Appl Biochem Biotechnol (2008) 144:225–235
DOI 10.1007/s12010-007-8017-y

Artificial Neural Network-Genetic Algorithm Approach to Optimize Media Constituents for Enhancing Lipase Production by a Soil Microorganism

M. A. Haider · K. Pakshirajan · A. Singh · S. Chaudhry

Received: 12 June 2007 / Accepted: 26 July 2007 /
Published online: 14 August 2007

Abstract Results of lipase production by a soil microorganism, expressed in terms of lipolytic activities of the culture were modeled and optimized using artificial neural network (ANN) and genetic algorithm (GA) techniques, respectively. ANN model, developed based on back propagation algorithm, were highly accurate in predicting the system with coefficient of determination (R^2) value being close to 0.99. Optimization using GA, based on the ANN model developed, resulted in the following values of the media constituents: 9.991 ml/l oil, 0.100 g/l $MgSO_4$ and 0.009 g/l $FeSO_4$. And a maximum value of 7.69 U/ml of lipolytic activity at 72 h of culture was obtained using the ANN-GA method, which was found to be 8.8% higher than the maximum values predicted by a statistical regression-based optimization technique-response surface methodology.

Keywords Artificial neural networks · Genetic algorithms · Response surface methodology · Optimization · Lipase production · Soil microorganism

Introduction

Despite abundance of information on fermentation conditions for lipase production by known and well-characterized microbes, there is always an ever-growing requirement of new sources of lipases for different applications. This consequently increases the demand for cost effective production of lipases from these newer sources. One of the major strategies in increasing the yield of lipase is by optimization of media constituents. Generally, optimization studies are carried out by conventional and classical methods, in which one factor is varied randomly keeping the other factors at an unspecified constant level. However, these techniques do not depict the combined effect of factors involved. In addition, these methods are time consuming and require a lot of experiments to determine optimum levels, which are unreliable. These limitations of a classical method can be eliminated by optimizing all the affecting factors collectively by nonconventional methods

M. A. Haider · K. Pakshirajan (✉) · A. Singh · S. Chaudhry
Department of Biotechnology, Indian Institute of Technology Guwahati, Guwahati 781039, India
e-mail: pakshi@iitg.ernet.in

such as response surface methodology [1] or artificial-intelligence-based methods such as artificial neural network combined with genetic algorithms [2]. Such methods are particularly significant in systems where the function relating output to input variables is not known. For example, in the system that we have adopted, no function is known a priori that can define the lipolytic activity in terms of the media composition.

Artificial Neural Networks

Artificial neural networks (ANN) are gaining importance in modeling of complex biological systems due to its excellent performance in pattern recognition, and in the modeling of nonlinear relationships involving a multitude of variables in place of conventional and statistical techniques [3, 4].

ANN architecture mimics the learning process of human brain. The basic architecture of ANN involves interconnected neurons, which are defined in three distinct categories: input layer neurons, output layer neurons and hidden layer neurons as shown in Fig. 1. The input data are presented through input layer neurons, and the response of the input data is presented at output layer neurons. Neurons are connected by scalar functions known as weights that take part in the learning process of networks. In back propagation algorithm, which is widely used in training of ANNs, a series of input and output data is presented to the system. Each hidden layer neuron and output layer neuron process this input data by

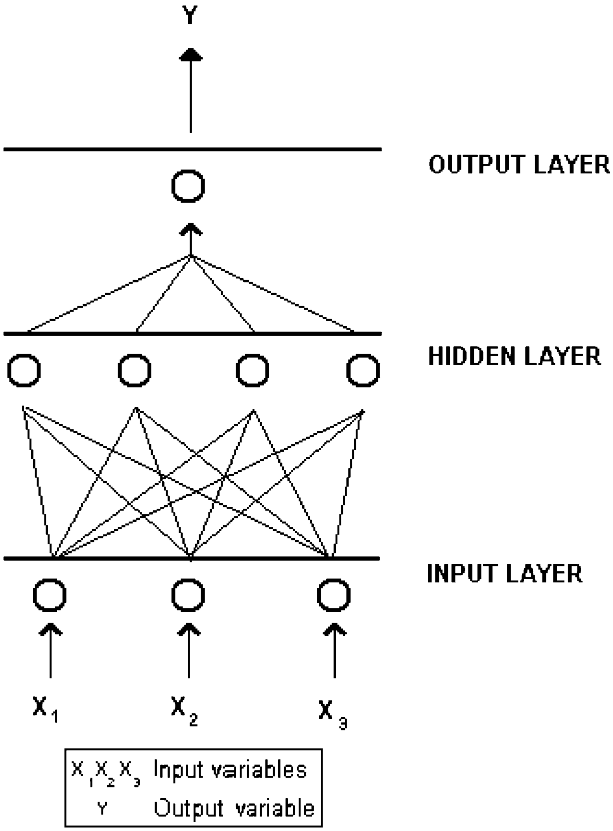


Fig. 1 Schematic representation of artificial neural network architecture

multiplying its corresponding weights, and using a transfer function. The S-shaped sigmoidal curve, which is mostly used as the transfer function is given as:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

The learning of the network is carried out through adjusting the weights by continuous iterations and minimizing the error between experimentally measured response and ANN-model-predicted response [5]. ANNs such as the three-layer back propagation network have been proven to be universal function approximators [6,7]. ANNs have already been applied to solve and predict a variety of problems in biological systems. The back propagation network is the most prevalent supervised ANN learning model, which uses the gradient decent algorithm to correct the weights between interconnected neurons [8].

Good network architecture generally involves selection of the most dependable values of network internal parameters like: number of hidden neurons in each layer, the activation function, the learning rate of the network, epoch size, momentum term, tolerance and training count. The best values for these parameters are usually estimated by a trial and error approach. But only for the parameter on number of hidden layer neurons, certain guidelines have been proposed for choosing its upper limit [5]; in this study, we chose its value between 4 and 8, which is less than the suggested upper limit. For choosing the values of the other network parameters, the most recommended trial and error approach has been adopted. A detailed study on the effect of internal network parameters on the performance of back propagation-neural network, and the procedure for selecting the best network topology may be found in Maier and Dandy (1998) [8].

An ANN model thus developed for a biological system can be used as an objective function for optimization studies. Conventional optimization techniques cannot provide any help in such a case as ANN models give complex functions, which are difficult to differentiate; however, ANN model equation can easily be optimized with a nontraditional optimization technique, such as using genetic algorithms [2].

Genetic Algorithms

Genetic algorithms (GA) mimic the survival of the fittest principle of nature while searching in an objective function that helps in its natural maximization. GA use string coding of variables for dividing the search space into discrete ones, which further helps in searching for global maxima in the entire solution space. In ANN-GA method of optimization, fitness function of input string is calculated using an ANN model equation. Reproduction, mutation and crossover operations are then performed to look for different search directions for optimal solutions. Once the convergence criterion is satisfied, the algorithm is terminated. A complete description on the working of a simple genetic algorithms used in this study has been given in Deb (1995) [9].

The literature is replete with studies that demonstrate the effectiveness of response surface methodology (RSM), which is essentially a collection of statistical and regression techniques. However, nonstatistical techniques such as ANN and genetic algorithms reported in this manuscript are found to outperform RSM in modeling and optimization of certain processes. However, in recent years, only a limited number of researchers have investigated the possibility of using such nonstatistical techniques in biological processes [10]. Moreover, optimization using such artificial intelligence techniques for enhancing lipase production has not been studied, which could be helpful even when a statistical based optimization technique fail.

In this paper, we present the results of optimization of media constituents, previously known to influence lipase production by a soil microorganism, using ANN-GA method.

Methodology

Background About the Previous Results on Lipase Production by the Soil Microorganism

In our previous study [11], microorganism isolated from a soil sample contaminated with vegetable cooking oil was used in lipase production. It was reported that the media constituents—oil, magnesium sulfate and ferrous sulfate among others were screened to be the most significant factors affecting lipase production by the soil microorganism in batch shake flask; the three factors were then optimized for maximum lipase production using a statistical technique-response surface methodology (RSM). The results on lipase production, expressed in terms of lipolytic activities at 72 h of culture, were obtained by performing a 2^3 central composite design of experiments with six center point replicates.

Artificial Neural Network Modeling

Using the previously obtained results, a neural network model based on back propagation algorithm was developed for predicting the lipolytic activities of the culture. Two third of the data were used for training the network and the remaining data were used for validating (testing) the model developed [5]. For developing the model, the training and validation data sets were randomly separated, and presented to the input layer neurons of the network. Once the network was adequately trained, the test data were used to evaluate performance of the model. The optimal values of network architecture thus selected for modeling lipolytic activity of the culture were obtained by trial and error method. These optimized values are presented in Table 1.

Optimization using Genetic Algorithms

ANN model developed for the data on lipolytic activities of the culture was used for optimization of media components employing the GA approach. This was used mainly to enhance the lipase production by the culture and also to validate this method of optimization over a statistical regression-based optimization technique-response surface methodology (RSM) in bioprocesses.

Table 1 Optimal values of network parameters used in ANN modelling of lipolytic activities of the culture.

Parameters	Values
Training matrix	15
Test matrix	5
No. of hidden layers	4
Error tolerance	0.001
Theta	0.5
Learning rate	0.75

Input variables: Oil; $MgSO_4$; $FeSO_4$

Table 2 GA parameters set for optimization of lipolytic activities of the culture.

Population size	Total no. of generations	Cross over probability	Mutation probability	Total string length	No. of binary coded variables	Total no. of runs
100	250	0.9	0.05	45	3	5

The GA parameters chosen for this study are given in Tables 2 and 3. A chromosome made of three different genes with each gene representing a different medium constituent, namely, oil, magnesium sulfate and ferrous sulfate was constituted. An initial random population of 100 chromosomes were taken for the study and passed on to the next generation. Strings for mating were selected by a probability proportional to their fitness value. A new generation of medium composition was obtained by carrying out a single point crossover between parent chromosomes selected for mating; generally, a crossover operation is mainly responsible for searching new strings. To avoid any convergence to local maxima, point mutations were used, which help in introducing diversity in the population and therefore force the algorithm to search for entire space. The crossover and mutation probabilities were assigned to be 0.9 and 0.05, respectively [3]. The offspring generated were evaluated of their fitness using the ANN model developed.

Softwares Used

ANN modeling using back propagation method was performed using the software Neurosolutions, version 1.4, USA. A modified version of the C-implementation of a simple binary coded variable GA developed at Kanpur Genetic Algorithms Laboratory, Indian Institute of Technology Kanpur, India was used for GA optimization.

Results and Discussions

ANN Modeling of the Data on Lipolytic Activities of the Culture

Regression models have been successfully applied to biological systems for optimization of media constituents to enhance the yield of *Bacillus thuringiensis* PBT-372 [12]. Studies on optimization of media components and growth conditions by response surface methodology to enhance lipase production [1, 11] have also used regression models for predicting responses in the system. Using the experimental results of central composite design, regression model equation for the lipolytic activities of the culture was obtained (Eq. 2).

$$Y = -0.22X_1^2 - 0.36X_2^2 - 0.68X_3^2 - 0.13X_1X_2 + 0.03X_1X_3 + 0.21X_2X_3 + 1.29X_1 - 0.16X_2 - 0.76X_3 + 4.65 \tag{2}$$

Table 3 Lower and upper bounds of individual chromosomes used in GA.

Variable	Lower bound	Upper bound	String length
Oil	0.100 ml/l	10.000 ml/l	15
$MgSO_4$	0.100 g/l	0.900 g/l	15
$FeSO_4$	0.001 g/l	0.020 g/l	15

Table 4 ANN model developed for lipolytic activities of the culture.

	Oil	$MgSO_4$	$FeSO_4$	Theta	Lipolytic activity
Input to hidden layer weights					
Hidden 1	+5.06	+6.57	−8.18	−0.42	+2.77
Hidden 2	−0.73	−1.17	−6.49	+3.58	−7.58
Hidden 3	+2.19	+1.90	−6.57	−1.04	+6.99
Hidden 4	−4.73	−6.34	+7.08	+2.43	+2.32
Theta					+0.008
Input to outer layer weights	−0.35	−5.03	−2.19		

The terms in the equation depict a second order polynomial relationship between the lipolytic activity (Y) and the media constituents: oil (X_1), magnesium sulfate (X_2) and ferrous sulfate (X_3). The model was slightly accurate with a coefficient of determination (R^2) value of 0.97.

Using the optimal values of the ANN parameters, obtained by trial and error method, an ANN model was developed to predict the lipolytic activities of the culture. The model was trained and tested adequately with the experimental data, and evaluated by the R^2 value obtained, which was found to be about 0.99. Valdez-Castro *et al.* reported that ANN based model was able to predict the fed-batch fermentation kinetics of *Bacillus thuringiensis* [3] based on similar ANN modeling procedure. The ANN model for lipolytic activities of the culture is presented in Table 4. A comparative analysis of measured versus predicted lipolytic activity values due to ANN and regression models is shown in Fig. 2. The figure shows that the ANN model is highly successful in predicting the experimental values. The measured and predicted values (due to the regression and ANN models) of lipolytic activities are presented in Table 5. The results of lipolytic activities predicted by the regression and ANN models are also illustrated in Fig. 3 through radar plots. Radar plots appear more realistic as they indicate a spatial distribution of a less number of data points. In other words, they visually depict scattered data points better than conventional line plots. Clearly, the lipolytic activity values predicted by the ANN model showed better correlation with the experimental values than the regression model. These results clearly establish the advantage of an artificial-intelligence-based technique over a simple regression-based statistical technique in predicting the complex behavior of a biological system. Similar observation was made by Nagata and Chu [2] for optimization of a fermentation medium while comparing statistical regression and ANN modeling of their data.

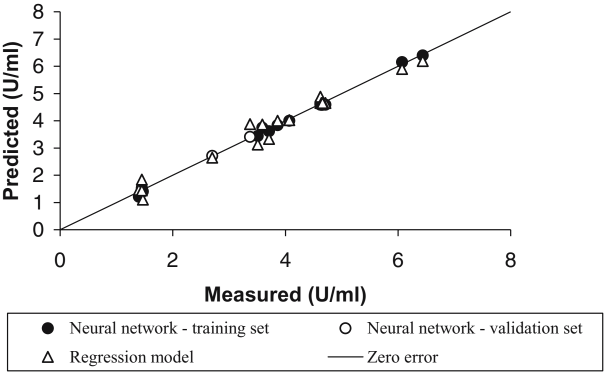


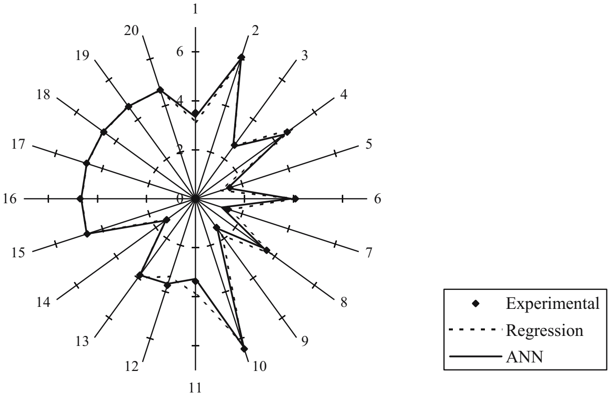
Fig. 2 Comparison between measured and predicted lipolytic activities of the culture due to ANN and regression models

Table 5 2^3 Central composite design showing experimental, regression model and ANN model predicted lipolytic activities of the culture.

Media constituents			Lipolytic activity values (U/ml)		
Oil (ml/l)	$MgSO_4$ (g/l)	$FeSO_4$ (g/l)	Experimental	Predicted	
				Regression	ANN
1.5	0.3	0.005	3.51	3.13	3.33
5.5	0.3	0.005	6.07	5.90	6.02
1.5	0.7	0.005	2.70	2.64	2.62
5.5	0.7	0.005	4.62	4.88	4.46
1.5	0.3	0.015	1.47	1.10	1.39
5.5	0.3	0.015	4.07	4.02	3.89
1.5	0.7	0.015	1.40	1.46	1.13
5.5	0.7	0.015	3.59	3.86	3.42
0.13	0.5	0.01	1.45	1.84	1.50
6.86	0.5	0.01	6.44	6.19	6.37
3.5	0.16	0.01	3.37	3.88	3.27
3.5	0.83	0.01	3.71	3.33	3.65
3.5	0.5	0.0016	3.86	4.00	3.86
3.5	0.5	0.018	1.45	1.44	1.50
3.5	0.5	0.01	4.64	4.65	4.66
3.5	0.5	0.01	4.71	4.65	4.66
3.5	0.5	0.01	4.67	4.65	4.66
3.5	0.5	0.01	4.62	4.65	4.66
3.5	0.5	0.01	4.64	4.65	4.66
3.5	0.5	0.01	4.66	4.65	4.66

The difference between the measured and ANN-predicted values of lipolytic activities was also calculated in the form of residuals for the entire experimental data set and is presented in Fig. 4. Generally, a positive or negative residual value of a parameter in a residuals plot reveals that model prediction of observed value is, respectively, high or low. In the present study, very low negative and positive residual lipolytic activity values (Fig. 4) show that the ANN model has predicted the experimental values very accurately.

Fig. 3 Radar plot for comparing measured and predicted lipolytic activities of the culture due to ANN and regression models. (Axis units—U/ml)



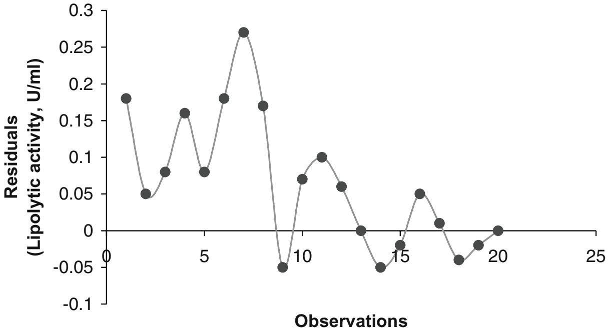


Fig. 4 Residual plot of measured and predicted lipolytic activities of the culture using the ANN model

Optimization by GA Technique

The ANN model developed for lipolytic activities of the culture was used for optimization of the media constituents using the GA technique. A total of five different runs employing the GA parameters were used to search for the entire space of objective functions given by the ANN model; the population profile of GA at various generations is shown in Fig. 5. In general terms, a population profile from one generation to the next generation in optimization using GA indicates a convergence to one or more level. The number of individuals with the same level of fitness function is directly correlated with its fit to the objective function in preceding generations. A string is usually selected for the mating pool with a probability proportional to its fitness. The generation ‘0’ initially contains a random population covering the whole parameter range. To get a better visualization of the whole system, the parameter range was divided into eight identical subunits for lipolytic activity, each of which summarizes the number of individuals in the respective concentration levels. As seen from the figure, irrespective of the type of initial population all chromosomes seem to converge to a single fitness level lying between 7 and 8 U/ml of lipolytic activity. Such a convergence to one concentration level usually indicates existence of a possible global maximum to the objective function. Weuster-Botz and Wandrey [13] have also reported similar types of convergence profile after four generations by using GA for media optimization in production of formate dehydrogenase.

Performance profile of the GA over generations for maximizing lipolytic activity is shown in Fig. 6. It could be seen from the figure that average fitness value of the objective function increases over the number of generations due to an increase in the number of healthy individuals. However, in later generations, the average fitness value became almost stagnant as the convergence to the maxima was achieved. It is also evident from the figure that all the runs converge to a single optimal value in each case. Bapat and Wangniker [14] found that average fitness value increased slightly with generations in their study involving optimization of Rifamycin B fermentation conditions using GA. A summary of results obtained in the present study by optimization using the ANN-GA method is presented in Table 6, and the optimized values of the media constituents were found to be: 9.991 ml/l oil, 0.100 g/l $MgSO_4$ and 0.009 g/l $FeSO_4$. These optimized values of oil and $MgSO_4$ were in the respective upper and lower levels of the ranges selected in the study; for $FeSO_4$, the value was found to be in middle of its chosen range. The optimized value of oil also revealed the significant role played by the sole carbon source in the media for enhancing maximum lipolytic activity of the culture obtained from experiments, an observation which was consistent with our previous study employing RSM technique [11]. A maximum value

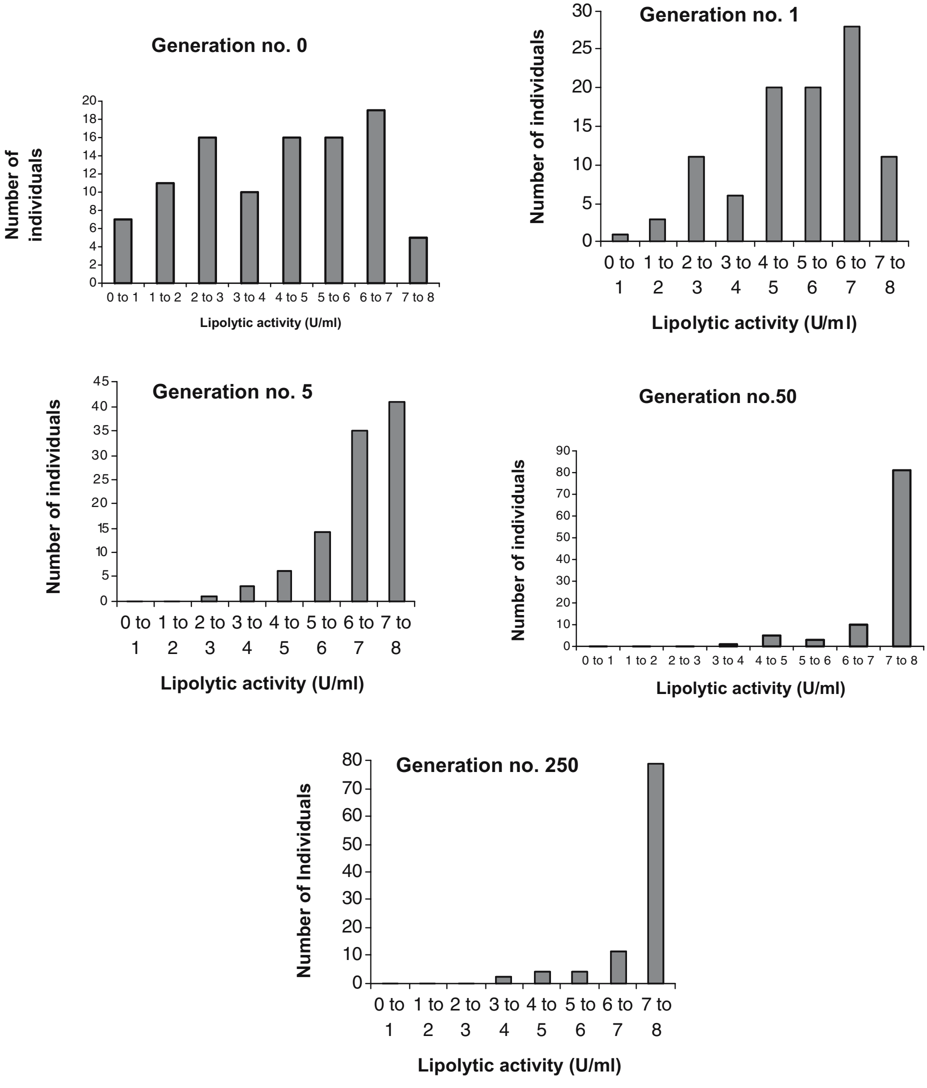


Fig. 5 Population profile at different generations showing convergence of GA for maximum lipolytic activity. The initial population contains chromosomes distributed in all levels of the fitness function

of 7.69 U/ml of lipolytic activity of the culture was obtained at the optimized levels of the media constituents. This maximum value was found to be 19.4% higher than the maximum lipolytic activity values observed in the experimental design (Table 5). In a study by Lakshmi et al. [15], the yeast *Candida rugosa* showed maximum lipolytic activity of 2.856 U/ml utilizing sunflower oil as the substrate. Compared to this value, the maximum lipolytic activity obtained in the present study is found to be considerably higher at optimized values of the media constituents.

There is no clear rule in the use of GA about the interactions between media constituents. It is generally observed that when different media constituents converge to a single optimal value, then there exist no significant interactions between them. In this study,

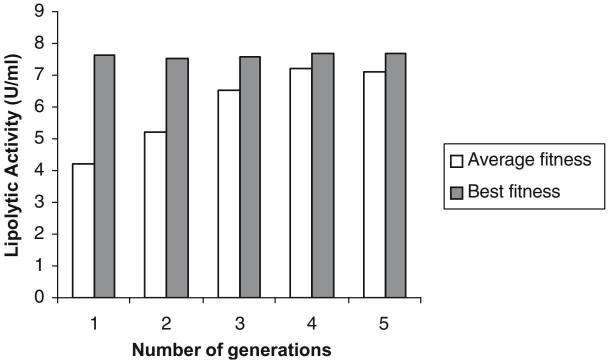


Fig. 6 Performance of ANN-GA-based optimization method for enhancing lipolytic activity

the optimal values of the media constituents, obtained using the ANN-GA method, also converge to a single concentration level, and hence, it could be surmised that there may exist least significant interaction between any two of the three variables: oil, magnesium sulfate and ferrous sulfate. Weuster-Botz and Wandrey [13] predicted similar noninteracting media constituents due to their convergence to a narrow concentration range.

The lipolytic activity value obtained by optimization of the media constituents using ANN-GA method was compared with that obtained by using RSM. For optimization using RSM, the quadratic regression model equation for predicting the lipolytic activities were partially differentiated and then equated to zero. Corresponding maxima was further checked by second order sufficient condition using Hessian matrix. The optimized value of media constituents thus obtained were 9.300 ml/l oil, 0.311 g/l $MgSO_4$ and 0.007 g/l FeSO4; using these optimal values of the media constituents, maximum predicted lipolytic activity was found to be 7.002 U/ml [11]. The value (7.69 U/ml) obtained using ANN-GA method was found to be 8.8% higher than the maximum RSM predicted value. From the results presented in the study, it could be well said that optimization of media constituents using ANN-GA method would be highly applicable for enhancing lipase production by other microorganisms, as well.

Conclusion

The data on lipolytic activities of a soil microbial culture obtained using a 2^3 central composite design were adopted for optimization of media constituents. Back propagation-ANN modeling and GA was successfully employed in this optimization study. The ANN model was found to be highly predictive of the system compared to a simple quadratic regression model. Using the ANN-GA method, a maximum lipolytic activity value of

Table 6 Summary of results obtained for maximum lipolytic activity of the culture using ANN-GA method.

Results	Lipolytic activity (U/ml)					
	Run1	Run2	Run3	Run4	Run5	Avg
Maximum	7.69	7.70	7.69	7.69	7.69	7.69
Minimum	3.36	4.20	3.62	4.49	2.08	3.78
Average	7.09	7.17	7.13	7.42	7.31	7.22

7.69 U/ml of culture were obtained. This value was 8.8% higher than the maximum output values obtained by optimizing the media constituents using response surface methodology indicating superior performance of ANN-GA method in optimizing the media constituents for enhancing lipase production by the soil microorganism

References

1. Kaushik, R., Saran, S., Isar, J., & Saxena, R. K. (2006). *Journal of Molecular Catalysis. B, 40,* 121–126.
2. Nagata, Y., & Chu, K. H. (2003). *Biotechnology Letters, 25,* 1836–1842.
3. Valdez-Castro, L., Baruch, I. S., & Barrera-Cortes, J. (2003). *Bioprocess and Biosystems Engineering, 25,* 229–233.
4. Baughman, D. R., & Liu, Y. A. (1995). in: *Neural Networks in Bioprocessing and Chemical Engineering*. San Diego: Academic.
5. Goh, A. T. C. (1995). *Artificial Intelligence in Engineering, 9,* 143–151.
6. Hornik, K., Stinchcombe, M. & White, H. (1989). *Neural Networks, 2,* 359–366.
7. Poggio, T. & Girosi, F. (1990). *Proceedings of IEEE, 78*(9), 1481–1497.
8. Maier, H. R., & Dandy, G. C. (1998). *Environmental Modelling & Software, 13,* 193–209.
9. Deb, K. (1995). *Optimization for Engineering Design: Algorithms and Examples*, Prentice Hall of India Private Limited.
10. Chen, L. Z., Nguang, S. K., Chen, X. D., & Li, X. M. (2004). *Biochemical Engineering Journal, 22,* 51–61.
11. Haider, M. A., & Pakshirajan, K. *Applied Biochemistry and Biotechnology, 141,* (in press)
12. Prabagaran, S. R., Pakshirajan, K., Swaminathan, T., & Jayachandran, S. (2004). *Chemical and Biochemical Engineering Quarterly, 18,* 183–187.
13. Weuster-Botz, D., & Wandrey, C. (1995). *Process Biochemistry, 30,* 563–571.
14. Bapat, P. M., & Wangikar, P. P. (2004). *Biotechnology and Bioengineering, 86,* 201–208.
15. Lakshmi, B. S., Kangueane, P., Abraham, B., & Pennathur, G. (1999). *Letters in Applied Microbiology, 29,* 66–70.