J Ind Microbiol Biotechnol (2009) 36:747–756
DOI 10.1007/s10295-009-0547-6

ORIGINAL PAPER

Media optimization for biosurfactant production by *Rhodococcus erythropolis* MTCC 2794: artificial intelligence versus a statistical approach

Moumita P. Pal · Bhalchandra K. Vaidya · Kiran M. Desai · Renuka M. Joshi · Sanjay N. Nene · Bhaskar D. Kulkarni

Received: 23 September 2008 / Accepted: 9 February 2009 / Published online: 13 March 2009

Abstract This paper entails a comprehensive study on production of a biosurfactant from *Rhodococcus erythropolis* MTCC 2794. Two optimization techniques—(1) artificial neural network (ANN) coupled with genetic algorithm (GA) and (2) response surface methodology (RSM)—were used for media optimization in order to enhance the biosurfactant yield by *Rhodococcus erythropolis* MTCC 2794. ANN and RSM models were developed, incorporating the quantity of four medium components (sucrose, yeast extract, meat peptone, and toluene) as independent input variables and biosurfactant yield [calculated in terms of percent emulsification index (% EI_{24})] as output variable. ANN-GA and RSM were compared for their predictive and generalization ability using a separate data set of 16 experiments, for which the average quadratic errors were ~3 and ~6%, respectively. ANN-GA was found to be more accurate and consistent in predicting optimized conditions and maximum yield than RSM. For the ANN-GA model, the values of correlation coefficient and average quadratic error were ~0.99 and ~3%, respectively. It was also shown that ANN-based models could be used accurately for sensitivity analysis. ANN-GA-optimized media gave about a 3.5-fold enhancement in biosurfactant yield.

Keywords Biosurfactant · Media optimization · Artificial neural network · Genetic algorithm · Response surface methodology · Rhodococcus

M. P. Pal · B. K. Vaidya · R. M. Joshi · S. N. Nene (✉) · B. D. Kulkarni
Chemical Engineering and Process Development Division, National Chemical Laboratory, Pune 411008, India
e-mail: sn.nene@ncl.res.in

B. K. Vaidya
e-mail: bk.vaidya@rediffmail.com

K. M. Desai
Food Engineering and Technology Department, Institute of Chemical Engineering, Matunga, Mumbai 400019, India

Introduction

Biosurfactants are surface-active metabolites produced extracellularly or as part of the cell membrane by a wide variety of microorganisms such as bacteria, yeasts, and fungi. Biosurfactants have been identified for several industrial applications particularly in cosmetic, pharmaceutical, and food processes as emulsifiers, humectants, preservatives, and detergents. Because of their structural diversity (i.e., glycolipids, lipopeptides, fatty acid esters), low toxicity, and biodegradability, biosurfactants have potential for replacing synthetic surfactants. Moreover, they are ecologically safe and can be applied in bioremediation and waste treatments [1].

Biosurfactants produced by a few *Rhodococcus* species have been reported to be more effective and efficient in reduction of surface and interfacial tensions than many synthetic surfactants [2]. Chemically, *Rhodococcus* biosurfactants are trehalose lipids. Although their commercial potential has been recognized in recent years, like other biosurfactants, of rhodococcal biosurfactants have yet to penetrate in the market [3]. The major obstacles to their commercialization are low fermentative yield and high production cost.

Formation of most microbial products is a complex, highly nonlinear process. Along with other process variables, the media components play a key role in controlling

yield and specific productivity. Thus in the fermentation process, media optimization is recognized as a simple but effective method for achieving high productivity of desired products. The limitations of the classical method (one factor at a time) are discussed in detail in many earlier reports [4, 5]. To overcome their inability to determine interactive effects between input variables and to predict the "true" optimum, two alternative approaches are commonly employed for media optimization: (1) artifical intelligence (AI)-based techniques and (2) statistics-based techniques.

In the past few decades, AI-based techniques—specifically artificial neural network (ANN) coupled with genetic algorithm (GA)—offer an attractive choice for nonlinear modeling and optimization. The ANN-based models have several advantages over different statistical methods. ANN-based models can be constructed solely from the historic process input–output data without any detailed knowledge of the process phenomenology. ANN has been successfully implemented in modeling a large number of biochemical processes such as pattern recognition [6], classification [7], process control [8], soft sensor applications [9], and reaction kinetics modeling [10, 11], including modeling of fermentation yield [12]. Though the obvious requirement for ANN is a large number of data points, being black-box models they don't reveal system information in a subtle way. However these drawbacks can be easily overcome, which makes ANN an attractive option for handling data with a wide range of nonlinearity. Moreover, the capability of ANN models to perform sensitivity analysis further expands their applicability [13]. The ANN-GA hybrid combinations are widely used in nonlinear optimization problems. GA belongs to class of evolutionary algorithm. It is a stochastic method that can be easily applied to nonlinear and noncontinuous functions, and it only requires a zeroth-order derivative [14]. GA has been successfully used to solve complex nonlinear problems in highly diverse fields [15–20], including media optimization [21, 22].

Among various statistical methods, response surface methodology (RSM) is the most widely used method in media optimization. The ability to search for an optimum condition from a relatively small number of experiments and the ability to interpret the interactive effects among input variables are some attractive features of RSM [23]. One drawback of RSM is that it is mainly restricted to quadratic nonlinear correlation, whereas biological process may show more complex nonlinear dependencies.

ANN and RSM are compared in a few earlier reports [10, 24–26], and in almost all cases ANN was found to perform better than RSM. But these reports mainly focused only on the predictive capability of models. Besides predictive capabilities, optimization and sensitivity analysis are essential criteria required to make a comprehensive comparison between ANN and RSM.

The present paper deals with comparison between two optimization techniques—ANN-GA and RSM—that were used to enhance the yield of *Rhodococcus* biosurfactant by media optimization. ANN-GA and RSM were compared for their predictive and generalization ability. Moreover the accuracy of ANN-GA and RSM predictions were estimated using a sensitivity analysis method.

Materials and methods

Materials

All media components were purchased from Hi-Media, India. All other chemicals were of analytical grade and procured from S.D. Fine Chemicals, India.

Microorganisms and growth conditions

Six microbial strains, *Rhodococcus erythropolis* MTCC 1526, 1548, 2794, and 3951 and *Rhodococcus* spp. MTCC 2678 and 2683, were purchased from MTCC-Chandigarh (India).

All *Rhodococcus* strains were maintained on nutrient agar slants for 48 h at 30°C. Pre-inoculum (5 ml) was prepared in tubes from the slants and incubated for 24 h at 30°C on a rotary shaker at 200 rpm. This was transferred to 45 ml of the growth medium in 250-ml erlenmeyer flasks and incubated under identical conditions. The liquid fermentation medium used for batch culture experiments (termed medium A) contained the following (g/l): glucose (10), yeast extract (3), meat peptone (7.5), Na_2HPO_4 (4.0), KH_2PO_4 (2.0), $MgSO_4{\cdot}7H_2O$ (0.2), $CaCl_2{\cdot}2H_2O$ (0.02), ammonium ferric citrate (0.05), and trace mineral solution (1 ml/l). The composition of trace mineral medium contained the following (g/l): H_3BO_3 (0.1), $MnCl_2{\cdot}4H_2O$ (0.1), $ZnSO_4{\cdot}H_2O$ (0.1), $FeCl_3{\cdot}6H_2O$ (0.1), $CaCl_2{\cdot}2H_2O$ (1), $CuCl_2{\cdot}2H_2O$ (0.05).

Selection of optimum nitrogen source, carbon source, and inducer

The effect of different nitrogen sources was studied by replacing the organic nitrogen sources (yeast extract and meat peptone) from medium A with different inorganic nitrogen sources (urea, ammonium sulphate, and ammonium phosphate) at equivalent nitrogen levels.

To evaluate the optimum carbon source, glucose was replaced by an equivalent amount of different carbon sources, namely sucrose, sorbitol, mannitol, and glycerol.

Seven inducers (3% v/v each) were screened to evaluate the corresponding enhancement in biosurfactant production. Biosurfactant production was calculated in terms of emulsification index (% EI_{24}) as described Below. Among seven different inducers, toluene was found to give maximum % EI_{24}. Hence toluene was selected for further experiments.

Media optimization using ANN-GA

ANN was used for obtaining the functional relationship between media component and % EI_{24}. The popular architecture multilayer perceptron (MLP) with sigmoidal function was used. The data set was divided into training set (80%) and test set (20%) to avoid over-parameterization. The input data were scaled within proper range to avoid any numerical overflow. The output parameter was scaled between 0 and 1, as output is produced by a sigmoidal transfer function. A fully connected feed forward neural network (FANN) architecture in which data always flow in a forward direction, i.e. from input layer to output layer, was used. A real number quantity, known as a weight, is associated with the connection of two neurons, which is analogous to a synapse in a brain neuron. The output of ANN was calculated as a function of input and weights using summation and transfer function. Weights are the adjustable parameters of the network.

An error back propagation (EBP) algorithm, which is a generalized form of least mean square (LSM) convergence, is used for adjusting the weights. It uses a gradient descent approach, in which weights are changed in proportion to the negative of the error gradient. The details of training an optimal MLP model possessing good prediction and generalization abilities are described in [12]. The EBP training algorithm makes use of two more adjustable parameters, the learning rate ($\eta\pm$) ($0 < \eta \leq 1$) and momentum coefficient (α) ($0 < \alpha \leq 1$). The magnitudes of both these parameters are optimized heuristically along with the number of hidden layer neurons. The average quadratic error (AQE) was chosen as performance index. The training iterations are stopped when the test AQE reaches a minimum, even though the training set AQE may continue to decrease with continuation of training. Initially a network with zero neurons in the hidden layer is used for training. The number of neurons in the hidden layer is increased subsequently and the AQE generated for networks with varying initialization, learning rate, and momentum coefficient are calculated. The topology, which gives a minimum AQE in all the above-mentioned heuristic training cycles, is chosen for final training. This optimum network is trained using the above procedure. The weights obtained after training are retained as model parameters.

After developing an ANN-based process model with good prediction accuracy and generalization ability, its input space can be optimized by using a genetic algorithm with a view of maximizing the yield. The objective function is to find a decision variable, i.e., ANN input vector (**x**), such that it maximizes the objective function, i.e., ANN output. The GA-based search for an optimal solution vector, x*, begins with a randomly initialized population of probable (candidate) solutions. The candidates are referred to as strings or chromosomes. Each chromosome is evaluated to measure its fitness using the ANN-based model. The steps involved in GA-based optimization algorithm are (1) selection: choosing fitter parent chromosomes to create a mating pool and (2) crossover: the production of offspring solutions, i.e., next generation solution by using genetic operators such as pair-wise crossing-over between fitter parent chromosomes and mutating of the offspring strings. This procedure creates a new population of chromosomes, which is then compared with the current pool of chromosomes. The best chromosomes evolve after repeating the above procedure until the termination criteria are met. The termination criteria could be a fixed number of generations or when the improvement in the fitness value of the subsequent generation is negligible.

An initial population of 40 chromosomes was generated randomly. Each chromosome was made up of a distinct media composition consisting of four different genes. Each gene represents a concentration of different medium components. The % EI_{24} value at the end of the batch was chosen as the fitness function. The ANN model built earlier from historic data was used to evaluate the fitness of each chromosome. After computing the fitness function, combinations producing high % EI_{24} were acted upon by the following genetic operators: selection, cross-over, and mutation. The roulette wheel scheme was used to determine a string with higher chance of surviving in subsequent generations. Selected chromosomes were used as parent chromosomes for single-point cross-over. Mutation was used to avoid premature termination due to entrapment local minima. Nevertheless, this parameter was used sparingly with a probability of 1% as compared to cross-over probability of 90%. Thus, the offspring of the next generation are generated. The fitness of the offspring was computed as output of the ANN model (% EI_{24}). This procedure was carried out for 100 generations to get the optimum solution.

Media optimization using RSM

The experimental design used for ANN was also used as input data for developing the RSM model. To examine the combined effect of four different medium components

(independent variables), a central composite factorial design of $2^4 = 16$ plus 6 *center points* plus 8 (i.e., 2×4) *star points* leading to a total of 30 experiments was performed in duplicate. The value of the dependent response (% EI_{24}) was the mean of two replications. The second-order polynomial coefficients were calculated and analyzed using a trial version of Design Expert software (version 6.0, Stat-Ease, USA). Statistical analysis of the model was performed to evaluate the analysis of variance (ANOVA).

Extraction of biosurfactant

Biosurfactant was extracted using methyl *ter*-butyl ether (MTBE) as described earlier by Kuyukina et al. [27]. To one volume of whole-cell broth, two volumes of MTBE were added and the extraction carried out for 3 h, at 28°C at 200 rpm. The top phase containing biosurfactant was collected using a separating funnel and dried on a rotary evaporator (Rotavap) to obtain a powder of crude biosurfactant.

Analytical methods

The emulsification index (% EI_{24}) provides a rapid and reliable measure of the quantity of biosurfactant. % EI_{24} was determined as described by Nitschke and Pastore [28]. The *Rhodococcus* cells were isolated by centrifugation at 10,000 rpm for 10 min at 10°C. The cell residue was suspended in distilled water (cell concentration was adjusted to 0.2 g/mL) and sonicated for 10 min to release biosurfactant from the cell wall. Then, 2 ml of sonicated sample was mixed with 3 ml of *n*-dodecane (hydrocarbon) in a 20-ml graduated, stoppered glass bottle. This was vortexed for 10 min and kept at room temperature. The percent ratio of height of emulsified zone to total height after 24 h gives % EI_{24} as in Eq. 1.

$$\%\ EI_{24} = \frac{\text{Height of emulsified zone}}{\text{Total height of liquid (sum of aqueous, oil and emulsified zones)}} \times 100 \quad (1)$$

Type of emulsion was determined using methyl orange (water soluble dye) and Sudan red III (lipid soluble dye) as described by Tian et al. [29]. Total carbohydrate and protein content of crude biosurfactant were estimated by phenol sulphuric acid method [30] and Folin Lowry assay [31], respectively. Surface tension measurement and critical micelle concentration were detected using a Kruss K-11 tensiometer (accuracy ± 0.1 mN/m).

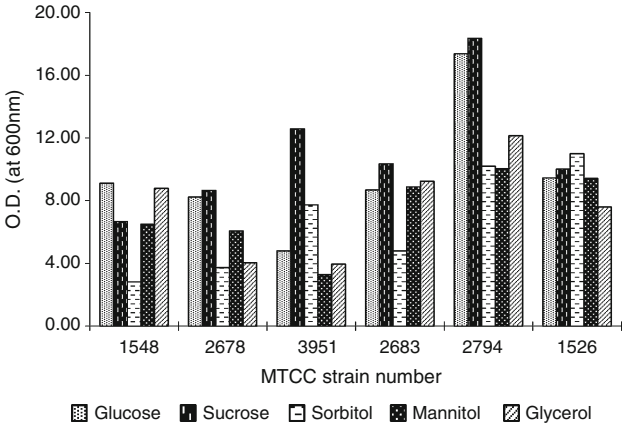

Fig. 1 Selection of *Rhodococcus* strain based on carbon source

Results and discussion

The *Rhodococcus* biosurfactant is a glycolipid that contains trehalose as the major carbohydrate along with (unsaturated and saturated) fatty acids and fatty alcohols. The glycolipid biosynthesis is predominantly cell-growth associated [32, 33]. Therefore initial attempts were made to increase the cell mass based on a one-factor-at-a-time strategy.

Selection of optimum nitrogen source, carbon source, and inducer

For all *Rhodococcus* strains, the complete substitution of organic by inorganic nitrogen resulted in very low cell mass (results not shown), therefore organic nitrogen sources (yeast extract and meat peptone) were selected for further studies.

Glucose was replaced by equivalent amounts of different carbon sources, namely sucrose, sorbitol, mannitol, and glycerol. The carbon source was found to affect the cell mass to a great extent. As the biosurfactant is cell-wall associated, high cell density is desirable [34]. The effect of carbon source on cell growth for six *Rhodococcus* strains is given in Fig. 1. The optimum carbon source was found to differ depending upon the *Rhodococcus* strain. Among different *Rhodococcus erythropolis*, MTCC 2794 gave maximum cell mass when grown on sucrose as carbon source. This combination was selected for further studies.

Hydrocarbons added to the fermentation medium are known to induce the production of biosurfactant [35]. Seven

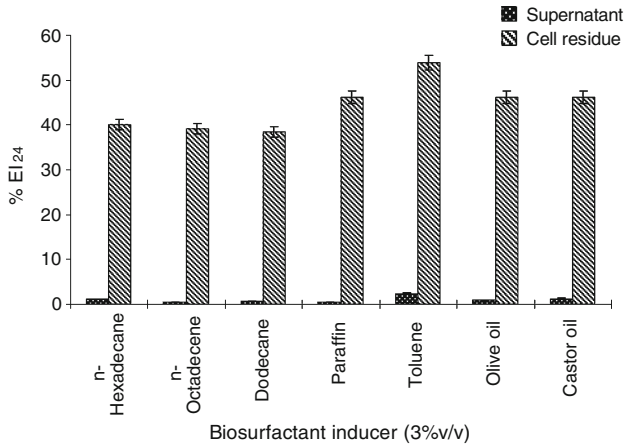

Fig. 2 Screening of biosurfactant inducer

such hydrocarbon inducers were screened for enhanced biosurfactant production by MTCC 2794. The effect of inducer on biosurfactant production is represented in Fig. 2. The % EI_{24} values of cell-free supernatant were very low (in the range of 0–4%) as compared with sonicated cell suspension. This indicated that the major portion of biosurfactant remained adhered to the cell surface. The % EI_{24} values of sonicated cell suspension and cell-free supernatant for different inducers are given in Fig. 2. Toluene gave maximum % EI_{24} (53.84%) and was therefore selected for further experiments.

ANN-GA-based modeling and optimization

The design of experiments (DoE) used as input data for developing an ANN based model is given in Table 1. The

Table 1 Experimental design and corresponding experimental and model-predicted values of % EI_{24}

No.	Sucrose	Yeast extract	Meat peptone	Toluene	% EI_{24}		
					Experimental values[a]	ANN-predicted	RSM-predicted
1	1.15	0.25	0.65	2.75	51.82	53.61	51.73
2	2.45	0.25	0.65	2.75	48.15	48.36	48.73
3	1.15	0.55	0.65	2.75	53.86	52.34	51.26
4	2.45	0.55	0.65	2.75	43.92	44.02	47.34
5	1.15	0.25	1.55	2.75	59.79	58.81	58.33
6	2.45	0.25	1.55	2.75	56.53	57.99	60.22
7	1.15	0.55	1.55	2.75	51.23	53.68	56.46
8	2.45	0.55	1.55	2.75	59.07	57.78	57.43
9	1.15	0.25	0.65	6.25	51.31	52.46	52.62
10	2.45	0.25	0.65	6.25	49.37	48.64	46.58
11	1.15	0.55	0.65	6.25	55.90	55.95	53.13
12	2.45	0.55	0.65	6.25	47.14	46.98	46.16
13	1.15	0.25	1.55	6.25	55.58	54.60	54.07
14	2.45	0.25	1.55	6.25	50.65	52.57	52.92
15	1.15	0.55	1.55	6.25	54.09	52.87	53.18
16	2.45	0.55	1.55	6.25	50.10	50.02	51.11
17	0.50	0.40	1.10	4.50	50.98	51.22	54.80
18	3.10	0.40	1.10	4.50	52.15	51.79	49.73
19	1.80	0.10	1.10	4.50	53.76	53.66	52.41
20	1.80	0.70	1.10	4.50	49.38	51.32	50.13
21	1.80	0.40	0.20	4.50	46.46	46.38	48.78
22	1.80	0.40	2.00	4.50	59.66	62.13	60.32
23	1.80	0.40	1.10	1.00	57.60	56.53	54.95
24	1.80	0.40	1.10	8.00	47.45	49.37	49.51
25	1.80	0.40	1.10	4.50	52.16	52.06	51.54
26	1.80	0.40	1.10	4.50	51.76	52.06	51.54
27	1.80	0.40	1.10	4.50	51.11	52.06	51.54
28	1.80	0.40	1.10	4.50	52.69	52.06	51.54
29	1.80	0.40	1.10	4.50	51.45	52.06	51.54
30	1.80	0.40	1.10	4.50	50.49	52.06	51.54

[a] Values indicate mean of duplicate observations

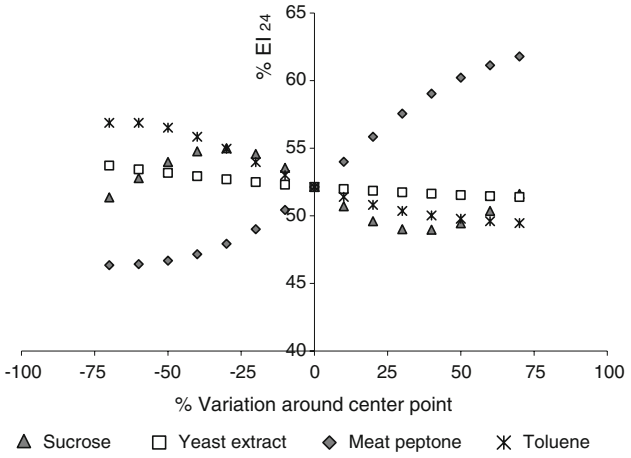

Fig. 3 Sensitivity analysis of ANN-based model

six center-point experiments were considered as a single data point with output as an average of outputs for these six replica experiments. Thus, the total data set of 25 points was divided into a training set of 20 and a test set of 5 data points. The star data points of DoE were kept in the training set as these points were the only data points for extreme values of input variables. The output, i.e. % EI_{24}, was used for obtaining the functional relationship between media component and biosurfactant yield. The momentum and learning rate were set to 0.1 and 0.8, respectively. The number of optimum hidden nodes was determined to be five. Thus, the final topological structure of the ANNs was 4-5-1. The correlation coefficients between predicted and experimental data for the training set and test set were 0.949 and 0.988, respectively, and average percent errors for training and test sets were 1.14 and 4.01, respectively. The overall correlation coefficient and average percent error were 0.96 and 1.79, respectively. The maximum error was ~4.79%. The small and comparable magnitudes of the average prediction error (%) and the high and comparable values of the correlation coefficient for both the training and test set outputs suggest that the MLP-based model possesses good approximation and generalization characteristics.

Even though ANN is a black-box model, useful information about the system can be obtained using simple sensitivity analysis. The center point of DoE data was used as the reference point. The data set was generated by changing the concentration of each component in steps on both sides of the reference point, keeping concentrations of all other components at the center composition. Figure 3 shows the simulated values of % EI_{24} for this data set using the ANN model. Each series represents the effect of each variable on the fermentation yield. The effect of each variable can be gauged from the extent of variation in response and also from the slope of each series. It can be seen that meat peptone has the most significant effect on the yield followed by sucrose and toluene. Yeast extract showed the least effect on biosurfactant yield. The positive slope suggests that biosurfactant yield is higher at the higher concentrations of meat peptone, whereas the negative slope of toluene implies higher yield at its lower concentration. The effect of sucrose on % EI_{24} was found to be highly nonlinear. The comparison of ANN sensitivity results with RSM are discussed below.

Since GA does not guarantee global optimum explicitly, it was necessary to search the entire input space rigorously. This was done by repeating the GA-based optimization procedure several times for different randomly initialized populations of chromosomes and for different GA-specific parameters. Almost all the varied initial conditions converged to similar solutions, suggesting it to be the global solution. The optimum solution with experimental verification is given in Table 2. The average percent error between predicted and experimental response for optimum conditions was less than 2% in all the cases.

RSM-based modeling and optimization

A second-order polynomial equation was used to correlate the independent process variables with biosurfactant production. The second-order polynomial coefficient for each term of the equation was determined through multiple regression analysis using the Design Expert. The experimental design used for ANN was also used as input data for developing the RSM model (Table 1).

The results were analyzed by using ANOVA. The results are shown in Table 3. The model *F*-value of 3.65 implies the model is significant. There is only a 0.90% chance that a model *F*-value this large could occur due to noise. Noise, which is responsible for most of the variability in the response, arises due to parameters that are hard and expensive to control in process settings (environmental conditions such as temperature and humidity,

Table 2 ANN-GA-based optimum solutions and experimental verification

No.	Sucrose	Yeast extract	Meat peptone	Toluene	Experimental % EI_{24}	ANN-GA-predicted % EI_{24}
1	1.95	0.10	1.99	4.30	65.20	63.95
2	2.02	0.30	2.00	3.57	63.60	62.92
3	1.98	0.27	1.99	2.33	63.76	62.57

Table 3 ANOVA of RSM model

No.	Model terms	Values
1	Standard deviation	2.69
2	Coefficient of variation (CV)	5.14
3	R^2	0.772
4	Adj. R^2	0.560
5	Adequate precision	7.43
6	Model F-value	3.65

variations in raw material, accuracy limits of instruments, etc.), and it varies randomly within the process. The process parameters other than the chosen independent variables are also a source of noise. The model error can be attributed to model lack of fit and experimental noise. The experimental noise can be estimated from replication experiments. This further confirms the significance of the model.

After regression analysis, the second-order response model was obtained as shown in Eq. 2:

$$\begin{aligned}\%\,\mathrm{EI}_{24} = &+ 51.54095 - 1.26788\mathrm{A} - 0.56988\mathrm{B} \\ &+ 2.88487\mathrm{C} - 1.35871\mathrm{D} + 0.18189\mathrm{A}^2 \\ &- 0.066751\mathrm{B}^2 + 0.75165\mathrm{C}^2 + 0.17197\mathrm{D}^2 \\ &- 0.23066\mathrm{AB} + 1.22192\mathrm{AC} - 0.76146\mathrm{AD} \\ &- 0.34898\mathrm{BC} + 0.24352\mathrm{BD} - 1.28694\mathrm{CD}\end{aligned} \quad (2)$$

where A, B, C, and D represent sucrose, yeast extract, meat peptone, and toluene, respectively.

The P values were used as a tool to determine the significance of each of the coefficients, which, in turn, are necessary to understand the pattern of mutual interactions between the test variables. The smaller the magnitude of P, the more significant is the corresponding coefficient. Values of P less than 0.05 indicate model terms that are significant. The coefficient and the corresponding P values suggest that, among the input variables, A (sucrose), C (meat peptone), and D (toluene) are significant model terms. The lack-of-fit F-value of 19.66 implies the lack of fit is significant. There is only a 0.21% chance that a lack-of-fit F-value this large could occur due to noise. The RSM-based optimum solutions and experimental verification are given in Table 4.

Comparison of ANN-GA and RSM in biosurfactant optimization

While RSM is the most widely used method in fermentation media optimization, ANN-GA has rapidly developed into one of the most efficient methods for modeling and optimization, especially for nonlinear systems. This section presents the comparison between predictive capabilities of ANN and RSM for two data sets: (1) the experimental data that are used for developing the model (DoE data) and (2) the separate unseen (validation) data. The generalization capability of the model can be verified by its prediction accuracy for a validation set. Table 5 shows the results for 16 experiments randomly performed to form a validation data set.

The predictive and generalization capability of the RSM and ANN models was compared on the basis of correlation coefficient, average percent error, and maximum error given by the model. The comparative results are shown in Fig. 4 and in Table 6. It can be observed from Table 6 that both the models performed reasonably well, but ANN performed consistently better than RSM. In the case of the DoE data, the correlation coefficients for the ANN and RSM models were 0.96 and 0.83, respectively. The average percent error and maximum error for ANN were less than half those for RSM. In the case of the validation data set, the correlation coefficients for the ANN and RSM models were 0.90 and 0.70, respectively. The average percent errors for the ANN and RSM models were 2.79 and 6.11, respectively, and the maximum errors for ANN and RSM models were 7.64 and 15.08, respectively, for the validation data set. The prediction performance of the ANN model for the validation data set confirms its superior generalization capacity for the given case.

The main limitation of RSM is that is assumes only a quadratic form of nonlinear correlation. So if we want to use RSM effectively, we need to narrow down the search window appropriately. (If we make the search window narrow enough, linear correlation may also suffice.) This makes the search process highly dependent upon the search

Table 4 RSM-based optimum solutions and experimental verification

No.	Sucrose	Yeast extract	Meat peptone	Toluene	% EI_{24}		
					Experimental values	ANN-predicted (% error in prediction)	RSM-predicted (% error in prediction)
1	2.42	0.38	2.02	2.19	61.26	60.54 (1.17)	67.01 (9.38)
2	2.25	0.27	2.03	2.82	60.78	62.87 (3.44)	66.30 (9.08)
3	1.61	0.20	1.98	1.53	57.59	59.24 (2.49)	63.22 (9.77)

Table 5 Validation data set with experimental as well as model-predicted % EI_{24}

No.	Sucrose	Yeast extract	Meat peptone	Toluene	% EI_{24}		
					Experimental values	ANN-predicted	RSM-predicted
1	1.00	0.20	0.50	1.00	53.14	52.46	49.96
2	1.00	0.20	0.25	1.00	48.50	48.88	49.13
3	0.50	0.10	1.00	1.00	53.55	53.81	52.27
4	2.00	0.10	1.00	1.00	53.00	52.57	51.70
5	0.50	0.20	0.50	1.00	51.92	51.76	50.55
6	0.50	0.10	2.00	1.00	58.33	54.79	56.69
8	0.50	0.40	0.25	1.00	52.00	50.50	49.76
9	2.00	0.40	1.00	1.00	51.44	51.22	51.35
10	0.50	0.60	2.00	1.00	52.17	48.75	56.10
11	1.00	0.40	0.5	3.00	50.00	52.98	45.84
12	2.00	0.10	0.25	1.00	45.26	45.85	47.99
13	0.50	0.10	0.25	8.00	51.85	52.67	46.51
14	2.00	0.40	0.25	1.00	41.66	42.55	47.72
15	0.50	0.40	1.00	1.00	49.35	47.58	52.02
16	2.00	0.20	0.50	3.00	43.24	46.54	43.46

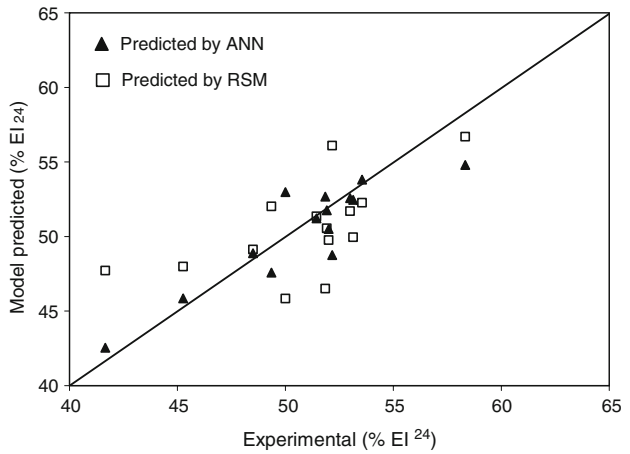

Fig. 4 Comparative parity plot

space. It will require either extra experiments or good a priori knowledge of the system to fix the search window. Since ANN can inherently capture almost any form of nonlinearity, it can easily overcome the limitation of RSM discussed above. Thus, in case of ANN, we can choose a more liberal search space, even if the correlation (in that search space) is more complex than quadratic.

In the present study, lack of fit and low correlation R^2 in RSM can be attributed to noncompliance of quadratic correlation between dependent and independent process variables in the given search space. This problem could be tackled by narrowing the search space or shifting it in the direction of steepest descent. But since ANN has captured the nonlinear correlation successfully (shown by low AQE and high R^2), there is no need for further amendment in RSM.

Table 6 Statistical comparison of predictive capability of ANN and RSM models

Parameter	DoE data		Validation data	
	ANN	RSM	ANN	RSM
Correlation coefficient	0.96	0.83	0.90	0.70
Average percent error	1.76	3.97	2.79	6.11
Maximum error	4.79	10.22	7.64	15.08

The RSM- and ANN-GA-predicted optimized compositions were comparable. However, RSM over-predicted the maximum yield. The average percent errors in maximum predicted yield by ANN-GA and RSM were approximately 2.5 and 9.5%, respectively. This difference in the optimum prediction can be attributed to the higher average and maximum percent error of RSM. The ANN has accurately predicted the yield for RSM-optimized conditions (see Table 4). This again exemplifies the superior generalization capacity of ANN and the accurate prediction of optimum by ANN-GA.

The sensitivity analysis, i.e., an effect of each variable on the system (as described in ANN section) shows that inferences derived from ANN and RSM are comparable. Both methods showed that the meat peptone is the most significant media component followed by sucrose and toluene. There are also methods available in literature to quantify the interactive effect of variable pairs on the

system using ANN models. But that study is beyond the scope of this report.

MTBE extraction of biosurfactant

MTBE was used as a method of extraction for recovery of biosurfactant from *Rhodococcus* spp. 2794 grown on optimum medium predicted by ANN model (given in Table 2, optimum solution no. 1). The yield of crude biosurfactant was expressed in grams per liter of fermentation broth. The yields of crude biosurfactant before and after optimization were 2.05 and 7.2 g/l, respectively. Thus, a significant increase (3.5-fold) in the yield of biosurfactant was achieved by ANN-GA optimization.

Characterization of crude biosurfactant

The type of emulsion can be determined by observing the mixture of emulsion and dye under a binocular light microscope. The two dyes used were methyl orange (water soluble) and Sudan red III (lipid soluble). The emulsion/methyl orange dye mixture, when observed under the microscope, appeared as colourless droplets on an orange background. This indicates that it is an o/w emulsion, as the water-soluble dye gave colour to an external aqueous phase, while oil droplets remained colourless. The o/w nature of the emulsion was confirmed using Sudan red III where droplets appeared red on a colourless background.

The crude biosurfactant was analyzed for the carbohydrate and protein contents. It was found to contain 22.62% protein and 50.13% total carbohydrates. The crude biosurfactant decreased the surface tension of water from 72 to 33.8 mN/m (at 120 mg/l) and achieved a CMC value of 100 mg/l.

Conclusions

The production of biosurfactant by *Rhodococcus* spp. MTCC 2794 was observed to be growth associated. It was observed that use of an organic nitrogen source gave higher cell mass than with inorganic nitrogen. Among six *Rhodococcus* strains selected for the work, *R. erythropolis* MTCC 2794 was found to give maximum cell mass when grown in a medium containing sucrose as carbon source. Among seven different inducers studied for MTCC 2794, toluene gave the best results.

Two optimization techniques, ANN-GA and RSM, were applied for media optimization in order to enhance the biosurfactant yield by *Rhodococcus erythropolis* MTCC 2794. A three-step systematic optimization approach comprised of (1) screening, (2) ANN-based modeling, and (3) GA-based optimization was reported for the first time to maximize the fermentative production of biosurfactant. Moreover, the present manuscript describes a comparative assessment between artificial intelligence (AI)-based and statistical methodologies in media optimization. Most such comparative studies have focused only on the predictive abilities of ANN and RSM. This study, however, compares optimization abilities and sensitivity analysis in addition to comparison between predictive capabilities of ANN-GA and RSM optimization methods. Thus the present manuscript gives a broader and more generalized comparison that can be used in designing media-optimization strategies.

ANN has better generalization capacity, and ANN-GA is more accurate in predicting the optimum than RSM. It was also demonstrated that ANN models could be used effectively for sensitivity analysis. In this particular case, ANN-GA proved to be consistently better than RSM.

Acknowledgements The authors would like to thank Dr. S. S. Bhagwat (ICT, Mumbai) for his help in measurement of surface tension and CMC. BKV would like to acknowledge the SRF grant from CSIR-India. RMJ would like to thank DST-India for the financial support.

References

1. Makkar RS, Cameotra SS (2002) An update on the use of unconventional substrates for biosurfactant production and their new applications. Appl Microbiol Biotechnol 58:428–434. doi:10.1007/s00253-001-0924-1
2. Bell KS, Philp JC, Aw DW, Christofi N (1998) The genus Rhodococcus. J Appl Microbiol 85:195–210. doi:10.1046/j.1365-2672.1998.00525.x
3. Mukherjee S, Das P, Sen R (2007) Towards commercial production of microbial surfactants. Trends Biotechnol 24:509–515. doi:10.1016/j.tibtech.2006.09.005
4. Lotfy WA, Ghanem KM, El-Helow ER (2007) Citric acid production by a novel *Aspergillus niger* isolate. II. Optimization of process parameters through statistical experimental designs. Bioresour Technol 98:3470–3477. doi:10.1016/j.biortech.2006.11.032
5. Tanyildizi MS, Özer D, Elibol M (2005) Optimization of a-amylase production by *Bacillus* sp. using response surface methodology. Process Biochem 40:2291–2296. doi:10.1016/j.procbio.2004.06.018
6. Jin H, Zheng Z, Gao M, Duan Z, Shi Z, Wang Z, Jin J (2007) Effective induction of phytase in *Pichia pastoris* fed-batch culture using an ANN pattern recognition model-based on-line adaptive control strategy. Biochem Eng J 37:26–33. doi:10.1016/j.bej.2007.03.009
7. Partovi FY, Anandarajan M (2002) Classifying inventory using an artificial neural network approach. Comput Ind Eng 41:389–404. doi:10.1016/S0360-8352(01)00064-X
8. Nagy ZK (2007) Model based control of a yeast fermentation bioreactor using optimally designed artificial neural networks. Chem Eng J 127:95–109. doi:10.1016/j.cej.2006.10.015
9. Desai KM, Badhe YP, Tambe SS, Kulkarni BD (2006) Soft-sensor development for fed-batch bioreactors using support vector regression. Biochem Eng J 27:225–239. doi:10.1016/j.bej.2005.08.002

10. Baş D, Boyaci IH (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. Baş D, Dudak FC, Boyaci IH (2007) Modeling and optimization III: reaction rate estimation using artificial neural network (ANN) without a kinetic model. J Food Eng 79:622–628. doi:10.1016/j.jfoodeng.2006.02.021
12. Desai KM, Vaidya BK, Singhal RS, Bhagwat SS (2005) Use of an artificial neural network in modeling yeast biomass and yield of β-glucan. Process Biochem 40:1617–1626. doi:10.1016/j.procbio.2004.06.015
13. Jaiswal S, Benson ER, Bernard JC, Van Wicklen GL (2005) Neural network modelling and sensitivity analysis of a mechanical poultry catching system. Biosyst Eng 92:59–68. doi:10.1016/j.biosystemseng.2005.05.007
14. Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading
15. Chen CR, Ramaswamy HS (2002) Modeling and optimization of variable retort temperature (VRT) thermal processing using coupled neural networks and genetic algorithms. J Food Eng 53:209–220. doi:10.1016/S0260-8774(01)00159-5
16. Cheng J (2007) Hybrid genetic algorithms for structural reliability analysis. Comput Struct 85:1524–1533. doi:10.1016/j.compstruc.2007.01.018
17. Nasseri M, Asghari K, Abedini MJ (2008) Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network. Expert Syst Appl 35:1415–1421. doi:10.1016/j.eswa.2007.08.033
18. Hassan MR, Nath B, Kirley M (2007) A fusion model of HMM, ANN and GA for stock market forecasting. Expert Syst Appl 33:171–180. doi:10.1016/j.eswa.2006.04.007
19. Kim HT, Kim BY, Park EH, Kim JW, Hwang EW, Han SK, Cho S (2005) Computerized recognition of Alzheimer disease-EEG using genetic algorithms and neural network. Future Gener Comp Syst 21:1124–1130. doi:10.1016/j.future.2004.03.012
20. Elalfi AE, Haque R, Elalami ME (2004) Extracting rules from trained neural network using GA for managing E-business. Appl Soft Comput J 4:65–77. doi:10.1016/j.asoc.2003.08.004
21. 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
22. Bapat PM, Wangikar PP (2004) Optimization of rifamycin B fermentation in shake flasks via a machine-learning-based approach. Biotechnol Bioeng 86:201–208. doi:10.1002/bit.20056
23. Box G, Hunter W, Hunter J (1958) Statistics for experiments. Wiley, New York
24. Huang J, Mei L-H, Xia J (2007) 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
25. Dutta JR, Dutta PK, Banerjee R (2004) Optimization of culture parameters for extracellular protease production from a 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
26. He L, Xu Y-Q, Zhang X-H (2008) Medium factors optimization and fermentation kinetics for phenazine-1-carboxylic acid production by *Pseudomonas* sp. M18G. Biotechnol Bioeng 100:250–259. doi:10.1002/bit.21767
27. Kuyukina MS, Ivshina I, Philip J, Christofi N, Dunbar S, Ritchkova M (2001) Recovery of *Rhodococcus* biosurfactants using methyl tertiary-butyl ether extraction. J Microbiol 46:149–156
28. Nitschke N, Pastore GM (2006) Production and properties of a surfactant obtained from *Bacillus subtilis* grown on cassava wastewater. Bioresour Technol 97:336–341. doi:10.1016/j.biortech.2005.02.044
29. Tian B, Dong C, Chen L (1998) Preparation of Konjac glucomannan ester of palmitic acid and its emulsification. J Appl Polym Sci 67:1035–1038. doi:10.1002/(SICI)1097-4628(19980207)67:6<1035::AID-APP10>3.0.CO;2-#
30. Dubois M, Gilles K, Hamilton J, Rebers P, Smith F (1956) Colorimetric method for determination of sugars and related substances. Anal Chem 28:350–356. doi:10.1021/ac60111a017
31. Lowry OH, Roseburgh N, Farr AL, Randall R (1951) Protein measurement with the Folin Phenol reagent. J Biol Chem 193:265–275
32. Ghurye GL, Vipulanandan C (1994) A practical approach to biosurfactant production using nonaseptic fermentation of mixed cultures. Biotechnol Bioeng 44:661–666. doi:10.1002/bit.260440514
33. Rodrigues L, Teixeira J, Oliveira R, van der Mei HC (2006) Response surface optimization of the medium components for the production of biosurfactants by probiotic bacteria. Process Biochem 41:1–10. doi:10.1016/j.procbio.2005.01.030
34. Bicca FC, Fleck LC, Ayub MAZ (1999) Production of biosurfactant by hydrocarbon degrading *Rhodococcus ruber* and *Rhodococcus erythropolis*. Rev Microbiol 30:231–236. doi:10.1590/S0001-37141999000300008
35. Bento FM, Gaylarde CC (1996) The production of interfacial emulsions by bacterial isolates from diesel. Int Biodeterior Biodegrad 38:31–33. doi:10.1016/S0964-8305(96)00021-2