ANN AND RSM BASED OPTIMIZATION OF CELLULASE PRODUCTION BY *HYPOCREA* SP. Z28 BY SUBMERGED FERMENTATION

YU ZHANG,[*] XIAOHUAN ZHANG,[*,**] WEI QI,[*] JINGLIANG XU,[*]
ZHENHONG YUAN[*] and ZHONGMING WANG[*]

[*]*Guangzhou Institute of Energy Conversion, Key Laboratory of Renewable Energy, Chinese Academy of Sciences; Guangdong Provincial Key Laboratory of New and Renewable Energy Research and Development, Guangzhou 510640, P.R. China*
[**]*University of Chinese Academy of Sciences, Beijing 100049, P.R. China*
✉ *Corresponding author: Jingliang Xu, xjl@ms.giec.ac.cn*

Received September 14, 2016

Response surface methodology (RSM) and artificial neural network (ANN) were used to simulate and optimize cellulase production by *Hypocrea* sp. Z28 by submerged fermentation. Results showed ANN had higher simulation accuracy than RSM. Cellulase production optimized by RSM was 5.48 U/mL, while the corresponding experimental value was 5.67±0.32 U/mL. Using ANN as a prediction function, a maximum cellulase production of 5.96 U/mL was searched by the genetic algorithm, and the corresponding experimental value was 6.01±0.43 U/mL. Compared to RSM, ANN brought higher and more accurate cellulase production values. The application of ANN to optimize cellulase production proved successful.

Keywords: cellulase production, submerged fermentation, artificial neural network-genetic algorithm, bioprocess optimization, response surface methodology, *Hypocrea* sp.

INTRODUCTION

The application of cellulase in many industries, including bioenergy and biobased materials, is seriously restricted by its expensive production.[1] Cellulase is mainly produced by microorganisms,[2] and in order to reduce its production costs, the first step is to screen out a high cellulase-production strain, no matter what techniques are used,[3] and then to optimize the cellulase production from the obtained strain.[4-6]

The conventional one-factor-at-a-time approach of optimization is not only laborious, but also ignores the combined interaction of the factors.[7-8] To solve this problem, model-based optimization techniques have been proposed. Due to the high complexity of microbial cellulase production, it is impossible to build a mechanistic model for the fermentation process. Thus, it is better to use empirical models to identify the relationship between cellulase production and fermentation conditions.[4] Such a model could be built by RSM (response surface methodology). RSM is a frequently used technique for building models and determining the optimal process conditions.[7-9] During RSM, a polynomial expression could be obtained from non-linear regression analysis of a pre-designed experimental matrix. In contrast, artificial neural network (ANN) was thought as another superior tool.[10-12] ANN could simulate an arbitrary bioprocess to any precision, and has made much progress in optimizing many bioprocess behaviors.[13-14]

In this study, ANN and RSM were used to simulate and optimize cellulase production by submerged fermentation.

EXPERIMENTAL

Materials

Hypocrea sp. Z28 was isolated in our laboratory previously. It was maintained on potato dextrose agar (PDA) slants and stored at 4 °C.

Rice straw was obtained from a farm in a local harvest and dried naturally. The dried straw was milled to small particles and the nominal sizes of <80 mesh were collected and used as the sole carbon source for microbial cellulase production.

Submerged fermentation

In order to prepare inoculums, Z28 slant cultures were incubated on liquid PDA medium at 30 °C and 110 rpm for 72 h. Then, 2% (v/v) inoculums were transferred into submerged fermentation medium for cellulase production at 120 rpm.

Temperature, pH and time are three important culture conditions for microbial fermentation, so the three factors were adjusted to improve cellulase production from Z28 by submerged fermentation. The range of each factor and their experiment design matrix are shown in Table 1.

Submerged fermentation medium (g/L): rice straw 5.0, $(NH_4)_2SO_4$ 2.5, KH_2PO_4 2.0, $MgSO_4 \cdot 7H_2O$ 0.3, $CaCl_2$ 0.3. The used buffer was $NaHPO_4$-$C_6H_8O_7$ buffer (0.2 M).

Cellulase assay

A rolled filter paper strip (1×6 cm, about 50 mg), as well as 1 mL acetate buffer (0.2 M, pH 5.0), was incubated with 1 mL diluted cellulase solution. The reaction was carried out in a 15×100 mm test tube at 50 °C without stirring. After 60 min, the DNS (3,5-Dinitrosalicylic acid) method was used to determine the reducing sugars (glucose equivalents) produced. One unit of cellulase activity was defined as the amount of enzyme required for the formation of 1 μmol glucose equivalents per minute.

Response surface methodology

The relationship amongst the three factors was expressed by the following second-order equation:

$$Y = a_0 + \sum_{i=1}^{3} a_i X_i + \sum_{i=1}^{3} a_{ii} X_i^2 + \sum_{i=1}^{3} \sum_{j=i+1}^{3} a_{ij} X_i X_j \quad (1)$$

where Y is the predicted cellulase production, a_0 is constant, a_i, a_{ii} and a_{ij} are the regression coefficients of the RSM model, X_i and X_j are the factor variables. Statistical analysis of the data from Box-Behnken Design (BBD) was performed to determine the values of a_0, a_i, a_{ii} and a_{ij}.

Artificial neural network

In this study, ANN consisted of only one hidden (four neurons) layer. There were three (temperature, pH and time) and one neuron (cellulase production) in the input and output layers of ANN, respectively (Fig. 1).

Table 1
BBD matrix of three factors and experimentally determined cellulase activity *versus* RSM and ANN simulated values

Trial	Levels	Factors			Y (U/ml)		
		X_1	X_2	X_3			
	-1.00	25 °C	4.5	3 days			
	0.00	29 °C	6.0	6 days	Experimental	RSM	ANN
	+1.00	33 °C	7.5	9 days			
1		+1.00	-1.00	0.00	3.39 ± 0.75	3.01	3.26
2		+1.00	0.00	-1.00	2.63 ± 0.20	2.75	2.78
3		+1.00	0.00	+1.00	1.36 ± 0.57	1.89	1.35
4		+1.00	+1.00	0.00	2.08 ± 0.44	1.81	2.26
5		0.00	-1.00	-1.00	2.48 ± 0.34	2.74	2.65
6		0.00	-1.00	+1.00	4.46 ± 0.73	4.31	4.63
7		0.00	0.00	0.00	5.26 ± 1.11	5.26	5.23
8		0.00	0.00	0.00	5.26 ± 1.11	5.26	5.23
9		0.00	0.00	0.00	5.26 ± 1.11	5.26	5.23
10		0.00	0.00	0.00	5.26 ± 1.11	5.26	5.23
11		0.00	0.00	0.00	5.26 ± 1.11	5.26	5.23
12		0.00	+1.00	-1.00	2.13 ± 0.22	2.28	1.99
13		0.00	+1.00	+1.00	2.08 ± 0.30	1.82	2.07
14		-1.00	-1.00	0.00	3.76 ± 0.50	4.04	3.83
15		-1.00	0.00	-1.00	2.63 ± 0.47	2.10	2.56
16		-1.00	0.00	+1.00	4.19 ± 0.84	4.07	4.17
17		-1.00	+1.00	0.00	1.93 ± 0.77	2.31	1.84

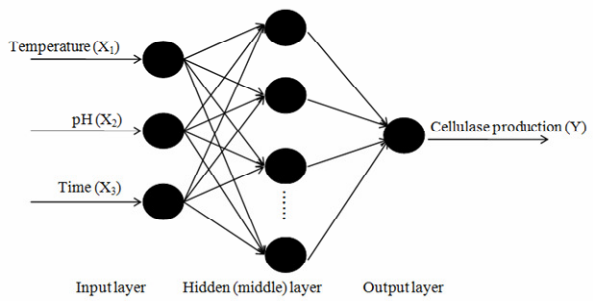


Figure 1: Schematic representation of ANN modelling the relationship between cellulase production and three factors (temperature, pH and time)

All the data (input and output ones) in Table 1 were scaled as follows:

$$X_i^* = 2\frac{X_i - X_{i,\min}}{X_{i,\max} - X_{i,\min}} - 1 \quad (2)$$

$$Y^* = \frac{Y - 0}{10 - 0}$$

where X_i^* and Y^* are new scaled data of input and output layers, respectively.

Back-propagation algorithm was used to train a random ANN by feeding the scaled data. In the ANN, the transfer functions of the hidden and output layers are tangent sigmoid and pure linear functions, respectively. The mean square error between actual and expected output neurons was calculated and propagated backward through the network. Then, the weight of each layer was adjusted accordingly. Backward propagation did not stop until the mean square error got to 1×10^{-4}.

Genetic algorithm

Using the trained ANN as the fitness function, a genetic algorithm (GA) was implemented to search the maximum output (cellulase production). The GA procedures consisted of the following steps:

- Assign a fitness value to each individual of a randomly generated population for guiding the search;
- Select individuals with higher fitness values and let them undergo genetic operation, such as crossover and mutation;
- Use the newly generated child population as the parent population for the next generation and then treat them with the same evolutional process continuously until the designed generation number is reached.[15-16]

Working parameters, namely total number of generation, population size, number of binary coded variables, cross-over probability and mutation probability were 50, 20, 3, 0.4 and 0.005, respectively.

Software

RSM was performed by Design Expert 7.0. Artifical neural network and genetic algorithm were developed by Matlab R2010b.

RESULTS AND DISCUSSION

RSM-based simulation and optimization

Based on the statistical analysis of the experimental data from Box-Behnken Design in Table 1, a quadratic polynomial was established to identify the relationship between activity yield and three culture conditions as follows:

$$Y = 5.26 + 0.38X_1 - 0.73X_2 + 0.28X_3 - 0.13X_1X_2 + 0.71X_1X_3 - 0.51X_2X_3 - 1.28X_1^2 - 1.19X_2^2 - 1.28X_3^2 \quad (3)$$

where Y is cellulase production, and X_1, X_2 and X_3 represent temperature, pH and time in coded values, respectively.

With t and *P* values shown in Table 2, the significance of the three factors, their interaction and quadratic terms could be considered as: $X_2 > X_1 > X_3$, $X_1X_3 > X_2X_3 > X_1X_2$ and $X_3^2 > X_1^2 > X_2^2$, respectively. The larger the t-value and the smaller the p-value was, the higher the significance of the corresponding coefficient.

The analysis of variance is shown in Table 3. The F value of the RSM regression equation is larger than $F_{0.01}$ (9.5), which indicates that the variance in RSM, in this case, is very significant. It was concluded from the F-test of linearity square and interaction terms that the main effects and the interaction of the examined three factors were significant. The effect of the three factors on cellulase production was very complicated. It was impossible to obtain the maximum cellulase production using only the one-factor-at-a-time approach. Model-based optimization in this study was required.

Based on Equation 3, the 3D response surface diagrams are presented in Figure 2. From the 3D diagrams, it is easy to understand the interactions between two factors and cellulase production, and also to locate their optimum levels. The obtained surfaces were very convex and symmetric,

suggesting that there were well-defined optimum operating conditions.

Table 2
Standard errors, *t* and *P* values of coefficients of regression equation

Model term	Standard errors	*t*	*P*
Intercept	0.19	28.31	< 0.001
X_1	0.15	2.60	0.036
X_2	0.15	-4.99	0.002
X_3	0.15	1.89	0.101
X_1X_2	0.21	-0.63	0.551
X_1X_3	0.21	3.41	0.011
X_2X_3	0.21	-2.44	0.045
X_1^2	0.20	-6.31	< 0.001
X_2^2	0.20	-5.89	0.001
X_3^2	0.20	-6.32	< 0.001

Table 3
RSM-based analysis of variance for the experimental data of the CCD

Source	Degree of freedom	Square sum	Mean square	F	*P*
Regression	9	31.26	3.47	20.12	< 0.001
Linearity	3	6.09	2.03	11.75	0.004
Square	3	22.08	7.36	42.62	< 0.001
Interaction	3	3.10	1.03	5.98	0.024
Residual error	7	1.21	0.17		
Lack of fit	3	1.21	0.40	*	*
Pure error	4	0.00	0.00		
Total	16	32.47			

After calculating the first-order partial derivate of Equation 3, three quadratic equations with three variables were obtained. By solving the equation set, the predicted maximum cellulase production was 5.48 U/mL, where X_1 = 29.9 °C, X_2 = 5.4 and X_3 = 6.8 days. Under these conditions, three replicated experiments were carried out. The obtained experimental cellulase production was 5.67 ± 0.32 U/mL.

ANN-based simulation

Via limited trials, an ANN was built successfully for simulating cellulase production. The weight and threshold values of each layer, which determined the structure of the built ANN, were as follows:

$$net.iw\{1\} = \begin{pmatrix} -0.1215 & 2.5161 & 1.7782 \\ -1.8766 & 0.0116 & -1.1379 \\ -0.2845 & 0.4047 & 2.0472 \\ -1.6421 & -0.6609 & 1.9489 \end{pmatrix} \quad (4)$$

$$net.lw\{2\} = (-0.2165 \quad 0.0108 \quad 0.2341 \quad -0.1412)$$

$$net.b\{1\} = (-1.4122 \quad 0.7468 \quad 0.2829 \quad -1.8170)^T$$

$$net.b\{2\} = (0.1252)^T$$

ANN-based optimization by GA

Figure 3 shows the evolution of the algorithm with successive generations. Starting from 3.43 U/mL, the average cellulase production apparently increased until about the 10th generation and was 5.75 U/mL at the end of 100 generations. The maximum cellulase production also increased quickly for the first 10 generations and reached 5.96 U/mL at the 15th generation, then kept invariant. Thus, the maximum cellulase production obtained from ANN could be considered as 5.96 U/mL. The corresponding experimental conditions were the following: X_1 = 29 °C, X_2 = 5.5 and X_3 = 7.2 days, where the experimental cellulase production was 6.01 ± 0.43 U/mL.

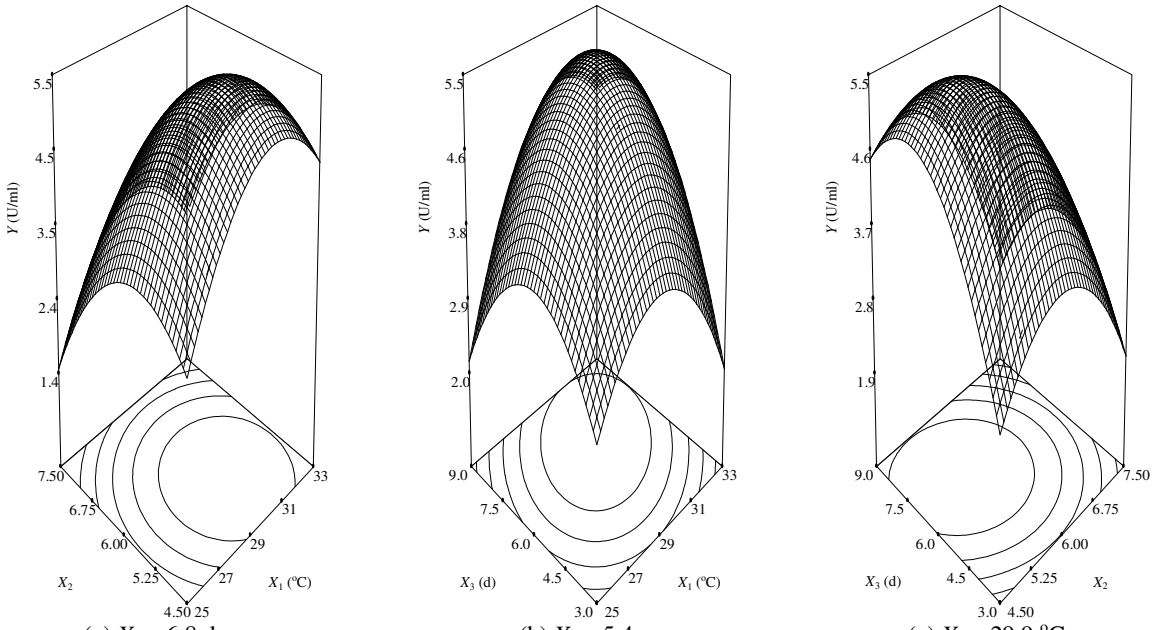
Comparison of ANN and RSM

Cellulase production simulation results obtained by RSM and ANN are shown in Table 1. The experimentally determined and ANN simulated values were almost identical, as compared to the values simulated by the RSM model. Except trials 2, 6 and 7~11 (central

points), the derivations produced by RSM were all larger than those performed by ANN.

The correlation coefficient, mean absolute, relative error, root-mean-square error and variance of the RSM-based simulation were 0.96, 0.21, 0.09%, 0.28 and 0.08, respectively, whereas for ANN, the corresponding values were 0.99, 0.08, 0.03%, 0.10 and 0.01. Mean absolute/relative error, root-mean-square error and variance of RSM were apparently higher than those of ANN, and *vice versa* for the correlation coefficient. These parameters are always used for evaluating the simulation performance of a model. The smaller the mean deviation, standard deviation and root-mean-square error and the larger the correlation coefficient were, the more precise the simulation. This stated further the simulation accuracy of ANN was higher than that of RSM. For optimization, ANN gave 6.01 U/mL maximum cellulase production, compared to 5.67 U/ml obtained from RSM. Similar results were obtained when simulating and optimizing other processes.[16-19]

The advantage of RSM resides in its ability to evaluate the factor contributions/significances according the regression analysis of factorial experiments, and thereby it can reduce the complexity of the process.[20-21] However, because of its limitation in simulating the data of an irregular experimental domain, RSM could only exhibit a low order non-linear behaviour to a regular experimental region.[22] The effective use of RSM requires a narrow search window (if we shrink the search window narrow enough, linear correlation may also suffice). The search process is highly dependent on the search space. If the search space is too wide, either extra experiments or good priory knowledge of the system is necessary for RSM simulation.[17]



(a) $X_3 = 6.8$ days (b) $X_2 = 5.4$ (c) $X_1 = 29.9$ °C

Figure 2: Three-dimensional response surfaces of cellulase production showing the interactions among temperature, pH and time

In contrast, ANN would not suffer from the limitation of experimental design and efficient ANN simulation requires relatively few experiments. Of course, the accuracy would be higher when a large number of experiments are used to create the non-linear behaviours.[23] Moreover, ANN represents the non-linearity in a much better way than RSM and can inherently capture the arbitrary form of non-linearity.[10-11] It can easily overcome the limitation of RSM discussed above. Thus, in the case of ANN, a more liberal search space can be chosen; although the correlation in that search space is more complex than an equation of higher degree.[17] On the other hand, ANN provides little information about the contribution/significance of each factor, unless further analysis has been done. Besides, the development of ANN requires a large number of

iterative calculations, whereas it is only a single step calculation for RSM.[22]

CONCLUSION

A maximum cellulase production value of 6.01 U/mL was obtained from ANN simulation, while RSM yielded a maximum value of only 5.67 U/mL. Therefore, ANN could be considered as a superior technique compared to RSM, as demonstrated in this study. It is believed that ANN-based optimization could be applied in more complicated systems due to its advanced non-linear analysis and mechanistic independence.

ACKNOWLEDGMENTS: This work was funded by the National Natural Science Foundation of China (21506215 and 51561145015).

REFERENCES

[1] V. Juturu and J. C. Wu, *Renew. Sustain. Energ. Rev.*, **33**, 188 (2014).
[2] R. K. Sukumaran, R. R. Singhania and A. Pandey, *J. Sci. Ind. Res.*, **64**, 832 (2005).
[3] F. Xu, J. Wang, S. Chen, W. Qin, Z. Yu *et al.*, *Appl. Biochem. Microbiol.*, **47**, 53 (2011).
[4] D. P. Maurya, D. Singh, D. Pratap and J. P. Maurya, *J. Environ. Biol.*, **33**, 5 (2012).
[5] D. Dutt and A. Kumar, *Cellulose Chem. Technol.*, **48**, 285 (2014).
[6] K. Matkar, D. Chapla, J. Divecha, A. Nighojkar and D. Madamwar, *Int. Biodeter. Biodegrad.*, **78**, 24 (2013).
[7] D. Bas and I. H. Boyaci, *J. Food Eng.*, **78**, 836 (2007).
[8] M. A. Bezerra, R. E. Santelli, E. P. Oliveira, L. S. Villar and L. A. Escaleira, *Talanta*, **76**, 965 (2008).
[9] G. Coman and G. Bahrim, *Cellulose Chem. Technol.*, **45**, 245 (2011).
[10] J. S. Almeida, *Curr. Opin. Biotechnol.*, **13**, 72 (2002).
[11] A. Tompos, J. L. Margitfalvi, E. Tfirst and K. Heberger, *Appl. Catal. A-Gen.*, **324**, 90 (2007).
[12] G. Astray, B. Gullon, J. Labidi and P. Gullon, *Ind. Crop. Prod.*, **92**, 290 (2016).
[13] A. Saraceno, S. Sansonetti, V. Calabro, G. Iorio and S. Curcio, *Chem. Biochem. Eng. Q.*, **25**, 461 (2011).
[14] A. Meszaros, A. Rusnak and K. Najim, *Chem. Biochem. Eng. Q.*, **11**, 81 (1997).
[15] M. Izadifar and M. Z. Jahromi, *J. Food Eng.*, **78**, 1 (2007).
[16] L. He, Y. Q. Xu and X. H. Zhang, *Biotechnol. Bioeng.*, **100**, 250 (2008).
[17] K. M. Desai, S. A. Survase, P. S. Saudagar, S. S. Lele and R. S. Singhal, *Biochem. Eng. J.*, **41**, 266 (2008).
[18] J. R. Dutta, P. K. Dutta and R. Banerjee, *Process Biochem.*, **39**, 2193 (2004).
[19] A. Singh, A. Majumder and A. Goyal, *Bioresour. Technol.*, **99**, 8201 (2008).
[20] A. Majumder and A. Goyal, *Bioresour. Technol.*, **99**, 3685 (2008).
[21] W. T. Su, W. J. Chen and Y. F. Lin, *Appl. Microbiol. Biotechnol.*, **84**, 271 (2009).
[22] A. K. Lakshminarayanan and V. Balasubramanian, *T. Nonferr. Metal. Soc.*, **19**, 9 (2009).
[23] S. D. Balkin and D. K. J. Lin, *Commun. Stat.-Theory Methods*, **29**, 2215 (2000).