of network architecture, a cross-validation technique was used to test the ANN models. Goodness-of-fit of the regression and ANN models was measured by the R^2 of cross-validated data and squared error of prediction. The PG activity and biomass were modeled with a 5-2-1 and 5-9-1 network topology, respectively. The models predicted enzyme activity with an R^2 of 0.84 and biomass with an R^2 value of 0.83, whereas the regression models predicted enzyme activity with an R^2 of 0.84 and biomass with an R^2 of 0.69.

F. Tokatli (✉) · C. Tari · N. Gogus Baysal
Department of Food Engineering, Izmir Institute of Technology, 35430 Urla-Izmir, Turkey
e-mail: figentokatli@iyte.edu.tr

S. M. Unluturk
Department of Software Engineering, Izmir University of Economics, Balçova-Izmir, Turkey

Keywords Artificial intelligence · Cross-validation · Filamentous fungi · Polygalacturonase production · Submerged culture

Introduction

The modeling of experimental or real-time data with artificial neural network (ANN) techniques has been applied in many research fields, such as biotechnology, pharmaceutical, medicine and food [1–8]. In the biotechnology field, ANN modeling is a commonly applied approach for carrying out estimations, control studies, classification analysis, and fault diagnosis. The major reason for the popularity of ANN techniques is their ability to model the complex and nonlinear behavior of the systems with a reliable set of input and output data pairs, without the need for prior information.

As in other data-based modeling strategies, ANN modeling requires two sets of data: (1) a larger set for the training (learning) of the network and (2) a smaller set for the testing of the model. In the training phase, the network learns from the known input–output data by adjusting its parameters (weights). The training data must include diverging sets of input–output pairs to be able to capture the dynamics of the system. The test data, on the other hand, has to be distinct and is used after the network architecture (topology), weights (w_{ij}) and biases (b_j) have been determined, since it will assess the performance of the network model with observations which are not used in the learning stage. One problem in ANN applications is the lack of sufficient input–output observations. This is common, especially in biological fields, where data collection in living systems may be challenging. Table 1 shows some examples of ANN modeling in cases of small

Table 1 A number of artificial neural network (ANN) modeling studies with a low number of data points

Source	Data-training	Data-testing	Input/output	Hidden layers	Neurons
Razmi-Rad et al. [9]	106	26	4/6	2	3 and 5
Bas and Boyaci [10]	13	22	2/1	1	4
Huang et al. [11]	13	8	3/1	1	4
Desai et al. [12]	44	10	4/1	1	4
Alonso-Salces et al. [13]	48	16	33/2	1	3
Hervas-Martinez et al. [14]	30	15	4/3	1	5 and 6
Yuste and Dorado [15]	45	45	5/1	1	13
Esnoz et al. [16]	70	12	3/1	1	3 and 5
Alonso-Salces et al. [17]	64	21	3/1 and 4/1	1	3 and 9
Hongwen et al. [18]	22	9	5/1	1	7
Spanila et al. [19]	16	3	3/1	1	3
Pazourek et al. [20]	35	–	7/2	–	–
Dutta et al. [21]	20	14	3/1	1	9
Perez-Magarino et al. [22]	107	37	7/3	1	4
Irudayaraj et al. [23]	32	16	46/3	1	40
Coleman et al. [24]	55	14	6/3	1	10
Castellanos et al. [25]	107	–	4/1	1	4
Iizuka and Aishima [26]	38	Cross-validation	20/3	1	6
Sun et al. [27]	170	Cross-validation	17/6	1	20

data sets in biotechnological studies. In cases where there is no luxury of splitting data into two parts, a cross-validation technique known as the leave-one-out (LOO) or leave-more-out procedure can be an alternative internal estimator of the model.

Pectinases are one of the most important groups of enzymes and are used extensively in the food, paper and textile industries and in wastewater treatments. These enzymes degrade the long and complex molecules of pectin, which are formed by galacturonic acid units linked by glycosidic bonds. Among the many pectinases identified to date, polygalacturonase (PG) is responsible for the hydrolytic cleavage of the polygalacturonic acid chain. Fungi are known to be a good source of PG enzyme. A number of *Aspergillus*, *Rhizopus* and *Penicillum* species are used for the large-scale production of PG enzymes [28, 29]. Another potential candidate for PG enzyme production is *Aspergillus sojae*, which is known as the organism used in the production of koji, a characteristic Japanese food product. The production of this pectinase enzyme and its activity studies with *A. sojae* ATCC 20235 are a new area of research, and an earlier study by our group focused on the development of low-cost nutrient media using statistical tools [30].

The scope of the study reported here is to model the activity of PG and its biomass formation at the end of a submerged fermentation by using a historical data set, which was created by two separate optimization studies. Artificial neural network modeling was used to combine both data sets to form an overall model. The LOO cross-validation technique was used to validate and measure the performance of the network models due to the low number of observations in the historical data sets. This study provides an example of ANN application in a biological process, which introduces the production of a commercially valuable enzyme by an organism that has not been considered in this context to date.

Materials and methods

Fermentation system

The data used in this study originate from a series of submerged fermentation studies with *A. sojae* ATCC 20235, which was purchased in lyophilized form (Procochem., Middlesex, UK). The details of inoculation, fermentation and measurement steps are given elsewhere [30]. After a preactivation step on YME agar and preparation of spore suspensions on molasses agar slants, the spores were incubated at 30°C for 1 week. Spore suspensions were collected and stored at 4°C. The fermentations were performed in flasks containing basal medium [in g/l: glucose, 25; peptone, 2.5; disodium phosphate, 3.2; monosodium phosphate, 3.3; corn steep liquor (CSL) and maltrin at changing concentrations]. The fermentation flasks containing 50 ml of production medium were agitated at different speeds (rpm) for 96 h at 30°C. The enzymatic

activity of PG was determined according to the procedure described by Panda et al. [31]. One unit of enzyme activity was defined as the enzyme that catalyzes the release of 1 μmol of galacturonic acid per unit volume per unit time (expressed as U/ml). Biomass determination was performed by the gravimetric method and expressed as grams per liter.

The effects of agitation speed, inoculum amount, maltrin and CSL concentration on PG activity and biomass were studied with the two-step response surface optimization method (RSM1 and RSM2). Table 2 presents the four-factor optimization experiments of the first face-centered central composite design (RSM1), which included 31 experiments (16 factorial points + 8 axial points + 7 center runs) and 12 validation experiments. In order to determine the best operating conditions, we ran a second face-centered central composite design with the same factors within different ranges. Table 3 presents this second set of four-factor optimization experiments (RSM2), which included 31 experiments (16 factorial points + 8 axial points + 7 center runs) and 17 validation experiments. Therefore, a total of 91 sets (points) of experimental data were collected (43 in RSM1 and 48 in RSM2) throughout these two optimization studies. The final pH of the fermentation medium, PG activity and biomass values were measured. The data set collected in the RSM2 study (Table 3) was analyzed to derive the final response surface Eqs. 1 and 2 for the PG activity and biomass responses in terms of significant main, interaction and quadratic effects ($P < 0.05$).

$$\begin{aligned}\text{Activity-RMS} = {} & 6.3 + 0.29\,x_1 - 0.15\,x_2 + 2.75\,x_3 \\ & - 2.66\,x_4 + 1.93\,x_4^2 + 0.58\,x_1x_2 \\ & - 0.52\,x_2x_4 + 1.11\,x_3x_4 \end{aligned} \quad (1)$$

$$\begin{aligned}\text{Biomass-RMS} = {} & 17.86 + 1.51\,x_1 + 1.94\,x_2 + 2.44\,x_3 \\ & + 5.74\,x_4 - 4.47\,x_4^2 + 1.85\,x_1x_2 \\ & - 1.23\,x_2x_3 \end{aligned} \quad (2)$$

where x_1, x_2, x_3 and x_4 are the coded values for the agitation speed, inoculum ratio, maltrin and CSL, respectively. The coding was done according to $x = [\text{Actual} - (\text{Low} + \text{High})/2]/(\text{High} - \text{Low})/2$. The details of the optimization experiments and data analysis can be found in Tari et al. [30].

Multiple linear regression

Multiple linear regression (MLR) is a common statistical technique to explain system outputs Y in terms of inputs or process variables X. The regression constants used to give the minimum model errors are determined by the least square technique. The fermentation outputs, PG activity and biomass were modeled by MLR using the 91 data points. R^2 of the model, R^2 of the test data (R^2 test) and lack-of-fit (LOF) P value were given as the model outputs. The R^2 test shows the prediction ability of the model and is determined by the cross-validation technique. The LOF P value should be insignificant ($P > 0.1$) in order to adequately define the variation in the system. Modde 7.0 was used to create the regression models (Umetrics, Umeå, Sweden).

Response surface methodology

The RSM is an experimental design technique in which the system is statistically investigated with several inputs and outputs for optimization purposes. In this study, RSM was used to optimize the ANN structure in terms of number of neurons, epochs and learning rate as the inputs and error measures and R^2 of the model as the outputs. The best combination of input variables that provide a minimum error and maximum prediction ability in terms of the R^2 of the test data was determined for PG activity and biomass in the *A. sojae* fermentation system. Details of the network optimization are provided in the following section.

Artificial neural network modeling and performance measures

The input variables used in the ANN models were agitation speed (x_1), inoculum ratio (x_2), maltrin concentration (x_3), CSL concentration (x_4) and final pH of the fermentation medium (x_5) (Fig. 1). The minimum and maximum values for the input variables are given in Table 4. Separate ANN models were generated for the output variables PG activity (y_1) and biomass (y_2). The ANN used in this work has three layers: an input layer, one hidden layer and an output layer. The hyperbolic tangent sigmoid function, $f(x) = \{2/[1 + \exp(-2x)]\} - 1$, was found to be the most suitable transfer function for both the hidden and output layers. The performance goal was set to 0.01. The data were scaled between $[-1\ +1]$ prior to the training in the network with the following expression:

$$Z_I = 2\left[\frac{X_I - X_{\min}}{X_{\max} - X_{\min}}\right] - 1 \quad (3)$$

Before training our ANN models, their weights and biases were initialized according to the Nguyen–Widrow initialization algorithm [32]. The optimization of a network can be accomplished by changing the network parameters (such as the number of neurons, number of hidden layers and number of epochs) one at a time. As an alternative approach, easy-to-use experimental design techniques were used to incorporate the combined effect of network parameters on error measures. A face-central composite design was run that included three factors: number of neurons in the single hidden layer, epochs and learning

Table 2 Mold fermentation data obtained by the first optimization study (RSM1)

Experiment	Agitation speed (rpm)	Inoculum ratio (total spore count)	Maltrin (g/l)	CSL (g/l)	pH	Activity (U/ml)	Biomass (g/l)	Low and high values of the factors[a]
1	225	5.00E + 05	50	8.75	3.70	2.213	17.481	Agitation: 150 and 300
2	150	7.50E + 05	25	2.5	3.51	0.765	10.668	Inoculum: 2.5×10^5 and 7.5×10^5
3	150	2.50E + 05	25	15	4.06	1.568	15.390	Maltrin: 25 and 75
4	300	2.50E + 05	25	15	3.84	1.890	14.060	CSL: 2.5 and 15
5	300	7.50E + 05	75	15	3.59	4.813	22.115	
6	300	5.00E + 05	50	8.75	3.37	2.983	12.185	
7	225	5.00E + 05	50	8.75	3.72	3.548	18.589	
8	150	7.50E + 05	25	15	4.15	0.317	19.165	
9	150	7.50E + 05	75	15	3.70	4.189	19.955	
10	150	7.50E + 05	75	2.5	3.33	5.492	14.983	
11	300	2.50E + 05	25	2.5	3.39	1.717	7.814	
12	225	5.00E + 05	50	8.75	3.52	2.495	13.887	
13	300	7.50E + 05	25	2.5	3.64	2.717	9.929	
14	300	7.50E + 05	75	2.5	3.33	7.351	13.047	
15	225	5.00E + 05	50	8.75	3.74	3.129	18.879	
16	150	2.50E + 05	75	15	3.85	4.351	17.261	
17	225	5.00E + 05	50	8.75	3.57	1.697	13.940	
18	150	5.00E + 05	50	8.75	3.55	1.668	13.949	
19	300	2.50E + 05	75	15	3.60	3.359	15.686	
20	225	7.50E + 05	50	8.75	3.49	2.433	12.811	
21	225	5.00E + 05	25	8.75	3.92	0.907	17.149	
22	300	2.50E + 05	75	2.5	3.49	8.620	10.227	
23	225	5.00E + 05	50	2.5	3.27	4.354	10.344	
24	150	2.50E + 05	25	2.5	3.44	1.576	10.188	
25	150	2.50E + 05	75	2.5	3.26	4.569	13.138	
26	300	7.50E + 05	25	15	3.88	0.290	12.186	
27	225	2.50E + 05	50	8.75	3.71	3.630	16.603	
28	225	5.00E + 05	50	8.75	3.59	1.899	16.629	
29	225	5.00E + 05	50	15	4.03	2.153	20.693	
31	225	5.00E + 05	75	8.75	3.68	5.582	20.077	
31	225	5.00E + 05	50	8.75	3.58	2.521	14.615	
32	300	1.25E + 04	75	2.5	3.60	11.585	21.697	
33	300	1.25E + 04	75	2.5	3.59	9.611	21.667	
34	300	2.50E + 05	75	2.5	3.52	10.352	21.057	
35	300	2.50E + 05	75	2.5	3.70	8.618	22.813	
36	300	7.50E + 05	75	2.5	3.52	7.341	21.547	
37	300	7.50E + 05	75	2.5	3.47	7.139	25.317	
38	300	1.00E + 07	75	2.5	3.78	4.786	29.561	
39	300	1.00E + 07	75	2.5	3.55	4.829	23.614	
40	300	1.00E + 07	100	2.5	3.71	8.312	33.045	
41	300	1.00E + 07	100	2.5	3.75	10.145	33.538	
42	300	1.25E + 04	100	2.5	3.59	12.777	23.720	
43	300	1.25E + 04	100	2.5	3.57	11.844	27.612	

CSL Corn steep liquor, *RSM* response surface optimization method

[a] The low (−1 in coded form) and high values (+1 in coded form) of the factor variables in the face-centered central composite design (the first 31 experiments)

Table 3 Mold fermentation data obtained by the second optimization study (RSM2)

Exp	Agitation speed (rpm)	Inoculum ratio (total spore count)	Maltrin (g/l)	CSL (g/l)	pH	Activity (U/ml)	Biomass (g/l)	Low and high values of the factors[a]
1	350	2.00E + 07	50	0	6.00	9.634	7.592	Agitation: 150 and 350
2	350	2.00E + 07	120	5	3.51	8.728	31.489	Inoculum: 1.25×10^4 and 2×10^7
3	250	1.00E + 07	85	2.5	3.47	6.950	22.543	Maltrin: 50 and 120
4	250	1.00E + 07	85	0	6.00	10.460	10.097	CSL: 0 and 5
5	250	1.00E + 07	85	2.5	3.48	6.133	18.276	
6	150	2.00E + 07	120	0	5.90	11.287	11.026	
7	350	1.25E + 04	120	0	5.92	10.551	6.819	
8	250	1.00E + 07	120	2.5	3.54	7.869	23.018	
9	150	1.25E + 04	50	5	3.66	1.895	13.285	
10	350	2.00E + 07	50	5	3.49	2.371	20.498	
11	250	2.00E + 07	85	2.5	3.43	5.332	18.500	
12	150	2.00E + 07	50	5	3.49	1.025	20.618	
13	150	1.25E + 04	120	0	5.87	13.503	7.562	
14	150	1.00E + 07	85	2.5	3.47	5.283	10.000	
15	250	1.00E + 07	85	2.5	3.50	6.360	13.128	
16	250	1.00E + 07	50	2.5	3.54	4.167	14.996	
17	150	1.25E + 04	120	5	3.53	10.474	26.631	
18	250	1.00E + 07	85	2.5	3.44	6.156	20.758	
19	350	1.25E + 04	120	5	3.64	11.023	18.462	
20	350	1.25E + 04	50	0	6.11	8.962	4.065	
21	350	2.00E + 07	120	0	5.43	15.429	12.043	
22	250	1.00E + 07	85	2.5	3.43	7.577	17.015	
23	150	2.00E + 07	50	0	6.17	9.321	7.841	
24	250	1.00E + 07	85	5	3.58	5.171	18.542	
25	250	1.25E + 04	85	2.5	3.44	7.369	13.503	
26	250	1.00E + 07	85	2.5	3.47	6.101	18.753	
27	150	2.00E + 07	120	5	3.42	8.132	10.087	
28	150	1.25E + 04	50	0	6.08	8.810	1.872	
29	350	1.00E + 07	85	2.5	3.42	6.865	22.558	
31	350	1.25E + 04	50	5	3.57	1.313	12.529	
31	250	1.00E + 07	85	2.5	3.47	5.697	19.136	
32	350	2.00E + 07	120	0	5.90	13.162	5.080	
33	350	4.00E + 08	120	0	5.79	8.687	6.778	
34	350	4.00E + 08	120	0	5.26	10.477	8.540	
35	350	2.00E + 07	150	0	6.03	11.760	3.758	
36	350	2.00E + 07	150	0	5.90	15.875	8.774	
37	350	2.00E + 07	180	0	5.90	13.190	9.175	
38	350	2.00E + 07	180	0	6.01	20.101	10.470	
39	150	1.25E + 04	120	5	3.70	9.511	9.018	
40	150	1.25E + 04	120	6	3.62	7.168	9.275	
41	350	2.00E + 07	120	4.10	3.46	5.384	37.698	
42	350	2.00E + 07	120	4.10	3.33	6.250	43.034	
43	350	2.00E + 07	150	4.10	3.43	9.259	42.369	
44	350	2.00E + 07	150	4.10	3.39	7.767	49.875	
45	350	2.00E + 07	180	4.10	3.35	11.384	56.167	
46	350	2.00E + 07	120	1.15	3.47	7.441	31.425	
47	350	2.00E + 07	120	1.00	3.56	5.699	25.046	
48	350	2.00E + 07	120	0.75	5.57	7.770	15.027	

[a] The low (−1 in coded form) and high values (+1 in coded form) of the factor variables in the face centered central composite design (the first 31 experiments)

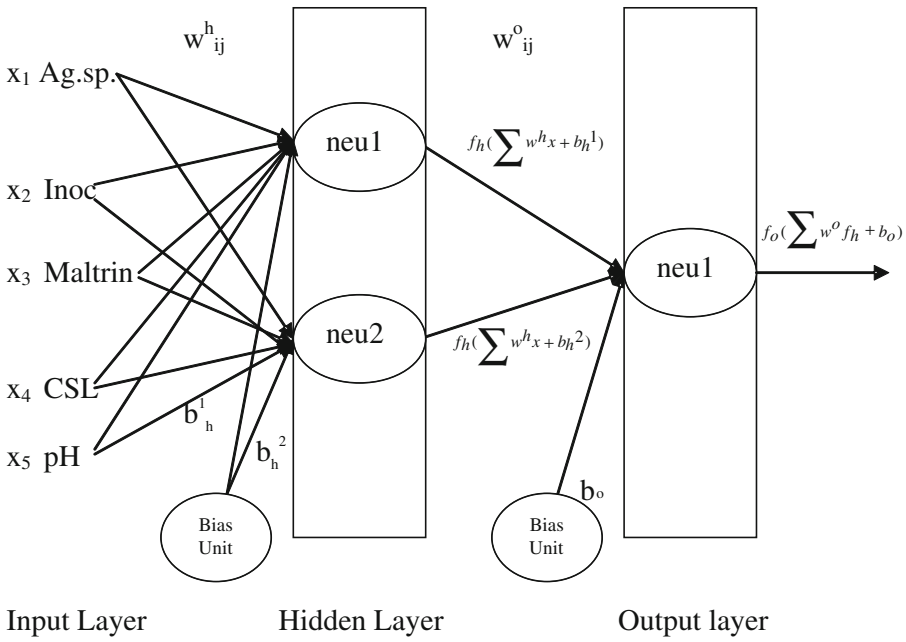

Fig. 1 Structure of the neural network. w_{ij}^{h}, the hidden weight from *j*th hidden neuron to *I*th input neuron (i.e., $j = 1, 2$ and $I = 1, 2\ldots5$); w_{ij}^{o}, the output weight from *j*th output neuron to *I*th hidden neuron (i.e., $j = 1$ and $I = 1, 2$); b_h^1, the bias weight to first hidden neuron; b_h^2, the bias weight to second hidden neuron. b_o, Bias weight to output neuron. *neu* Neuron, *Ag sp* agitation speed, *Inoc* inoculum amount, *f* activation function

Table 4 The boundaries of input variables used in ANN modeling

Boundaries	Agitation speed (rpm)	Inoculum ratio (total spore count)	Maltrin (g/l)	CSL (g/l)	pH
Minimum	150	1.25E + 04	25	0	3.27
Maximum	350	4.00E + 08	180	15	6.17

rate. The minimum and maximum levels of epochs and learning rate in the design were determined after a number of preliminary simulations: the levels chosen were 2 and 20 for the neurons, 1000 and 5000 for the epochs and 0.1 and 0.7 for the initial learning rate. It is not practical to keep the learning rate parameter constant as the optimal learning rate changes during the training process. Furthermore, a low learning rate makes the neural network learn very slowly, and a high learning rate makes the weights and biases diverge. We used the adaptive learning rate algorithm to train our neural networks [33]. An adaptive learning rate keeps the learning rate as large as possible while keeping the error stable.

A total of 18 simulations, including four center runs, were run. Before each simulation, the experimental data (Tables 2, 3) were randomly placed and then introduced to the network. The LOO cross-validation method was used to calculate the prediction errors of each model. As the responses of the design, the standard error of prediction percentage (SEP) and coefficient of determination (R^2) values for validation data were used.

$$\mathrm{SEP} = 100\frac{\sqrt{\frac{\sum(\hat{y}-y)^2}{m}}}{\bar{y}} \quad (4)$$

where $\sum(\hat{y}-y)^2$ is the sum of squared prediction errors and m is the number of predicted values. The $\hat{y}$ and $\bar{y}$ terms are the predictions and the mean value of the **y** vector, respectively. The SEP is a relative deviation of the mean prediction values and has the advantage of being not dependent on the magnitude and the number of the measurements. The appropriate neurons, epochs and learning rate were chosen to minimize the SEP and maximize the R^2 of the cross-validated data.

The ANN was implemented by Matlab 6.0 (The MathWorks, Natick, MA). Designs for simulations for the best network topology and minimum error were generated and data analyzed in Modde 7.0 (Umetrics, Umeå, Sweden).

Results and discussion

Sensitivity analysis

Fermentation data were collected according to two experimental designs, the details of which are given in the "Materials and methods". The factor variables were agitation speed, inoculum concentration, maltrin concentration and CSL concentration, and the response variables were PG activity and biomass concentration. Our results revealed that the most important factors directly associated with enzyme activity and biomass were the concentrations of maltrin and CSL ($P < 0.02$). Even though the agitation

speed and inoculum concentration were not found to be significant ($P > 0.1$), their interactions were significant ($P < 0.035$), and the inclusion of these variables and their interactions improved the regression (RSM) model. All four factor variables in RSM were also included in the neural network model. The results of this RSM analysis were considered as a sensitivity analysis that showed the importance of each input variable on the output. A number of published studies also considered RSM as the sensitivity analysis technique for providing insight into the main and interaction effects [34].

In the first and second optimization studies given in Tables 2 and 3, the pH values of final mold fermentation medium were also recorded at the end of each fermentation study. In order to study the effect of pH on the activity and biomass responses, we constructed general regression models using total of 91 data points. The R^2 values and LOF P value of models for activity and biomass are presented in Table 5. We found that the pH contribution led to a better explanation of the fermentation outputs and that it was an important factor, especially for the modeling of biomass. The MLR models are given as:

$$\begin{aligned}\text{Activity-MLR} = {} & 0.20 + 0.07x_1 + 0.69x_2 - 3.75x_3 \\ & - 0.45x_4 + 0.25x_5 - 0.27x_1^2 + 3.54x_2^2 \\ & + 0.32x_4^2 \end{aligned} \quad (5)$$

and

$$\begin{aligned}\text{Biomass-MLR} = {} & 16.58 + 0.32\,x_1 + 11.80x_2 + 0.46x_3 \\ & + 12.28x_4 - 1.01x_5 - 0.26x_1^2 - 1.53x_5^2 \\ & + 0.7x_1x_3 - 0.37x_1x_5 + 12.26x_2x_4 \\ & - 0.76x_3x_5 \end{aligned} \quad (6)$$

where x_1 is the agitation speed; x_2 is the inoculum ratio; x_3 is the concentration of maltrin; x_4 is the concentration of CSL; x_5 is the pH. The final pH of the fermentation medium, as an easy-to-measure variable, was taken as the fifth input variable in ANN model.

Table 5 Results of the multiple linear regression model with/without the pH parameter

MLR parameter	PG activity			Biomass		
	R^2	R^2 test	LOF P value	R^2	R^2 test	LOF P value
pH included	0.86	0.84	0.43	0.77	0.69	0.21
pH not included	0.83	0.8	0.07[a]	0.54	0.46	0.001[a]

PG Polygalacturonase, *MLR* multiple linear regression, *LOF* lack of fit

[a] Parameter is significant at the 5% (0.05) significance level

ANN modeling

The key issue in ANN modeling is to decide on the network topology. In the construction of an ANN model, several parameters, such as the number of hidden layers, the number of neurons in each layer, the transfer function in each layer, the epochs and the learning rate, can be optimized. Based on the results of our preliminary analysis of the mold fermentation modeling, we used the single-layer network and hyperbolic tangent function in the network models. In this study, the optimization of the network was achieved in terms of number of hidden layer neurons, number of epochs and the learning rate, with a central composite design for both network outputs: PG activity and biomass concentration. The number of hidden neurons is one of the most important parameters of ANN modeling, and the improper selection of hidden neurons results in over-fitting and under-fitting problems. A high number of neurons performs satisfactorily for training data but may fail for testing data (over-fitting), while a few hidden neurons cause unsatisfactory convergence (under-fitting). According to the Kolmogorov theorem, the number of neurons can be taken as 2N + 1, with N dimensional input vector as a starting point [35]. The usual practice is to decrease the number of neurons gradually. With the *A. sojae* fermentation system, the neurons can be taken as 11 initially. In the network optimization study, the number of neurons was changed between 2 and 20 (the center value is 11).

The results of the statistical analysis are given in Table 6 in terms of P values. Significant parameters have P values <0.1. For the PG activity data, the neurons and learning rate do not affect the R^2 of the cross-validation (R^2 test). However, in accordance with SEP, a low number of neurons, epochs and learning rate has an increasing effect on R^2. This also means that SEP values are minimum under these conditions. For the biomass data, all three parameters were found to be significant in terms of their interactions.

Table 6 Results of the experimental design for neural network structure: P values of factors for the outputs of the artificial neural network model

	PG activity	Biomass
Neuron (neu)	0.25[a]	0.08
Epochs (epo)	0.78[a]	0.09
learning rate (lea)	0.04	0.03
neu × neu	–	0.003
lea × lea	0.27[a]	–
neu × epo	0.3[a]	0.002
neu × lea	0.08	0.001
epo × lea	0.25[a]	0.03

[a] Insignificant parameter ($P > 0.1$)

We observed that high number of neurons (20) and epochs (5,000) in training results in high R^2 values of trained data and lower R^2 of the cross-validated data. The optimum network by which to attain a maximum R^2 of test data was determined to be nine neurons, 1000 epochs and a learning rate of 0.7. Different topologies of the network models for activity and biomass can be explained by comparing the R^2 values of the linear models and neural network models. In terms of activity, the R^2 values of the ANN and MLR models are close (0.84 vs. 0.86), whereas, for biomass, ANN yielded a higher R^2 value than MLR (0.83 vs. 0.77).

The cross-validation technique is used in empirical modeling techniques to determine the generalization power of the models. In those cases particularly where a large enough data set is not available for training and testing procedures, a LOO cross-validation can be used to estimate the prediction error from the learning data itself [36]. Each time, one of the data points (I) is left outside, and the remaining N − 1 data points are used to model the system. Then, the data point left outside is predicted with the existing model to see how it performs with a new set of input combination that was not used in the modeling step. The error is calculated by subtracting the predicted value of the Ith output data from its observed value. The same procedure is repeated with all of the input set–target pairs in the data. At the end of the process, the sum of squares of all N error components are calculated and expressed in terms of SEP. This is advantageous since model validation is performed with N data points that were not used in the training process. Instead of using only a certain amount of data in testing, all available data are used, especially if a low number of data points is available for both training and validation. Validation results for PG activity and biomass are presented in Figs. 2 and 3 and Table 7. Figure 2 presents the scatter plot of actual PG activity readings versus the ANN-predicted activity values: the ANN model predicts the enzyme activity with an R^2 of 0.84. Figure 3 shows the actual versus predictions for biomass data: the ANN model predicts actual biomass readings with an R^2 of 0.83.

The advantage of using ANN here is the ability of using data coming from two different designs of the same system. Even though data belong to different studies, ANN successfully found the relationship between inputs and outputs and estimated the validation data with high R^2 values. The same cannot be said for regression models of a RSM study since the empirical equation of a design belongs to the investigated region generated with that design and gives no guarantee of producing good estimates of outputs generated at different levels of input variables. The PG enzyme activity and biomass concentrations of the first optimization study, including its validation experiments, were predicted by the regression equations of the second optimization

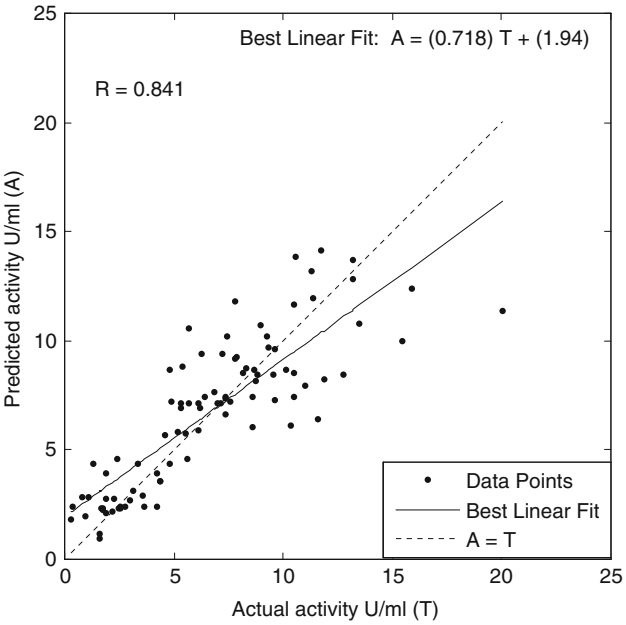

Fig. 2 Actual versus predicted polygalacturonase (*PG*) activity values of artificial neural network (*ANN*) model: 5-2-1 topology

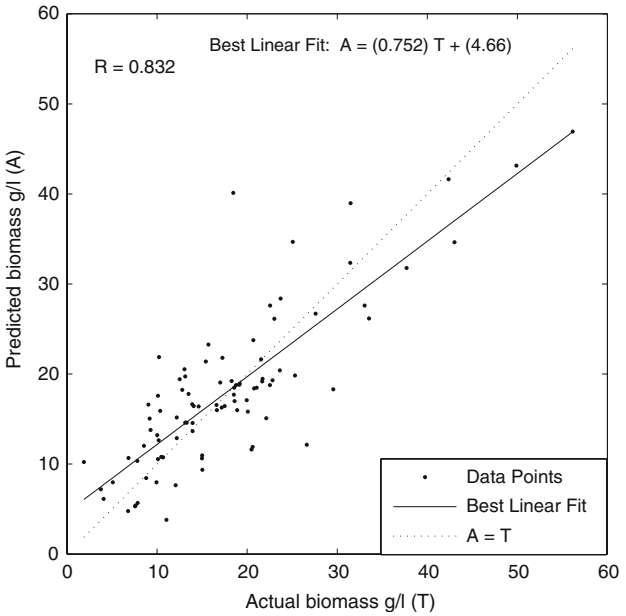

Fig. 3 Actual versus predicted biomass values of the ANN model: 5-9-1 topology

Table 7 Results of ANN models

Parameter	Network	Learning rate	Epochs	SEP	R^2 test
PG activity	5:2:1	0.1	1000	32.61	0.841
Biomass	5:9:1	0.7	1000	30.51	0.832

study (Eq. 1, 2) with low R^2 values, such as 0.06 and 0.02, respectively. When the ANN models were compared to the general MLR models (Eqs. 5, 6), biomass data were predicted with higher R^2 by the ANN model (Tables 5, 7).

The performance of ANN models in terms of R^2 are better than those of RSM models, which were performed beforehand. The ANN modeling, with its non-linear features, is superior to the RSM in estimating the fermentation outputs, which belong to different studies of that particular system. Therefore, neural networks can be considered to a practical prediction tool, especially for biological measurements, such as biomass and enzyme activity, that require laborious analytical procedures and produce a certain amount of laboratory waste. On the other hand, we conclude that the use of fermentation data collected based on experimental design techniques clarified the interactions between inputs and outputs of the black box, i.e., neural network. As a result, the integration of experimental design and ANN techniques help researchers understand the complex systems better.

References

1. Aguado D, Ferrer A, Seco A, Ferrer J (2006) Comparison of different predictive models for nutrient estimation in a sequencing batch reactor for wastewater treatment. Chemom Intell Lab Syst 84:75–81. doi:10.1016/j.chemolab.2006.03.009
2. Barthus RC, Mazo LH, Poppi RJ (2005) Simultaneous determination of vitamins C, B6 and PP in pharmaceutics using differential pulse voltammetry with a glass carbon electrode and multivariate calibration tools. J Pharm Biomed Anal 38:94–99
3. Chen CR, Ramaswamy HS (2003) Analysis of critical control points in deviant thermal processes using artificial neural networks. J Food Eng 57:225–235. doi:10.1016/S0260-8774(02)00301-1
4. Garcia-Gimeno RM, Hervas-Martinez C, Rodriguez-Perez R, Zurera-Cosano G (2005) Modeling the growth of *Leuconostoc mesenteroides* by artificial neural networks. Int J Food Microbiol 105:317–332. doi:10.1016/j.ijfoodmicro.2005.04.013
5. Kiviharju K, Salonen K, Leisola M, Eerikäinen T (2006) Modeling and simulation of *Streptomyces peucetius* var. *caesius* N47 cultivation and ε-rhodomycinone production with kinetic equations and neural networks. J Biotechnol 126:365–373. doi:10.1016/j.jbiotec.2006.04.034
6. Kenedy M, Krouse D (1999) Strategies for improving fermentation medium performance: a review. J Ind Microbiol Biotechnol 23:456–475. doi:10.1038/sj.jim.2900755
7. Torrecilla JS, Aragon JM, Palancar MC (2005) Modeling the drying of a high-moisture solid with an artificial neural network. Ind Eng Chem Res 44:8057–8066. doi:10.1021/ie0490435
8. Moreira G, Micheloud GA, Beccaria AJ, Goicoechea HC (2007) Optimization of the *Bacillus thuringiensis* var. *kurstaki* HD-1 δ-endotoxins production by using experimental mixture design and artificial neural networks. Biochem Eng J 35:48–55. doi:10.1016/j.bej.2006.12.025
9. Razmi-Rad E, Ghanbarzadeh B, Mousavi SM, Emam-Djomeh Z, Khazaei J (2007) Prediction of rheological properties of Iranian bread dough from chemical composition of wheat flour by artificial neural networks. J Food Eng 81:728–734. doi:10.1016/j.jfoodeng.2007.01.009
10. Bas D, Boyaci I (2007) Modeling and optimization II: comparison of estimation capabilities of response surface methodology with artificial neural networks in a biochemical reaction. J Food Eng 78:846–854. doi:10.1016/j.jfoodeng.2005.11.025
11. Huang J, Mei L, Xia J (2006) Application of artificial neural network coupling particle swarm optimization algorithm to biocatalytic production of GABA. Biotechnol Bioeng 96:924–931. doi:10.1002/bit.21162
12. Desai KM, Akolkar SK, Badhe YP, Tambe SS, Lele SS (2006) Optimization of fermentation media for exopolysaccharide production from *Lactobacillus plantarum* using artificial intelligence-based techniques. Process Biochem 41:1842–1848. doi:10.1016/j.procbio.2006.03.037
13. Alonso-Salces R, Herrero C, Barranco A, Lopez-Marquez D, Berrueta L, Gallo B, Vicente F (2006) Polyphenolic compositions of basque natural ciders: chemometric study. Food Chem 97:438–446. doi:10.1016/j.foodchem.2005.05.022
14. Hervas-Martinez C, Garcia-Gimeno R, Martinez-Estudillo A, Martinez-Estudillo F, Zurera-Cosano G (2006) Improving microbial growth prediction by product unit neural networks. J Food Sci 71:M31–M38. doi:10.1111/j.1750-3841.2006.00029.x
15. Yuste A, Dorado P (2006) A neural network approach to simulate biodiesel production from waste olive oil. Energy Fuels 20:399–402. doi:10.1021/ef050226t
16. Esnoz A, Periago PM, Conesa R, Palop A (2006) Application of artificial neural networks to describe the combined effect of pH and NaCl on the heat resistance of *Bacillus stearothermophilus*. Int J Food Microbiol 106:153–158. doi:10.1016/j.ijfoodmicro.2005.06.016
17. Alonso-Salces RM, Herrero C, Barranco A, Berrueta LA, Gallo B, Vicente F (2005) Classification of apple fruits according to their maturity state by the pattern recognition analysis of their polyphenolic compositions. Food Chem 93:113–123. doi:10.1016/j.foodchem.2004.10.013
18. Hongwen C, Baishan F, Zongding H (2005) Optimization of process parameters for key enzymes accumulation of 1, 3-propanediol production from *Klebsiella pneumaniae*. Biochem Eng J 25:47–53. doi:10.1016/j.bej.2005.03.011
19. Spanila M, Pazourek J, Farkova M, Havel J (2005) Optimization of solid-phase extraction using artificial neural networks in combination with experimental design for determination of resveratrol by capillary zone electrophoresis in wines. J Chromatogr A 1084:180–185. doi:10.1016/j.chroma.2004.10.007
20. Pazourek J, Gajdosova M, Spanila M, Farkova M, Novotna K, Havel J (2005) Analysis of polyphenols in wines: correlation between total phenolic content and antioxidant potential from photometric measurements: prediction of cultivars and vintage from capillary zone electrophoresis fingerprints using artificial neural network. J Chromatogr A 1081:48–54. doi:10.1016/j.chroma.2005.02.056
21. Dutta J, Dutta P, Banerjee R (2004) Optimization of culture parameters for extracellular protease production from newly isolated *Pseudomonas* sp. using response surface and artificial neural network models. Process Biochem 39:2193–2198. doi:10.1016/j.procbio.2003.11.009
22. Perez-Magarino S, Ortega-Heras M, Gonzalez-San Jose ML, Boger Z (2004) Comparative study of artificial neural network and multivariate methods to classify Spanish DO rose wines. Talanta 62:983–990. doi:10.1016/j.talanta.2003.10.019
23. Irudayaraj J, Xu F, Tewari J (2003) Rapid determination of invert cane sugar adulteration in honey using FTIR spectroscopy and multivariate analysis. J Food Sci 68:2040–2045. doi:10.1111/j.1365-2621.2003.tb07015.x
24. Coleman MC, Buck KKS, Block DE (2003) An integrated approach to optimization of *Escherichia coli* Fermentations using historical data. Biotechnol Bioeng 84:274–285. doi:10.1002/bit.10719

25. Castellanos JA, Palancar MC, Aragon JM (2002) Designing and optimizing a neural network for the modeling of a fluidized-bed drying process. Ind Eng Chem Res 41:2262–2269. doi:10.1021/ie000950t
26. Iizuka K, Aishima T (1997) Soy sauce classification by geographic region based on NIR spectra and chemometrics pattern recognition. J Food Sci 62:101–104. doi:10.1111/j.1365-2621.1997.tb04377.x
27. Sun L, Danzer K, Thiel G (1997) Classification of wine samples by means of artificial neural networks and discrimination analytical methods. Fresenius J Anal Chem 359:143–149. doi:10.1007/s002160050551
28. Alkorta I, Garbisu C, Llama MJ, Serra JL (1998) Industrial application of pectic enzymes: a review. Process Biochem 33:21–28. doi:10.1016/S0032-9592(97)00046-0
29. Nighojkar S, Phanse Y, Sinha D, Nighojkar A, Kumar A (2006) Production of polygalacturonase by immobilized cells of *Aspergillus niger* using orange peel as inducer. Process Biochem 41:1136–1140. doi:10.1016/j.procbio.2005.12.009
30. Tari C, Gogus N, Tokatli F (2007) Optimization of biomass, pellet size and polygalacturonase production by *Aspergillus sojae* ATCC 20235 using response surface methodology. Enzyme Microb Technol 40:1108–1116. doi:10.1016/j.enzmictec.2006.08.016
31. Panda T, Naidu GSN, Sinha J (1999) Multiresponse analysis of microbiological parameters affecting the production of pectolytic enzymes by *Aspergillus niger*: a statistical approach. Process Biochem 35:187–195. doi:10.1016/S0032-9592(99)00050-3
32. Nguyen D, Widrow B (1990) Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. Proc IJCNN 3:21–26. doi:10.1109/IJCNN.1990.137819
33. Masters T (1993) Practical neural network recipies in C++. Academic Press, San Francisco
34. Gevrey M, Dimopoulos I, Lek S (2006) Two-way interaction of input variables in the sensitivity analysis of neural network models. Ecol Modell 195:43–50. doi:10.1016/j.ecolmodel.2005.11.008
35. Molga EJ (2003) Neural network approach to support modeling of chemical reactors: problems, resolutions, criteria of application. Chem Eng Process 42:675–695. doi:10.1016/S0255-2701(02)00205-2
36. Mevik B, Cederkvist HR (2004) Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). J Chemometr 8:422–429. doi:10.1002/cem.887