CrossMark

ORIGINAL ARTICLE

Multiple linear regression and artificial neural networks for delta-endotoxin and protease yields modelling of *Bacillus thuringiensis*

Karim Ennouri[1] · Rayda Ben Ayed[1] · Mohamed Ali Triki[2] · Ennio Ottaviani[3] ·
Maura Mazzarello[3] · Fathi Hertelli[1] · Nabil Zouari[4]

Abstract The aim of the present work was to develop a model that supplies accurate predictions of the yields of delta-endotoxins and proteases produced by *B. thuringiensis* var. *kurstaki* HD-1. Using available medium ingredients as variables, a mathematical method, based on Plackett–Burman design (PB), was employed to analyze and compare data generated by the Bootstrap method and processed by multiple linear regressions (MLR) and artificial neural networks (ANN) including multilayer perceptron (MLP) and radial basis function (RBF) models. The predictive ability of these models was evaluated by comparison of output data through the determination of coefficient (R^2) and mean square error (MSE) values. The results demonstrate that the prediction of the yields of delta-endotoxin and protease was more accurate by ANN technique (87 and 89% for delta-endotoxin and protease determination coefficients, respectively) when compared with MLR method (73.1 and 77.2% for delta-endotoxin and protease determination coefficients, respectively), suggesting that the proposed ANNs, especially MLP, is a suitable new approach for determining yields of bacterial

products that allow us to make more appropriate predictions in a shorter time and with less engineering effort.

Keywords Delta-endotoxins · Proteases · *Bacillus thuringiensis* · Bootstrap method · Multiple linear regression · Artificial neural networks

Introduction

Bacillus thuringiensis is a Gram-positive bacterium used as a biopesticide in agriculture. Compared to chemical pesticides, its benefits are a particular toxicity against specific insects and safety to non-target organisms. The insecticidal characteristics of *B. thuringiensis* are due to entomopathogenic proteins produced through sporulation and proteolytic enzymes considered as primary metabolites.

Bacillus genus bacteria are considered as the most entomopathogenic microorganisms used as biocontrollers. Thanks to spore-forming and toxin secretion, *Bacillus* is appropriate to scale-up production and field use. Toxins are proteins synthesized during the sporulation stage (Charles et al. 1997) and are efficient against insects such as lepidoprans, coleopterans and dipterans (Van Frankenhuyzen 2013). Thus, the *Bacillus thuringiensis* var. *kurstaki* HD-1 strain is successfully commercialized as an agrochemical pesticide (Navon 2000).

Insecticidal activity of *Bacillus thuringiensis* strains is based on the production of a crystalline parasporal inclusion (named delta-endotoxin) during sporulation. The protoxin proteins, also called Cry proteins, are classified according to their toxicity level and the types of insects which are specifically sensitive to each of them. Moreover, suitable selection of medium ingredients and culture conditions are decisive to the commercial production success achieved through high toxic activity per fermentation broth volume. After exhaustion of one or more

Senior author: Nabil Zouari.

✉ Karim Ennouri
1karimennouri1@gmail.com

[1] Centre of Biotechnology of Sfax, Sfax University, 3038 Sfax, Tunisia

[2] Laboratoire Ressources et Amélioration Génétiques de l'Olivier, du Pistachier et de l'Amandier, Institut de l'Olivier, Sfax 3000, Tunisia

[3] On Air s.r.l., via Carlo Barabino 26/4B, Genoa, Italy

[4] Biological and Environmental Sciences Department, College of Arts and Sciences, Qatar University, Doha, Qatar

Springer

nutrients, *B. thuringiensis* produces spores and parasporal crystal proteins (also called delta-endotoxins) exhibiting insecticidal activity with high specificity to insect larvae. The harvested mixture of spores and crystals after the completion of sporulation are formulated as bioinsecticides, and quality is related to its potency and stability (Ghribi et al. 2007). The most important components of culture medium are carbon and nitrogen sources, as well as trace elements. Added to the production of insecticidal proteins, *B. thuringiensis* has also been considered as an excellent producer of protease enzymes. In fact, *B. thuringiensis* var. *israelensis*, *B. thuringiensis* var. *kurstaki* and *B. thuringiensis* var. *tenebrionis* are the major proficient strains of *B. thuringiensis* in producing proteases (Hotha and Banik 1997; Tyagi et al. 2002). Microbial proteases catalyze the hydrolysis of proteins (Haq et al. 2006). There are two main protease types: intracellular and extracellular. Extracellular proteases hydrolyze proteins in the medium and allow the assimilation of the degraded products by the cell, whereas, intracellular proteases are required to support cellular and metabolic pathways, such as sporulation and cell differentiation, protein turnover and also the transformation of protoxins into mature endotoxins (Hartley 1960). In general, *Bacillus* proteases are extracellular and could be accumulated in the fermentation medium. The production of extracellular proteases by microorganisms is influenced by medium components, essentially carbon and nitrogen sources (McKellar and Cholette 1984), and metal ions (Adinarayana et al. 2003). In addition, some studies had shown that *B. thuringiensis* bioinsecticides production is achieved by important levels of enzymes in the growth media (Zouari and Jaoua 1999; Chen et al. 2004). The production of enzymes by microorganisms is induced by medium components (Singh et al. 2004). Proteolytic activities are necessary to assimilate proteins as nitrogen source, but they are involved in the stability of proteins produced by the micro-organism as metabolites. Some microorganisms produce low amounts of these enzymes, thus limiting their biotechnological application. Nevertheless, by rather simple manipulations, the use of a definite and optimized medium could enhance the yield productions. The protease production in *Bacillus* species are affected by various complex mechanisms especially during the transition state between exponential growth, when the production is difficult (Frankena et al. 1985), and the stationary phase (Strauch and Hoch 1993). In our previous work, we showed that a significant relationship exists between the accumulation of delta-endotoxins of *B. thuringiensis* and the available proteolytic activities in the medium (Ennouri et al. 2013c).

The selection of an appropriate fermentation medium to develop an industrial fermentation is a decisive assignment since its product concentration and total yield are significantly influenced by the fermentation medium composition. This design can be arduous, costly, and often time consuming, involving many trials (Chen et al. 2010). To specify the optimal conditions for product formation in fermentation process optimization, it is crucial to evaluate the medium component concentrations and to choose the best process condition combination (Stanbury et al. 1997). In our study, it is essential to establish a balanced system in which the synthesis and accumulation of delta-endotoxins in the fermentation broth is not counteracted by proteolytic activities of *B. thuringiensis*. It is still complicated to predict simultaneously yields of several bioproducts using a given set of cultivation factors.

In this regard, several regression models could be built from yield analyses, and then be used for yield predictions. Multivariate linear regression (MLR) techniques are suitable statistical tools to approximate complex relationships of the prediction variables (Baffi et al. 1999). MLR is a statistical method used to investigate the relationship between one response variable (dependent variable) with two or more variables (independent variables). If the relation between the dependent and other independent variables could be found using multiple regressions, a better control approach could be sought. Usually, the modelling used in bioprocesses is based on equation equilibrium combined with substrate consumption and product formations. However, these methods have been proven to be inefficient to reveal the nonlinear relationships between influenced variables and responses (Lee and Park 1999). An alternative approach based on artificial neural networks (ANN) has been utilized in modelling industrial fermentation processes (Khaouane et al. 2013). Such approaches are able to solve specific problems through learning, by typical inputs and corresponding desired responses, unlike usually employed methods that consist of the construction of an algorithm and its execution as a computer program (Tadeusiewicz 1993).

The aim of the present work was to identify the most important medium ingredients for the improvement of delta-endotoxin and protease yields produced by *B. thuringiensis* var. *kurstaki* HD-1. Moreover, mathematical models are constructed by means of two different artificial intelligence techniques, namely MLR and ANN. The obtained predictions of delta-endotoxin and protease yields are compared in terms of accuracy to decide upon the most efficient technique.

Materials and methods

Microorganisms

Microbial strain HD-1, identified as *B. thuringiensis* var. *kurstaki* strain (ATCC39756), was provided kindly by

Dr. Daniel R. Zeigler (*Bacillus* Genetic Stock Center, Columbus, OH, USA). The strain was maintained by streak inoculation Luria Broth (LB) nutrient plates containing (g l^{-1}): yeast extract 5, peptone 10, NaCl 5 and agar 15. Plates were incubated at 30 °C for 24 h and stored at 4 °C for future use.

Inocula preparation method

Inocula preparation had previously been optimised based on delta-endotoxin production yields (Ghribi et al. 2004). One isolated colony was dispensed in 3 ml of LB medium and incubated overnight at 30 °C. Aliquots of 0.5 ml were used to inoculate 250-ml shake flasks containing 50 ml of LB medium. After 6 h of incubation at 30 °C in a rotary shaker set (New Brunswick Scientific^{TM}, Edison, NJ, USA) at 200 rpm, the O.D.600 was estimated using a Smart-Spec^{TM} 3000 UV–visible spectrophotometer (Bio-Rad Laboratories, Hercules, CA, USA). The obtained broth should correspond to mid-log period culture (Ghribi et al. 2004). 1 O.D.600 Unit in *B. thuringiensis* inoculum has been estimated to approximately 2×10^8 CFU ml^{-1}. The culture broth was used to inoculate the studied culture to start with an initial O.D.600 of 0.15 corresponding to 3×10^7 CFU ml^{-1} and 0.05 g l^{-1} dry biomass, which is required to start growth of *B. thuringiensis* cultures for delta-endotoxin production (Zouari et al. 1998; Ghribi et al. 2004).

Cultural conditions

For delta-endotoxin production, trials were carried out on the basis of a complex medium (Ghribi et al. 2007). Runs are shown in Table 2 to provide information for the determination of the settings of variables in the experimental design. The pH was adjusted to 7.0 and CaCO$_3$ (20 g l^{-1}) was added to maintain pH stability. All flasks were sterilised at 121 °C for 20 min. The 250 ml flat-bottom flasks, containing 20 ml of culture medium, were incubated for 72 h at 30 °C in a rotary shaker at 200 rpm.

Delta-endotoxin determination

Crystal proteins were solubilised before protein concentration assay as described by Zouari et al. (1998). Crystal spore pellets were washed twice with 1 M NaCl and twice with bidistilled water. Then samples were incubated in 0.05 M NaOH (pH 12.5) for 2 h at 30 °C in a rotary shaker (200 rpm). The soluble fractions were collected by centrifugation at 13000 rpm, for 10 min. Protein concentration in the supernatant containing the alkali-soluble insecticidal crystal proteins was determined by the Bradford method (Bradford 1976) using bovine serum albumin (BSA) as a standard. The presented values are the averages (±SD) of three determinations of two independent experiments.

Protease assay

Proteolytic activity was estimated with casein as a substrate (Ennouri et al. 2013a). Casein was suspended in 0.1 M Tris–HCl buffer (pH 7.0) at a concentration of 1%. The assay mixture consisted of enzyme solution diluted with 0.1 M Tris–HCl buffer (pH 9.0). The reaction mixture was incubated at 60 °C for 20 min and the reaction was ended by the addition of 3 ml of 5% trichloroacetic acid (TCA), and then centrifuged at $5000 \times g$ for 10 min to eliminate the resulting precipitate. Protease activity was calculated as released tyrosine from the supernatants according to a modified Lowry method. One unit of enzyme activity was defined as the amount of the enzyme producing the release of 1 µg of tyrosine per min at 60 °C. The calculated values were the mean of three values of two independent experiments.

Determination of sporulation

Viable spores were counted as colony forming unit (CFU). Appropriately diluted samples were treated at 80 °C for 10 min then transferred to LB solid agar plates and incubated at 30 °C for 24 h. The colonies counted on plates were corrected for dilution factor and registered as spore viable count (Lachhab 2001; Ennouri et al. 2013b). Values are given as the mean of three values of two experiments.

Experimental designs and data analyses

Plackett–Burman design

The Plackett–Burman statistical experimental design is helpful for screening the most important factors from a number of parameters (Krishnan et al. 1998). This design is used to screen the important variables affecting the production of delta-endotoxins and proteolytic enzymes. The design matrix was established according to Montgomery (1997). Each variable is represented at two levels, a high level expressed by (+) and a low level expressed by (−). The high level of each variable is far enough from the low level so that a significant effect, if it exists, is probably going to be detected. The experimental design for the screening of medium components in our work is shown in Table 1. The horizontal rows in Table 1 represent the twelve different experiments and each column represents a different variable. For each experimental variable, high (+) and low (−) levels were evaluated. All experiments were performed in separate duplicates and the values of the

Table 1 Medium components (variables) and the respective high (+) and low (−) concentration levels used in Plackett–Burman design

Variable	Medium components	Lower level (−1) (g l^{-1})	Higher level (+1) (g l^{-1})
X_1	KH_2PO_4	0.5	1.5
X_2	K_2HPO_4	0.5	1.5
X_3	$MgSO_4$	0.1	0.5
X_4	$MnSO_4$	0	0.002
X_5	$FeSO_4$	0	0.002
X_6	Starch	25	35
X_7	Soybean meal	20	30

yields of delta-endotoxins and proteases are the averages of five measurements. The effect of each variable was determined with the following equation:

$$E_{(X_i)} = \left(\sum M_{i+} - M_{i-} \right) \Big/ N,$$

where $E_{(X_i)}$ is the concentration effect of the medium component variable, $\sum M_{i+} - M_{i-}$ are the sum of the obtained values at high and low levels, respectively, the yields of delta-endotoxins or proteases in trials, in which the variable (X_i) determined was present at the high or low level, respectively, and N is the number of experiments. Experimental error was calculated by estimating the variance among the dummy variables as follows:

$$V_{\text{eff}} = \sum (E_d)^2 / n,$$

where V_{eff} is the variance of the concentration effect, E_d is the concentration effect for the variable and n is the number of the shown variables. The standard error of the mean (S.E.M.) of the concentration effect is the square root of the concentration and the p value (significance level) of concentration effect was estimated by Student's t test:

$$t_{(X_i)} = E_{(X_i)} / \text{S.E.M.}$$

where $E_{(X_i)}$ is the effect of variable X_i (Pujari and Chandra 2000).

Bootstrap method

Experimental data produced using Plackett–Burman are small sample design. The application of ANN requires large data sets for the learning and testing procedures, and we, therefore, used the statistical method of bootstrap to generate random samples with replacement from a data set where each observation is selected separately at random from the original data set. The number of elements (components of the medium and the response) in each bootstrap sample equals the number of elements in the original data set. The procedure analyzes each sample the same way. To compare the results of the regression analysis to those of

the ANN, we used the same sample of data generated by the bootstrap method.

Multiple linear regression (MLR)

MLR techniques based on least-square procedures are usually used for estimating the variable effects involved in the model (Miller and Miller 2000). In this study, MLR was carried out on the training data set, using yields of delta-endotoxins or proteases as the response variable, and the seven medium components as predictor variables. In all cases, the response was expressed as a function of the seven nutrient concentrations. The success of the MLR can be measured by evaluating the magnitude of the determination coefficient R^2, the residual standard error (RSE) for the regression, and the results of the Student's t test results for the individual predictor variables.

Artificial neural networks (ANN)

ANN are a type of artificial intelligence based on the brain's neural operations (Pachepsky et al. 1996). A neural network is composed of simultaneously processing elements, i.e. neurons (Malinova and Guo 2004). The network consists of a set of input units, hidden units, and output units, which connect the inputs to the outputs. The selection of the input factors is the main aspect of neural network modelling (Song et al. 1995). The number of the hidden neurons depends on the character of the investigated problem. The determination coefficient (R^2) is used to control the accuracy of the predictive capacity of the constructed models. The training data set is applied to teach the ANN to find the global comprehensive model between its inputs and outputs. The test data are used to verify and confirm the predictive quality of the extended networks. The multilayer perceptron (MLP) and radial basis function (RBF) neural network architectures are possibly the most used ANNs (Nabney 2002). MLP and the RBF neural network structures have been employed in making predictions (Shafizadeh-Moghadem et al. 2015; Zare et al. 2013). Due to the non-linear efficiencies of these networks, they are considered good estimators providing very accurate results.

In this study, the MLP network hidden layer consists of non-linear activation functions: hyperbolic tangent and exponential functions. The expanded RBF network included a Gaussian activation function in the hidden layer and a linear activation function within the output layer. It has been argued that a network with a single hidden layer, with sufficient data, can be used to model any function (Beale and Jackson 1990). Therefore, the used MLP and RBF neural network structures consisted of only one hidden layer. Thus, designing the ANN implies the selection of a

Table 2 Plackett–Burman experimental design for the evaluation of factors affecting delta-endotoxin and protease yields of *Bacillus thuringiensis* HD-1

Run	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Delta-endotoxin ($\mu g\ ml^{-1}$)	Protease (IU ml^{-1})	CFU ($10^7\ ml^{-1}$)	Yield delta-endotoxins	Yield proteases
1	−	−	−	+	+	+	−	2458.8 ± 80	113.6 ± 9	78 ± 6	31.52	1.46
2	+	+	+	−	+	+	−	2032.7 ± 65	227.2 ± 8	70 ± 4	29.04	3.25
3	+	−	−	−	+	+	+	4139 ± 95	145.4 ± 5	70 ± 5	59.13	2.08
4	−	−	−	−	−	−	−	1898.7 ± 75	250 ± 8	128 ± 4	14.83	1.95
5	−	+	−	−	−	+	+	3697.3 ± 90	159.1 ± 7	120 ± 7	30.81	1.33
6	−	+	+	−	+	−	−	2096.7 ± 70	466 ± 10	91 ± 3	23.04	5.12
7	+	+	−	+	−	−	−	2341.5 ± 60	109.1 ± 8	84 ± 4	27.88	1.30
8	−	−	+	+	+	−	+	3925.1 ± 65	263.6 ± 6	78 ± 2	50.32	3.38
9	+	−	+	−	−	−	+	3775.3 ± 95	220.4 ± 9	77 ± 5	49.03	2.86
10	+	+	−	+	+	−	+	3484 ± 80	803 ± 11	127 ± 2	27.43	6.32
11	−	+	+	+	−	+	+	4800.5 ± 85	152.2 ± 8	86 ± 4	55.82	1.77
12	+	−	+	+	−	+	−	2790.6 ± 55	165.6 ± 9	93 ± 6	30.01	1.78

satisfactory number of hidden neurons and suitable network organizations in concordance with type and nature of inputs (discrete, continuous, categorical, quantitative, etc.). The number of hidden neurons was optimized by reducing an error function that mapped the number of hidden nodes to the accuracy of the expanded networks. The neural network module of STATISTICA 8.0 software was used in modelling the ANN. Data were categorized into the two parts: training (80%), and testing (20%). In the network, there were seven inputs and one output, corresponding to medium ingredients and cellular yield, respectively. The hidden neurons were optimized by building various MLP and RBF ANNs with hidden nodes of 7–70. Networks with hidden nodes greater than 70 were not developed due to the predictive capabilities decreasing as the number of intermediate units. The determination coefficient and mean square error (MSE) method were employed in selecting the best model.

The mean square error (MSE) and the determination coefficient (R^2) can be considered standard criteria for the estimation of statistical performance and used to control the precision of the predictive capacity of the constructed models. One hundred different combinations of activation functions and neuron numbers were tried by considering the fitted model MSE.

Results and discussion

Evaluation of the effect of cultivation conditions on delta-endotoxin yields by Plackett and Burman design

Fractional factorial design (FFD) is an experimental design that enables the user to evaluate the most important factors influencing a given process with the least number of trials. Plackett–Burman design is an FFD which was used in our case to study seven factors allowing a better understanding of the effect of each culture medium component. The evaluation of process variables was carried out according to the experimental matrix presented in Table 2, in which the delta-endotoxin yield was the measured response. Variation in delta-endotoxin yields are also shown in Table 2. The highest delta-endotoxin yield of 59.13 ($\mu g/10^7$ cells) was obtained in combination 3, and the lowest, of about 15 ($\mu g/10^7$ cells) in combination 4. Statistical analyses of these data revealed that the determination coefficient R^2, an indicator of the goodness of the model fitting, is close to 76%. In fact, this indicates that less than 25% of the total variations were not explained by the model, which thus delivers a good approximation. The R^2 describes the variance amount by calculating data explained by the model. Indeed, R^2 ranges from 0 to 1, with higher values signifying less error variance and values superior to 0.5 are considered suitable (Van Liew et al. 2003).

The analysis of regression coefficients and t value (Table 3) reveals that mainly soybean meal, starch, $MgSO_4$, $MnSO_4$, $FeSO_4$, and KH_2PO_4 were found to be the variables that improved delta-endotoxin production of *B. thuringiensis* HD-1. In fact, the presence of some ions such as Mg^{2+}, Fe^{2+} and Mn^{2+} in the culture of *B. thuringiensis* is fundamental (El-Bendary 2006). Potassium ion is essential for toxin production by *B. thuringiensis* (El-Bendary 1999). Moreover, starch and soybean meal are considered, respectively, as adequate carbon and nitrogen sources for increasing delta-endotoxin production (Avignone Rossa et al. 1990). Furthermore, soybean meal appeared to be the most significant factor for delta-endotoxin yield as indicated by a p value <0.05. Whereas,

Table 3 Estimated effects for delta-endotoxin and protease yields (coded units)

	Delta-endotoxin			Protease		
	Effect	t	p value	Effect	t	p value
Constant		10.73	0.000		7.54	0.002
KH_2PO_4	2.694	0.40	0.707	0.4302	0.60	0.583
K_2HPO_4	−6.804	−1.02	0.365	0.9288	1.29	0.267
$MgSO_4$	7.609	1.14	0.317	0.6213	0.86	0.437
$MnSO_4$	2.849	0.43	0.691	−0.0964	−0.13	0.900
$FeSO_4$	2.018	0.30	0.777	1.7687	2.45	0.070
Starch	7.299	1.10	0.335	−1.5465	−2.15	0.098
Soybean meal	19.371	2.91	0.044	0.4803	0.67	0.542

$R^2_{\text{Delta-endotoxin}} = 75.69\%$

$R^2_{\text{Protease}} = 77.59\%$

Fig. 1 Effects of different medium components for delta-endotoxin yield
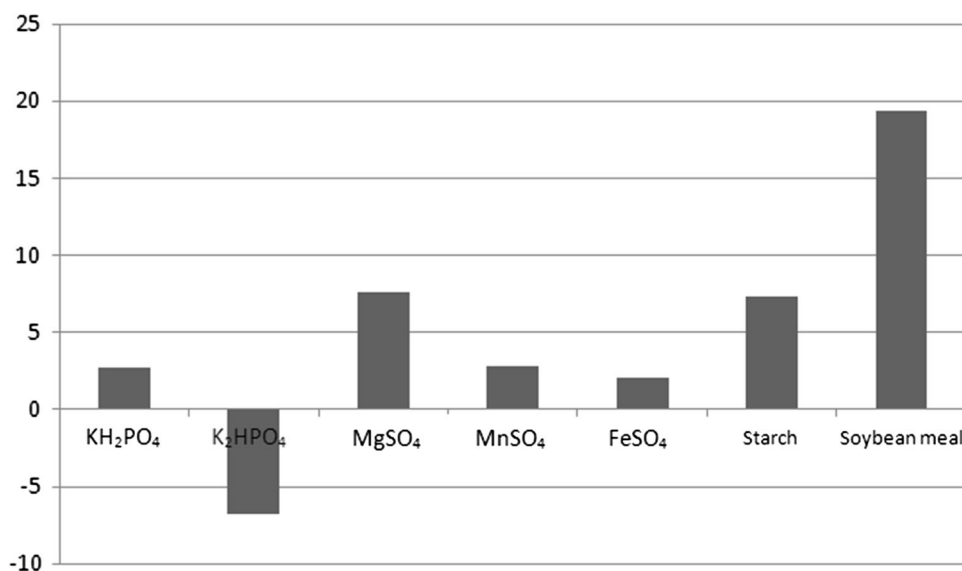
K_2HPO_4 was the lowest variable that hinders delta-endotoxin production. This finding is in line with results obtained by Ozkan et al. (2003) who found that an effective synthesis of Cry4Ba by *B. thuringiensis* var. *israelensis* HD-500 occurred only when inorganic phosphate was available in high concentrations. The main effects of the examined variables on delta-endotoxin yield are presented in Fig. 1.

Evaluation of the effect of cultivation conditions on protease yields by Plackett and Burman design

Generally, the production of bacterial extracellular enzymes during fermentation is influenced by nutrients such as carbon, nitrogen and minerals (Chao et al. 2010). In this study, seven medium ingredients were selected as variables and two concentrations of each variable were

assigned to 12 trials according to a Plackett–Burman design. The protease yield of each trial is summarized in Table 2. The highest yield of 6.32 (IU/10^7 cells) was obtained with combination 10, while the lower yield of 1.30 (IU/10^7 cells) was obtained with combination number 7. The R^2 value was close to 78%. The effects of variables and their confidence levels are shown in Table 3. Soybean meal, $FeSO_4$, KH_2PO_4, K_2HPO_4 and $MgSO_4$ affected positively the protease production. This effect was also observed for protease production by other *Bacillus* species in which an improved supply of nitrogen, such as soybean meal and phosphorus stimulated protease excretion (Moon and Parulekar 1991; Chu 2007). Furthermore, supplementation of the culture medium with Mg^{2+} and K^+ salts allowed higher protease production (Ellaiah et al. 2002). On the other hand, starch and $MnSO_4$ had a negative effect on proteolytic activity. In fact, the presence of carbon

source, such as starch, had a retroactive effect on enzyme production in some *Bacillus* species (Johnveshy and Naik 2001; Joo et al. 2002). Several studies also demonstrated that the presence of Mn^{2+} is necessary for enhanced protease activity of *Bacillus subtilis* PE-11 (Adinarayana et al. 2003). Indeed, present results were in disagreement with these latter findings and this may be explained by the possibly different nature of studied bacterial strain.

The main effects of medium component on proteolytic activity of *B. thuringiensis* var. *kurstaki* HD-1 are presented in Fig. 2. This plot is useful in determining the protease production at intermediate levels of different combinations of independent variables. The statistical design of experiments offers an efficient methodology to select the decisive variable and to optimize the factors with minimum experiment numbers for response estimation. The most important advantage of the Plackett–Burman design was the possibility to rank the different variable effects on the measured response irrespective of its nature or trend.

Because the dataset in our study was relatively small, it was important to perform a bootstrap test for the entire available data set to obtain an estimate of the generalization power of the analysis that is better than that obtained using only a separate test set.

Multiple linear regression for delta-endotoxin yields

Multiple regression analysis was carried out to get an estimate of the predictive value of the considered model composed by seven dependent variables. The standardized coefficients (also named b coefficients), the standard error of coefficients, the t values and their related p values are all presented in Table 4. The large t value ($t = 43.3$) and corresponding low p value ($p < 0.01$) support the significance of soybean meal. On the other hand, there was a significant opposite relationship between K_2HPO_4 and delta-endotoxin yield, with a negative t value ($t = -16.87$) and corresponding low p value ($p < 0.01$).

The results of the model fitting by analysis of variance (ANOVA) are given in Table 5. The Fisher variance ratio, (F value), is a statistically suitable measure of how well the factors explain the variation in the data about its mean.

The better the F value, the more efficiently the factors describe the variation in the data about its mean, and the more valid calculated factor effects are. The ANOVA of the regression model reveals that the model is well significant, as is obvious from the Fisher's F test ($F_{model} = 384.33$) and low probability value (p value = 0.000 < 0.05).

The goodness of the fit of the model was checked by the determination coefficient (R^2). The R^2 value provides a measure of how much variability in the observed response values can be explained by the experimental variables. In this case, the value of the determination coefficient ($R^2 = 0.731$) indicates that 73.1% of the variability in the response could be explained by the model. In addition, the value of the adjusted determination coefficient (adj $R^2 = 0.729$) is also high enough to advocate for a high significance of the model. The adjusted coefficient of determination (adj R^2) is a statistical measure that shows the proportion of *variation* explained by the estimated regression line. The closer adjusted R^2 is to 1, the better the estimated regression equation fits or explains the relationship between X and Y. Olori et al. (1999) stated that $R^2 > 0.70$ implies a very good fit of a model, while if

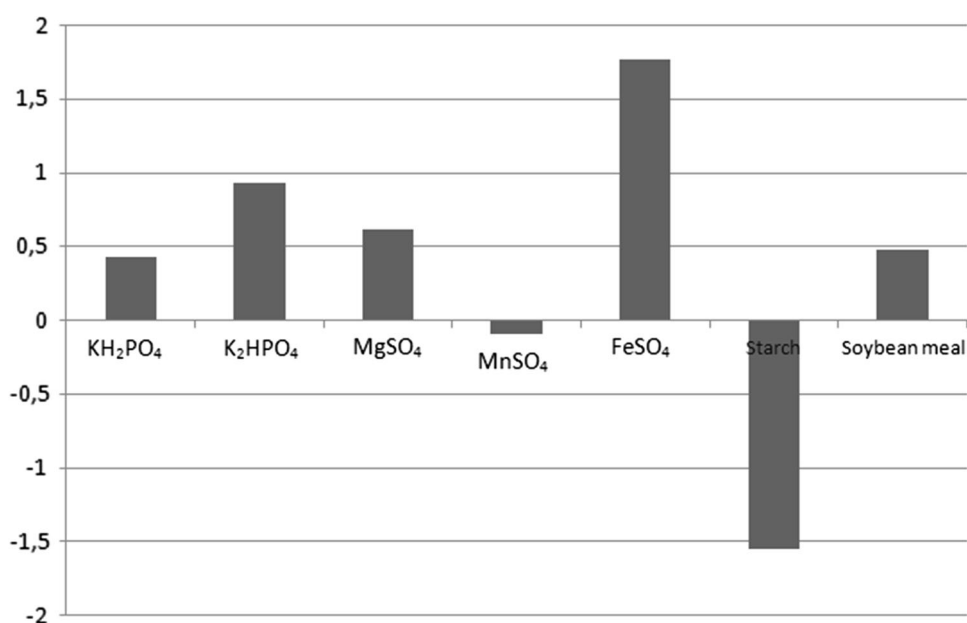**Fig. 2** Effects of different medium components for protease yield

Table 4 model coefficients estimated by multiple linear regression for delta-endotoxin and protease yields (significance of regression coefficients)

	Delta-endotoxin			Protease		
	Coefficient	t value	p value	Coefficient ($\times 10^{-3}$)	t value	p value
Constant	−34.914	−18.80	0.000	3229.2	15.92	0.000
KH_2PO_4	0.002	4.99	0.000	0.4890	10.80	0.000
K_2HPO_4	−0.0068	−16.87	0.000	0.9144	20.55	0.000
$MgSO_4$	0.0171	16.41	0.000	1.7069	15.01	0.000
$MnSO_4$	0.1488	6.94	0.000	−4.223	−1.80	0.072
$FeSO_4$	0.0656	3.14	0.002	90.489	39.61	0.000
Starch	0.0006	15.72	0.000	−0.1523	−32.11	0.000
Soybean meal	0.0019	43.30	0.000	0.0512	10.68	0.000

$R^2_{\text{Delta−endotoxin}} = 73.1\%, \quad R^2_{\text{Protease}} = 77.2\%$

$R^2(\text{adj})_{\text{Delta−endotoxin}} = 72.9\%, \quad R^2(\text{adj})_{\text{Protease}} = 77.0\%$

$R^2 < 0.40$, such a model should not be used for prediction. The Normal probability plot showed a satisfactory correlation between the experimental and predicted values of delta-endotoxin yields, wherein, the points cluster around the diagonal line indicated the good fit of the model because the deviation between the experimental and predicted values is low.

Multiple regression results for protease yield

Tables 4 and 5 present the results from the multiple regressions carried out using the medium components as the independent variables and protease yield as the dependent variable. This was done to determine the best linear combination of estimated models for predicting delta-endotoxin and protease yield responses. From Table 4, it can be seen that 77.2% of the variance in the model can be predicted using the independent variables.

Table 4 shows the standardized Beta Coefficients that present the contributions of each variable to the model. The t and p values show the impact of the independent variables on the dependent variable. All independent variables are significant except for $MnSO_4$ where $p > 0.05$. The large t value ($t = 39.61$) and corresponding low p value ($p < 0.01$) support the result for $FeSO_4$ which had the highest beta coefficient. On the other hand, there was a significant negative relationship between starch and protease yield, with a negative t value ($t = -32.11$) and corresponding low p value ($p < 0.01$). The value of the determination coefficient ($R^2 = 0.772$) indicates that 77.2% of the variability in the response could be explained by the model. Besides, the value of the adjusted determination coefficient (adj $R^2 = 0.77$) is also high enough to suggest the high significance of the model.

Table 5 presents the ANOVA report on the general significance of the model. As p value is less than 0.05, the model is significant. Thus, the combination of the variables

Table 5 ANOVA for the estimated models of delta-endotoxin and protease yield responses

Source	Degree of freedom	Sum of square	MS	F	p value
Delta-endotoxin					
Regression	7	9595.2	1370.7	384.33	0.000
Residual error	992	3538.0	3.6		
Total	999	13,133.2			
Protease					
Regression	7	142.557	20.365	478.67	0.000
Residual error	992	42.205	0.043		
Total	999	184.762			

Table 6 Performance indices (R^2 and MSE) for different models of delta-endotoxin prediction

Type of neural network	Neuron number in hidden layer	Training performance	Test performance	Hidden function	Training MSE ($\times 10^{-3}$)	Test MSE ($\times 10^{-3}$)
MLP	25	0.882137	0.867755	Tanh	4.514	5.146
MLP	36	0.879237	0.867722	Tanh	4.622	5.144
MLP	49	0.884775	0.868974	Tanh	4.418	5.112
MLP	34	0.881570	0.868693	Exponential	4.538	5.100
MLP	23	0.881989	0.869266	Exponential	4.518	5.088
MLP	22	0.881273	0.871280	Exponential	4.546	5.004
RBF	29	0.835893	0.829874	Gaussian	6.128	6.502
RBF	35	0.847601	0.832053	Gaussian	5.728	6.438
RBF	43	0.859240	0.830132	Gaussian	5.324	6.496
RBF	50	0.866231	0.846248	Gaussian	5.078	5.908
RBF	58	0.850053	0.848736	Gaussian	5.644	5.836
RBF	69	0.863802	0.837950	Gaussian	5.164	6.234

Fig. 3 Observed and predicted data of delta-endotoxin yields
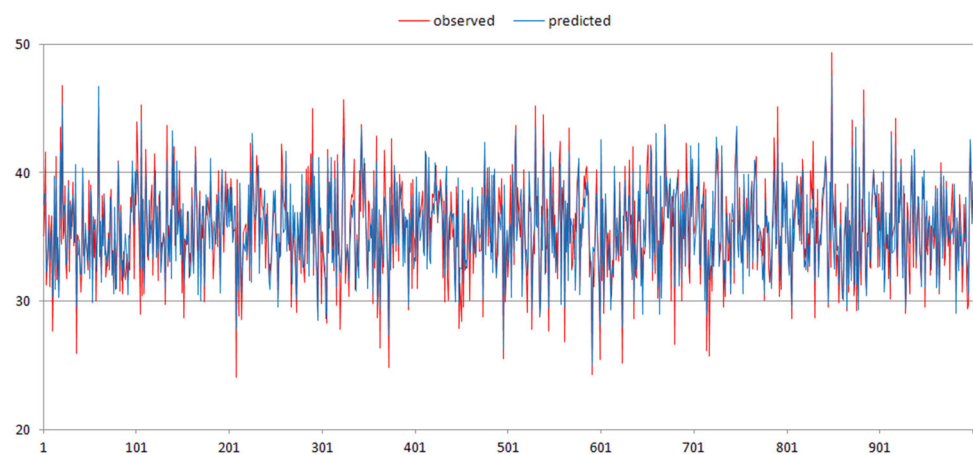
significantly predicts the dependent variable ($F = 478.67$; $p < 0.05$).

Application of artificial neural network for delta-endotoxin yields

To develop a model based on Neural Network performance for delta-endotoxin prediction, several ANN networks in MLP and RBF structure were constructed and tested to determine the optimum number of neurons, hidden layers and transfer functions. In fact, establishing an adequate structure with the suitable number of hidden layers, and neurons is important, since a larger number could result in over-fitting, while a smaller number, may not process the data sufficiently. Several MLP and RBF networks were developed and trained using the learning data set of delta-endotoxin concentration and then they were validated with the test data set (Table 6). The optimal obtained network model with maximum coefficient of determination (R^2) and minimum training and testing MSE was selected.

In this study, ANNs were used to forecast delta-endotoxin yields, In fact, after several model runs, the best structure with the highest R^2 and lowest MSE was depicted. This structure constitutes the selected ANN model for the delta-endotoxin prediction. It is an MLP with 3 layer perceptron described as follows: 1 input layer with 7 input variables, 1 hidden layer with 22 neurons and one output layer with one output variable. This model explains the variation of about 87% of delta-endotoxin production by the medium strain composition with small training and test errors (MSE were 4.546×10^{-3} and 5.004×10^{-3}, respectively). The predicted and observed production of delta-endotoxins is shown in Fig. 3.

Application of artificial neural network for proteases

To develop a model for protease production, the ANN was composed of seven neurons in the input layer each one corresponded to a medium component, and one neuron in

Table 7 Performance indices (R^2 and MSE) for different models for protease predictions

Type of neural network	Neuron number in hidden layer	Training performance	Test performance	Hidden function	Training MSE ($\times 10^{-3}$)	Test MSE ($\times 10^{-3}$)
MLP	70	0.896712	0.888043	Tanh	4.456	4.292
MLP	64	0.899958	0.889069	Tanh	4.318	4.252
MLP	55	0.894867	0.891583	Tanh	4.524	4.184
MLP	53	0.903104	0.884769	Exponential	4.194	4.420
MLP	56	0.905071	0.885385	Exponential	4.108	4.400
MLP	60	0.903214	0.887035	Exponential	4.184	4.350
RBF	70	0.879007	0.856292	Gaussian	5.162	5.386
RBF	68	0.874135	0.846058	Gaussian	5.356	5.764
RBF	58	0.890203	0.848720	Gaussian	4.712	5.730
RBF	55	0.883876	0.859407	Gaussian	4.968	5.274
RBF	54	0.869819	0.851357	Gaussian	5.528	5.602
RBF	51	0.894291	0.863359	Gaussian	4.548	5.170

Fig. 4 Observed and predicted data of protease yields
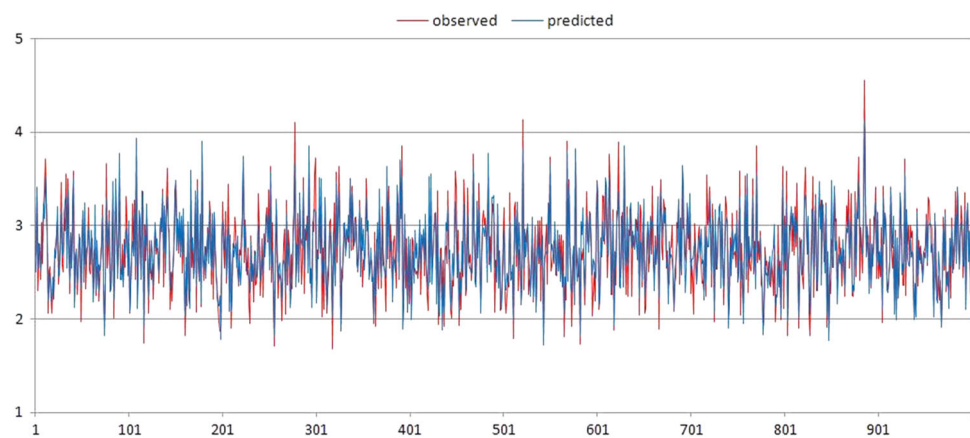
the output layer that stood for protease production. As the modelling of delta-endotoxin production, a process of trial and error in determining the number of hidden layer neurons and the threshold is also used. In this study, the neural network training function in MLP (exponential and tangent hyperbolic functions) and RBF architectures were carried out. To determine an optimal number of hidden layer neurons that is required to obtain the most satisfactory application performance, the results are compared based on the determination coefficient R^2 that shows the model performances.

On the basis of results in the Table 7, it can be concluded that the most suitable model is an MLP trained with 7 medium components for the input parameters (7–55–1) including 7 input neurons, 55 hidden neurons and 1 output. The results of the model trained with a threshold function tangent hyperbolic created high determination coefficients ($R^2 = 0.89$) and small errors (MSE = 4.524×10^{-3} and 4.184×10^{-3} for training and test, respectively). The

predicted protease yields can be considered successful. Figure 4 shows the observed and predicted results.

Table 8 showed the Paired samples t test between observed and predicted data. The paired samples t test is commonly used to test the Statistical difference between two measurements. Mean represented the average difference between the two variables. The standard error (standard deviation divided by the square root of the sample size) is generally used in computing both the test statistic and the upper and lower bounds of the confidence interval. The paired t test statistic was denoted (t) and the p value corresponding to the given test statistic t. From the results, we can say that observed and predicted endotoxin and protease yields were highly and positively correlated [($r = 0.879$, $p < 0.05$) and ($r = 0.888$, $p < 0.05$), respectively]. There was insignificant average difference between observed and predicted delta-endotoxin yields ($t_{999} = -0.111$, $p = 0.911$) and observed and predicted protease yields ($t_{999} = -0.647$, $p = 0.518$). These results

Table 8 Paired samples test of observed and predicted values

		Paired differences					t	df	p
		Mean	Std. deviation	Std. error mean	95% Confidence interval of the difference				
					Lower	Upper			
Pair 1	Observed–predicted endotoxins	−0.006	1.728	0.054	−0.113	0.101	−0.111	999	0.911
Pair 2	Observed–predicted proteases	−0.004	0.197	0.006	−0.016	0.008	−0.647	999	0.518

confirmed the robustness of established ANN models with insignificant difference with observed and predicted data. On average, observed delta-endotoxin yields were similar to predicted yields (95% CI [−0.113, 0.101]). Furthermore, observed protease yields were similar to predicted yields (95% CI [−0.016, 0.008]).

In this work, an attempt was made to analyze and compare MLR and ANN including MLP and RBF models to develop transfer function for predicting bacterial yields using medium ingredients. The statistical prediction performances of used models were measured in terms of determination coefficient (R^2) and MSE. Hence, the results obtained from the MLR models indicated that the measured determination coefficient between the observed and the predicted data were acceptable with a determination coefficient higher than 70% for both delta-endotoxin and protease yields. Furthermore, ANN models showed higher accuracy (87 and 89% for delta-endotoxin and protease determination coefficients, respectively) when compared with MLR models (73.1 and 77.2% for delta-endotoxin and protease determination coefficients, respectively). Moreover, the MLP models exposed more stable forecast when compared with the RBF model, according to descriptive performance indices (R^2 and MSE) between forecasted and experimental data. Subsequently, with the use of the proposed process with statistical bootstrap that enlarged the sample of data and MLP network learning, the *B. thuringiensis* var. *kurstaki* yield prediction could be determined by performing a limited number of experiments and test operations, therefore, saving engineering effort and time.

Conclusion

This work proposes a method to analyze simultaneously multiple factors with a limited number of experiments using the Plackett–Burman analysis. To enhance delta-endotoxin production in *B. thuringiensis* var. *kurstaki* HD-1, we used a statistical experimental design. Findings demonstrate that medium components play a significant role in the growth of *B. thuringiensis* var. *kurstaki* and

proteolytic enzyme production. In fact, KH_2PO_4, $MgSO_4$, $MnSO_4$, $FeSO_4$, starch and soybean meal increase delta-endotoxin yield, whereas K_2HPO_4 decreases delta-endotoxin yield. Remarkably, the obtained results in this study facilitate the large-scale, economically viable delta-endotoxin production using *B. thuringiensis*. Moreover, KH_2PO_4, $MgSO_4$, K_2HPO_4, $FeSO_4$ and soybean meal increase protease yield, whereas starch and $MnSO_4$ decrease protease yield.

Besides, we found that a neural network analysis is potentially more successful than MLR in predicting response. In fact, we developed an artificial neural network that yielded a higher level of correct forecast for response than the MLR method. Thus, the ANN analysis seems to be a more efficient method for the prediction of studied responses. This finding implies that the prediction of delta-endotoxin and protease yields responses may involve a complicated non-linear relationship. Furthermore, based on the obtained results, MLP seems to be the most adequate ANN model for *B. thuringiensis* yield predictions. However, the activation function and the number of hidden neurons are specific for each type of forecasting responses (delta-endotoxin or protease yields).

The objectives of the present study were initially to find the adequate topology of the ANN and regression models for prediction of delta-endotoxin and proteases yields. Secondly, the current work aimed to select the best method in prediction of microbial yields and thus selecting the optimized topology. The obtained results showed that ANNs are able to identify complex parameters in datasets which may not be well explained by a simple mathematical formula. In fact, this work demonstrated how modern computational tools such as ANN can be successfully used to address this type of problems.

Compliance with ethical standards

Conflict of interest The authors did not declare any conflict of interest.

References

Adinarayana K, Ellaiah P, Prasad DS (2003) Purification and partial characterization of thermostable serine alkaline protease from a new isolated *Bacillus subtilis* PE-11. AAPS PharmSciTech 4:440–448. doi:10.1208/pt040456

Avignone Rossa CA, Yantorno OM, Arcas JA, Ertola RJ (1990) Organic and inorganic nitrogen source ratio effects on *Bacillus thuringiensis* var. *israelensis* delta-endotoxin production. World J Microbiol Biotechnol 6:27–31. doi:10.1007/BF01225351

Baffi G, Martin EB, Morris AJ (1999) Non-linear projection to latent structures revisited (the neural network PLS algorithm). Comp Chem Eng 23:1293–1307. doi:10.1016/S0098-1354(99)00291-4

Beale R, Jackson T (1990) Neural computing—an introduction. CRC Press, New York

Bradford M (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72:248–254. doi:10.1016/0003-2697(76)90527-3

Chao YY, Hong CY, Kao CH (2010) The decline in ascorbic acid content is associated with cadmium toxicity of rice seedlings. Plant Physiol Biochem 48:374–381. doi:10.1016/j.plaphy.2010.01.009

Charles JF, Silva-Filha MH, Nielsen-LeRoux C, Humphreys M, Berry C (1997) Binding of the 51- and 42-kDa individual components from the *Bacillus sphaericus* crystal toxin on mosquito larval midgut membranes from *Culex* and *Anopheles* sp. (Diptera: Culicidae). FEMS Microbiol Lett 156:153–159. doi:10.1111/j.1574-6968.1997.tb12721.x

Chen FC, Shen IF, Chak KE (2004) A facile analytical method for the identification of protease gene profiles in *Bacillus thuringiensis* strains. J Microbiol Methods 56:125–131

Chen Y, Huang Y, Lin R, Shang N (2010) A continuous-flow biodiesel production process using a rotating packed bed. Bioresour Technol 101:668–673. doi:10.1016/j.biortech.2009.08.081

Chu WH (2007) Optimization of extracellular alkaline protease production from species of *Bacillus*. J Ind Microbiol Biotechnol 34:241–245. doi:10.1007/s10295-006-0192-2

El-Bendary MA (1999) Growth physiology and production of mosquitocidal toxins from *Bacillus sphaericus*. Dissertation, Ain-Shams University

El-Bendary MA (2006) *Bacillus thuringiensis* and *Bacillus sphaericus* biopesticides production. J Basic Microbiol 46:158–170. doi:10.1002/jobm.200510585

Ellaiah P, Srinivasulu B, Adinarayana K (2002) A review on microbial alkaline proteases. J Sci Ind Res 61:690–704

Ennouri K, Ben Hassen H, Ben Khedher S, Zouari N (2013a) Concomitant production of proteases and delta-endotoxins of *Bacillus thuringiensis* subsp. *kurstaki* in a low-cost medium: effect of medium components. Acta Biol Szeged 57:13–19

Ennouri K, Ben Hassen H, Zouari N (2013b) Optimization of bioinsecticides overproduction by *Bacillus thuringiensis* subsp. *kurstaki* using linear regression. Pol J Microbiol 62:287–293

Ennouri K, Ben Khedher S, Jaoua S, Zouari N (2013c) Correlation between delta-endotoxin and proteolytic activities produced by *Bacillus thuringiensis* subsp. *kurstaki* growing in an economic production medium. Biocontrol Sci Technol 23(7):756–767

Frankena J, Van Verseveld HW, Stouthamer AH (1985) A continuous culture study of the bioenergetic aspects of growth and production of exocellular protease in *Bacillus licheniformis*. Appl Microbiol Biotechnol 22:169–176. doi:10.1007/BF00253604

Ghribi D, Zouari N, Jaoua S (2004) Improvement of bioinsecticide production through mutagenesis of *Bacillus thuringiensis* by

U.V. and nitrous acid affecting metabolic pathways and/or delta-endotoxin synthesis. J Appl Microbiol 97:338–346. doi:10.1111/j.1365-2672.2004.02323.x

Ghribi D, Zouari N, Trigui W, Jaoua S (2007) Use of sea water as salts source in starch- and soya bean-based media, for the production of *Bacillus thuringiensis* bioinsecticides. Process Biochem 42:374–378. doi:10.1016/j.procbio.2006.09.006

Haq IU, Mukhtar H, Umber H (2006) Production of protease by *Penicillium chrysogenum* through optimization of environmental conditions. J Agric Soc Sci 2:23–25

Hartley BS (1960) Proteolytic enzymes. Annu Rev Biochem 29:45–72

Hotha S, Banik RM (1997) Production of alkaline protease by *Bacillus thuringiensis* H14 in aqueous two-phase systems. J Chem Tech Biotechnol 69:5–10. doi:10.1002/(SICI)1097-4660(199705)69:1<5:AID-JCTB661>3.0.CO;2-I

Johnveshy B, Naik G (2001) Production of bleach-stable and halo-tolerant alkaline protease by an alkalophilic *Bacillus pumilus* JB-05 isolated from cement industry effluents. J Microbiol Biotechnol 11:558–563

Joo HS, Kumar CG, Park GC, Kim KT, Paik SR, Chang CS (2002) Optimization of the production of an extracellular alkaline protease from *Bacillus horikoshii*. Process Biochem 38:155–159. doi:10.1016/S0032-9592(02)00061-4

Khaouane L, Benkortbi O, Hanini S, Si-Moussa C (2013) Modeling of an industrial process of pleuromutilin fermentation using feed-forward neural networks. Braz J Chem Eng 30:105–116. doi:10.1590/S0104-66322013000100012

Krishnan S, Prapulla SG, Rajalakshmi D, Misra MC, Karanth NG (1998) Screening and selection of media components for lactic-acid production using Plackett–Burman design. Bioprocess Eng 19:61–65. doi:10.1007/PL00009003

Lachhab K (2001) Production of *Bacillus thuringiensis* biopesticides using wastewater sludge as a raw material: effect of inoculum and sludge solids. Process Biochem 37:197–208. doi:10.1016/S0032-9592(01)00198-4

Lee DS, Park JM (1999) Neural network modeling for on-line estimation of nutrient dynamics in a sequentially-operated batch reactor. J Biotechnol 75:229–239. doi:10.1016/S0168-1656(99)00171-6

Malinova T, Guo ZX (2004) Artificial neural network modelling of hydrogen storage properties of Mg-based alloys. Mater Sci Eng A 365:219–227. doi:10.1016/j.msea.2003.09.031

McKellar RC, Cholette H (1984) Synthesis of extracellular proteinase by *Pseudomonas fluorescens* under conditions of limiting carbon, nitrogen, and phosphate. Appl Environ Microbiol 47:1224–1227

Miller JN, Miller JC (2000) Statistics and chemometrics for analytical chemistry. Prentice Hall, London

Montgomery DC (1997) Design and analysis of experiments, 4th edn. Wiley, New York

Moon SH, Parulekar SJ (1991) A parametric study of protease production in batch and fed-batch cultures of *Bacillus firmus*. Biotechnol Bioeng 37:467–483

Nabney IT (2002) Netlab: algorithms for pattern recognition. Springer, Great Britain

Navon A (2000) *Bacillus thuringiensis* application in agriculture. In: Charles JF, Delécluse A, Nielsen-Le Roux C (eds) Entomopathogenic bacteria: from laboratory to field application. Kluwer Academic Publishers, Dordrecht, pp 355–369

Olori VE, Brotherstone S, Hill WG, McGuirk BJ (1999) Fit of standard models of the lactation curve to weekly records of milk production of cows in a single herd. Livest Prod Sci 58:55–63. doi:10.1016/S0301-6226(98)00194-8

Ozkan M, Dilek FB, Yetis U, Ozcengiz G (2003) Nutritional and cultural parameters influencing antidipteran delta-endotoxin

production. Res Microbiol 154:49–53. doi:10.1016/S0923-2508(02)00006-2

Pachepsky YA, Timlin D, Varallyay G (1996) Artificial neural networks to estimate soil water retention from easily measurable data. Soil Sci Soc Am J 60:727–733. doi:10.2136/sssaj1996.03615995006000030007x

Pujari V, Chandra TS (2000) Statistical optimization of medium components for improved synthesis of riboflavin by *Eremothecium ashbyii*. Bioprocess Biosyst Eng 23:303–307. doi:10.1007/PL00009127

Shafizadeh-Moghadem H, Hagenauer J, Farajzadeh M, Helbich M (2015) Performance analysis of radial basis function networks and multi-layer perceptron networks in modeling urban change: a case study. Int J Geogr Inf Sci 29:606–623. doi:10.1080/13658816.2014.993989

Singh J, Vohra RM, Sahoo DK (2004) Enhanced production of alkaline proteases by *Bacillus sphaericus* using fed-batch culture. Process Biochem 39:1093–1101

Song RG, Zhang QZ, Tseng MK, Zhang BJ (1995) The application of artificial neural networks to the investigation of aging dynamics in 7175 aluminium alloys. Mater Sci Eng C 3:39–41. doi:10.1016/0928-4931(95)00068-2

Stanbury PF, Whitakar A, Hall SJ (1997) Principles of fermentation technology. Aditya Books, New Delhi

Strauch MA, Hoch JA (1993) Transition-state regulators: Sentinels of *Bacillus subtilis* post-exponential phase gene expression. Mol Microbiol 7:337–342. doi:10.1111/j.1365-2958.1993.tb01125.x

Tadeusiewicz R (1993) Neural networks. AOW, Warsaw

Tyagi RD, Sikati Foko V, Barnabe S, Vidyarthi A, Valero JR, Surampalli RY (2002) Simultaneous production of biopesticide and alkaline protease by *Bacillus thuringiensis* using waste water as a raw material. Water Sci Technol 46:247–254

Van Frankenhuyzen K (2013) Cross-order and cross-phylum activity of *Bacillus thuringiensis* pesticidal proteins. J Invertebr Pathol 114:76–85. doi:10.1016/j.jip.2013.05.010

Van Liew MW, Arnold JG, Garbrecht JD (2003) Hydrologic simulation on agricultural watersheds: choosing between two models. Trans ASAE 46:1539–1551. doi:10.13031/2013.15643

Zare M, Pourghasemi HR, Vafakhah M, Pradhan B (2013) Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: a comparison between multi-layer perceptron (MLP) and radial basic function (RBF) algorithms. Arab JGeosci 6:2873–2888. doi:10.1007/s12517-012-0610-x

Zouari N, Jaoua S (1999) Production and characterization of metalloproteases synthesized concomitantly with δ-endotoxin by *Bacillus thuringiensis* subsp. *kurstaki* strain grown on gruel-based media. Enzyme Microb Technol 25(3):364–371

Zouari N, Dhouib A, Ellouz R, Jaoua S (1998) Nutritional requirements of strain of *Bacillus thuringiensis* subsp. *kurstaki* and use of gruel hydrolysate, for the formulation of a new medium for delta-endotoxin production. Appl Biochem Biotechnol 69:41–52