Soft-Sensors for Monitoring *B. Thuringiensis* Bioproduction

C. E. Robles Rodriguez[1], J. Abboud[2], N. Abdelmalek[3], S. Rouis[4], N. Bensaid[3], M. Kallassy[5], J. Cescut[2], L. Fillaudeau[1], and C. A. Aceves Lara[1(✉)]

[1] Toulouse Biotechnology Institute, Bio & Chemical Engineering, and the CNRS, UMR5504, Université de Toulouse, UPS, INSA, INP, TBI F31077, the INRA, UMR792, 135 avenue de Rangueil, 31077 Toulouse, France
{roblesro,luc.fillaudeau,aceves}@insa-toulouse.fr
[2] Toulouse White Biotechnology (UMS INRAE/INSA/CNRS), 135 Avenue de Rangueil, 31077 Toulouse CEDEX 04, France
{joanna.abboud-1,Julien.Cescut}@inrae.fr
[3] Laboratoires Pharmaceutiques MédiS, Nabeul, Tunisia
nadia.bensaid@labomedis.com
[4] Centre of Biotechnology of Sfax, B.P 1177, 3018 Sfax, Tunisia
souad.rouis@cbs.rnrt.tn
[5] Faculty of Sciences Saint, Joseph University, Riad El Solh, Beirut, Lebanon
mireille.kallassy@usj.edu.lb

Abstract. One of the main challenges in ferentation cultures is the monitoring of key variables that can indicate the performance and help in the optimization of the bioprocess. On-line estimation could be a challenging task when an accurate process model is not available. Support Vector Machine (SVM) is an attractive and relatively simple method that can be used as an alternative to predict key variables by using several physical parameters measured online. In this paper, we show the application of SVM to the production of protein by *B. thuringiensis*. The soft-sensor was trained and validated with independent data sets of batch fermentations evaluating the impact of different predictor variables and kernels. Results show that protein production can be predicted only using online measurements.

Keywords: Soft-sensors · Monitoring · Support vector machine · Proteins and Spores optimization · B. thuringiensis bioproduction

1 Introduction

B. thuringiensis is a facultative anaerobic gram-positive sporulating bacterium, frequently used in the production of some biopesticides and as a source of genes for transgenic expression in plants [1]. It is usually found in different environments, among which soil, settled dust, insects, water, and others have been identified [2]. *B. thuringiensis* has been shown to be toxic to various organisms such as lepidopterans, coleopterans, dipterans, or nematodes, but is considered safe for mammals. Thus, the products based on *B. thuringiensis* (Bt) provide effective and environmentally benign control of

K. Matsui et al. (Eds.): DCAI 2021, LNNS 327, pp. 129–136, 2022.
https://doi.org/10.1007/978-3-030-86261-9_13

several insects in agricultural, forestry and disease-vector applications [3]. This insecticidal activity is mainly due to the production of some intracellular inclusions (called δ-endotoxins) during the post-exponential phase of *B. thuringiensis* cells. Most of the biopesticides distributed in the world are principally based on *B. kurstaki* HD1 strain. However, a recent strain, identified as *B. kurstaki* Lip, has been isolated and described to be more efficient than HD1 [4]. Therefore, this last strain will be studied in this work.

The monitoring of certain variables of *B. thuringiensis* culture is complicated due to several changes of cell physiology during growth, which will hamper the optimization of the fermentation. Usually, it is difficult to measure online substrate, biomass, and product concentrations in the bio-process. The so-called "soft-sensors" are an alternative for online estimation. Soft sensors are software based sophisticated monitoring systems, which can relate the infrequently measured process variables with the easily measured [5]. In this way, these soft-sensors assist in obtaining a real-time prediction of the unmeasured variables [6].

Several software sensors have been proposed for fermentation such as Support Vector Machine (SVM) [7] and Decision Trees DT [8]. A recent review paper [9] has shown the use of methods like neural networks, fuzzy logic, SVM, genetic algorithms and probabilistic latent variable models in fermentation, where the authors highlighted that SVM has become an indispensable method to measure internal variables, especially when small amount of data exists [10]. Furthermore, SVM shares many of its features with the artificial neural networks, but it proposes some additional characteristics [5]. It has good generalization ability of the regression function, robustness of the solution, and sparseness of the regression [6].

In this context, this work proposes some SVM soft-sensors to monitoring the production of a protein by *B. thuringiensis* with the purpose of showing the application of SVM for microorganisms with physiology changes during fermentation. The remainder of the paper is as follows: the experiments are presented in Sect. 2, the methodology of soft-sensors is presented in Sect. 3. Section 4 holds the results of the soft-sensor with the training and validation data. Finally, Sect. 5 reports the conclusions and some perspectives of this work.

2 Material and Methods

2.1 Organism and Culture Media

B. thuringiensis Lip is a Lebanese strain [4]. Luria broth (LB) was used for inoculum production, whereas, a semi-synthetic medium (SSM). For the SSM, concentrated glucose (Sol 2) and all salts solutions (Sol 3, 4, 5) were prepared and sterilized separately and added before inoculation to the rest of medium (Sol 1) previously sterilized.

2.2 Fermentation Conditions

Several fermentations were conducted in batch mode at 30 °C in a 3 L Biostat B plus fermenter (Sartorius; Germany) containing 1.8 L of the SSM medium and with continuous regulation of pH at 6.8 using 1 M H3SO4 and 3 M NaOH. Dissolved oxygen was

continuously monitored by an optical oxygen sensor and maintained at 25% pO2sat with a constant aeration rate (0.18min/L) and variable stirring. Foaming was controlled by the use of an antifoam (Emultrol DFM DV-14 FG), through the fermentation process.

2.3 Total Cell and Spores Count

The follow up of the Bt culture was performed by estimating total cell counts (TC) and spore counts (SC) by plate counts. To determine TC, the withdrawn samples were serially diluted, spread on LB plates and incubated at 30 °C for 16–18 h. As for SC, the appropriately diluted samples were heated at 85 °C for 15 min and cooled for 5 min before spreading onto LB plates and then incubated at 30 °C.

2.4 Dry Matter

Biomass dry weight (DW) concentration was determined by filtering a known amount of sample and differential weighing of the filter before and after oven drying at 70 °C.

2.5 Quantification of Delta Endotoxins Production

The concentration of the delta-endotoxins was determined by Bradford assay using bovine serum albumin (BSA) as a protein standard. Samples were measured after 10 min at 595 nm. The obtained value was the average of three measures of the same sample.

2.6 Sugar Analysis

The sugar concentrations were determined using HPLC-UV with a (Ultimate 3000 RSLC/MWD/RI/CAD). A mobile phase of 5 mM H2SO4 with a flow rate of 0.6 mL/min was used. The mobile phase was filtered and degassed through a 0.2 μm cellulose nitrate membrane. The samples and standards were also filtered before injection into the HPLC.

3 Support Vector Machine

Support vector machine (SVM) is a kernel-based tool for solving pattern recognition and regression problems. The idea of SVM relies on mapping the input data or features $x \in R^{M \times N}$ into a nonlinear space in order to predict a desired vector of outputs $y \in R^{M}$ [9]. The objective of the regression analysis is to determine a function that predicts accurately the desired outputs y in the form

$$y = w^{T} \varphi(x) + b \tag{1}$$

where $\varphi(x)$ is the nonlinear mapping of the inputs x into a high dimensional feature space. The vector w represents the support vectors and b is the bias term. The determination of the support vectors is performed solving the following optimization problem:

$$\underset{w,\, \xi,\, \xi^{*}}{min}\, J = \frac{1}{2} w^{2} + C \sum\nolimits_{i=i}^{N} \left(\xi_{\iota} + \xi_{i}^{*}\right) \tag{2}$$

$$\text{Subject to} \begin{cases} d_i \leq \varepsilon + \xi_i \\ -d_i \leq \varepsilon + \xi_i^* \\ \xi_i,\ \xi_i^* \geq 0 \end{cases} \tag{3}$$

with $d_i = y_i - w^T \varphi(x_i) - b$.

In Eq. (2) the first term is the regularized term, and the second term is the empirical error (risk) measured by the insensitive ε-loss function enabling to use less data points to represent the decision function given by Eq. (1). The variables ξ_i and $\xi_i{}^*$ are the slack variables that measure the deviation of the support vectors from the boundaries of the ε -zone and determine how strictly the model fits the data. The constant C is the regularization constant. It determines the trade-off between the empirical risk and the regularized term. The term ε is called the tube size and it is equivalent to the approximation accuracy placed on the training data points. Both parameters will determine the efficiency of the estimation [5].

In order to simplify the dual minimization problem, Lagrange multipliers are introduced as follows:

$$L = J - \sum_{i=1}^{N} \alpha_i \{\xi_i + \varepsilon - d_i\} - \sum_{i=1}^{N} \alpha_i^* \left\{\xi_i^* + \varepsilon + d_i\right\} - \sum_{i=1}^{N} \left(\eta_i \xi_i - \eta_i^* \xi_i^*\right) \tag{4}$$

where the parameters α_i, α_i^*, η_i, and η_i^* are the Lagrange multipliers. According to the Karush-Kuhn-Tucker (KKT) of quadratic programming, the dual equation that can be obtained [6] is:

$$min_{\alpha,\, \alpha^*}\, W = \frac{1}{2} \sum_{i,j=i}^{N} \left(\alpha_\iota - \alpha_i^*\right).\left(\alpha_j - \alpha_j^*\right) K\left(x_i,\ x_j\right) + \varepsilon \sum_{i=i}^{N} \left(\alpha_\iota + \alpha_i^*\right) - \sum_{i=i}^{N} \left(\alpha_\iota - \alpha_i^*\right) y_i \tag{5}$$

$$\text{Subject to} \begin{cases} \sum_{i=i}^{N} \left(\alpha_\iota - \alpha_i^*\right) \\ 0 \ \leq\ \alpha_\iota,\ \alpha_i^* \leq C; \quad i = 1, 2, \ldots, N \end{cases} \tag{6}$$

Therefore, the final regression function given in Eq. (1) is rewritten as

$$y = \sum_{i=i}^{N} \left(\alpha_\iota - \alpha_i^*\right) K\left(x_i,\ x_j\right) + b \tag{7}$$

where $K\left(x_i,\ x_j\right) = \varphi(x_i)\varphi\left(x_j\right)$ is the kernel function that corresponds to any symmetric function satisfying the Mercer's condition. The most typical examples of the kernel functions are the polynomial kernels and the Radial Basis Function (Gaussian) kernels whose mathematical representation is

$$K\left(x_i,\ x_j\right) = [(x.x_i) + 1]^d \tag{8}$$

$$K\left(x_i,\ x_j\right) = \exp\left(-\left(x - x_i^2 / 2\sigma^2\right)\right) \tag{9}$$

where d is the order of the polynomial and σ represents the width of the RBF. The most used Kernel function is the Radial Basis Function (RBF) because it can classify multidimensional data. As the prediction depends on the type of kernel and their involved parameter, in this work we will compare three types of kernels and assess their prediction capability for protein production.

4 Results

The support vector machine algorithm has been applied to predict Protein production by *B. thuringiensis*. The available data correspond to ten batch experiments performed with three different strains. The available online measurements of the experiments were partial oxygen pressure (pO2), pH, agitation (Agit) and aeration (Aer); whereas the off-line measurements regarded optical density (OD), Biomass (Bio), Glucose (Glc), flora (Fl), and spores (Sp). The nine variables have been used as inputs to SVM. Additionally, the logarithms of Flora and Spores have been added as inputs as well as the number of strains. Before performing SVM, the inputs were ranked according to the Neighborhood Component Analysis in order to guide feature selection. The most important variables corresponded to pO2, strain number, Biomass and Spores. The least relevant inputs for protein production were the Flora and the logarithm of the flora. As the kernel type is a limiting factor in the performance of SVM, we have explored three kernels to assess their application.

Table 1. Combinations to determine the best model for Protein prediction.

	Model							
	1	2	3	4	5	6	7	8
	Gaussian Kernel							
RMSEt	189.35	64.96	66.18	57.41	65.18	66.22	61.26	66.42
RMSEv	207.50	174.57	133.62	152.31	129.26	126.65	165.01	156.05
	Quadratic Kernel							
RMSEt	200.74	128.01	103.62	94.75	50.45	53.40	191.29	185.30
RMSEv	194.30	316.80	166.32	175.14	138.05	151.54	228.61	235.21
	Linear Kernel							
RMSEt	218.98	208.41	186.31	175.43	122.89	134.42	145.46	148.65
RMSEv	199.95	184.66	174.75	189.71	218.33	162.90	168.17	172.29
Predictors								
OD	0	0	0	1	1	1	1	0
Bio	0	1	1	1	1	1	0	0
Glc	0	0	1	1	1	1	0	0
pO2	1	1	1	1	1	1	1	1
Agit	0	0	0	0	1	1	1	1
pH	0	0	0	1	1	1	1	1
Aer	0	0	0	0	1	1	1	1
Fl	0	0	0	0	0	1	0	0
Sp	0	0	1	1	1	1	0	0
Log(Fl)	0	0	0	0	0	1	0	0
Log(Sp)	0	1	1	1	1	1	0	0
Strain	1	1	1	1	1	1	1	1

The training of the SVM has been performed in MATLAB using the fitrsvm command from the regression learner application. Seven data sets were used for training with an 8 k-fold cross validation. The training sets comprised batch experiments: 2, 3, 4, 5, 7, 8, 10. The remaining three data sets consisting on batches 1, 6 and 9 one per strain number) were considered for the validation of the soft-sensor.

Several combinations of input variables have been explored with three different kernels: linear, quadratic and gaussian. The eight combinations have been assessed via the RMSE. The performance analysis of the SVM models are reported in Table 1 where RMSEt and RMSEv correspond to the errors for the training and validation, respectively. Table 1 indicates that the best prediction for the training data is achieved by Model 5 with a quadratic kernel. Furthermore, this model provides a good between the training and validation sets. However, it needs 10 predictor variables. The prediction with Model 5 is displayed in Fig. 1 where it can be appreciated that most of the data points are well predicted and some points could be removed to improve the performance since they seem to be outliers in the prediction. For instance, see the two last points of Batch 10, which are at a very low value. Furthermore, it is worth noting that the predictions are well adapted for the different types of strains producing lower titers of protein (*i.e.* batch 8, 9,10).

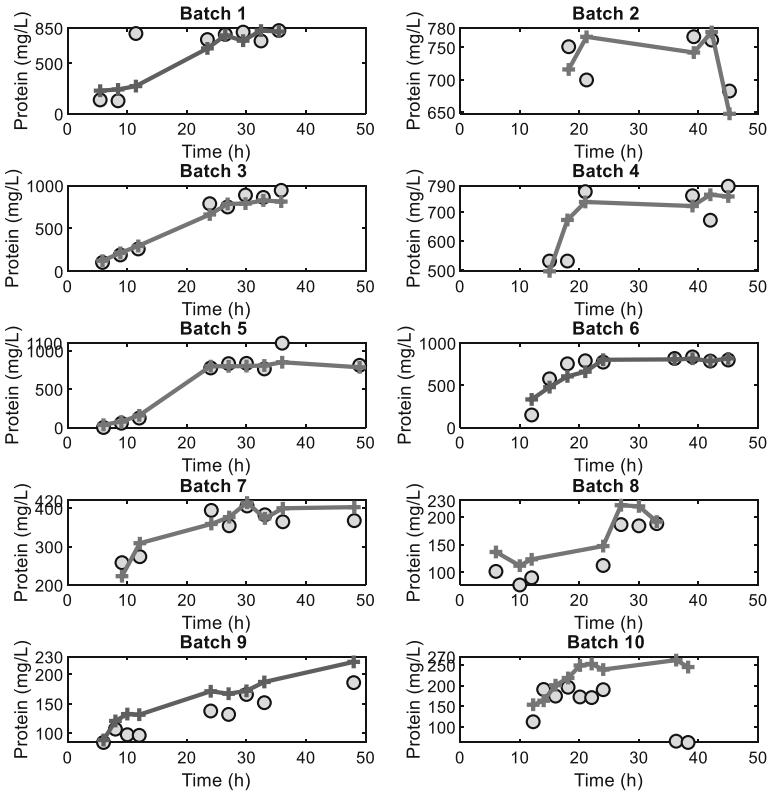

Fig. 1. Results of the SVM model 5 with the training sets (red lines) and the validation sets (blue lines). The dots represent the experimental data points.

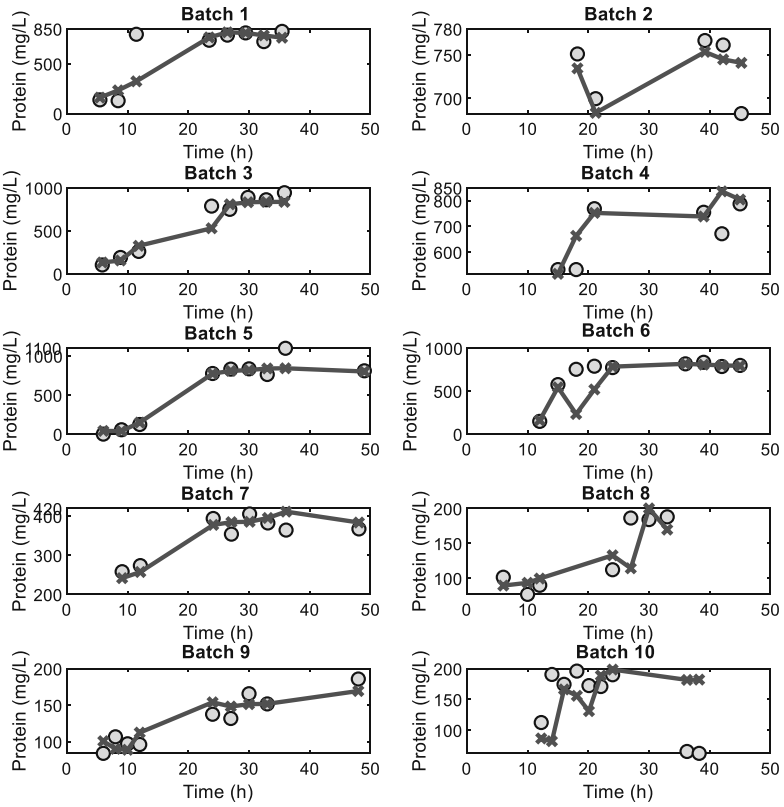

Fig. 2. Prediction of protein production with the SVM model 8 and gaussian kernel against the training sets (green lines) and the validation sets (purple lines). The dots represent the experimental data points.

Although model 5 provides good prediction, one of the goals of the development of soft-sensors for fermentation is their ability to be used at different conditions and scales. In this context, we have explored models only using online measurements. This is the case of model 8 and model 7 assuming that OD can sometimes be measured online. It is worth noting that model 8 which uses a gaussian kernel also provides a good compromise between the training and validation sets. Additionally, the RMSE values are similar to the results obtained by model 5. The prediction of model 8 is shown in Fig. 2. It is important to note that the one if the most important input is the strain number which allows to predict correctly among the different cultures. Model 8 is chosen as the best one because it presents higher online applicability, which is one of the objectives of the proposition of an online software sensor. These results highlight the fact that SVM can represent accurately the nonlinearities of the process.

5 Conclusions

This work introduced a soft-sensor based on SVM for the prediction of protein by different strains of *B. thuringiensis*. The SVM algorithm was successfully implemented to generate an online monitoring of the concentration of protein production, which is normally measured off-line. The results proved that SVM is an attractive for monitoring providing a good tradeoff between the quality of the approximation of the given data and

the complexity of the approximating function. Results have shown that diverse combinations of input variables can produce accurate predictions. However, the model only using online data is preferred due to the potential for extrapolability to other conditions and, especially, for industrial application. Future work will focus on increasing the quality of the prediction, the application of the soft-sensor in other experimental working conditions and the coupling of the soft-sensor for in-situ experiments.

References

1. Schnepf, E.: Bacillus thuringiensis and its pesticidal crystal proteins. Microbiol. Mol. Biol. Rev. **62**(3), 775–806 (1998). Chen, W.-K.: Linear Networks and Systems (Book style), pp. 123–135. Wadsworth, Belmont (1993)
2. Iriarte, J., Porcar, M., Lecadet, M.-M., Caballero, P.: Isolation and characterization of bacillus thuringiensis strains from aquatic environments in Spain. Curr. Microbiol. 40, 402–408 (2000). Smith, B.: An approach to graphs of linear forms (Unpublished work style). unpublished
3. Rowe, G.E., Margaritis, A.: Bioprocess design and economic analysis for the commercial production of environmentally friendly bioinsecticides from bacillus thuringiensis HD-1 kurstaki. Biotechnol. Bioeng. **86**(4) (2004)
4. El Khoury, M., Azzouz, H., Chavanieu, A., Abdelmalak, N., Chopineau, J., Awad, M.K.: Isolation and characterization of a new Bacillus thuringiensis strain Lip harboring a new cry1Aa gene highly toxic to Ephestia kuehniella (Lepidoptera: Pyralidae) larvae. Arch. Microbiol. **196**(6), 435–444 (2014)
5. Vapnik, V., Golowich, S.E., Smola, A.: Support vector method for function approximation, regression estimation, and signal processing. Annu. Conf. Neural Inf. Process. Syst. 281–287 (1996)
6. Liu, G., Zhou, D., Xu, H., Mei, C.: Model optimization of SVM for a fermentation soft sensor. Expert Syst. Appl. **37**, 2708–2713 (2010)
7. Ou Yang, H.-B., Li, S., Zhang, P., Kong, X.: Model penicillin fermentation by least squares support vector machine with tuning based on amended harmony search. Int. J. Biomath. **08**, 1550037 (2015)
8. Ahmad, M.W., Reynolds, J., Rezgui, Y.: Predictive modelling for solar thermal energy systems: a comparison of support vector regression, random forest, extra trees and regression trees. J. Clean. Prod. **203**, 810–821 (2018)
9. Zhu, X., Rehman, K.U., Wang, B., Shahzad, M.: Modern soft-sensing modeling methods for fermentation processes. Sensors(Switzerland), **20**(6), 1771 (2020).
10. Jianlin, W., Tao, Y.U., Cuiyun, J.I.N.: On-line estimation of biomass in fermentation process using support vector machine. Chinese J. Chem. Eng. **14**, 383–388 (2006)