Dynamic hybrid model for biopesticides production using *Bacillus thuringiensis* strains

Sergio Figueroa-Cardona [0009-0009-0745-2417], Carlos E. Robles-Rodríguez [0000-0002-2436-3653], Rim El-Jeni [0000-0002-3246-7890], Luc Fillaudeau [0000-0002-6389-5441], César A. Aceves-Lara*[0000-0001-6291-3655].

TBI, Université de Toulouse, CNRS, INRAE, INSA, Toulouse, France
*aceves@insa-toulouse.fr

Abstract. Commercial biopesticides are primarily based on the bacterium *Bacillus Thuringiensis* which produces insecticidal proteins. Dynamic models have been proposed to describe the production of these proteins, but their approximation can still be improved. Dynamic hybrid modelling combines kinetic models with data-driven algorithms which could be promising to improve the existing dynamic models. This work presents a dynamic hybrid model using Support Vector Machine (SVM) to predict the specific production of protein and spore by different strains of *B. thuringiensis*. The dynamic hybrid model was trained and validated with independent datasets of batch fermentations using the production rates of biomass, protein and spores computed with a kinetic model and the strain type as predictors. Additionally, Shapley values were calculated to determine the pertinence of each predictor. The Normalized root mean squared errors (NRMSE) revealed an improvement of the dynamic hybrid model of 12% for proteins and 7% for spores over the dynamic model. Although good, the model could be further improved by training the model with a larger quantity of data.

Keywords: *B.thuringiensis*, Support Vector Machine, hybrid modelling, optimization

1 Introduction

Bacillus thuringiensis is a facultative anaerobic Gram-positive sporulating bacterium, commonly used in the production of biopesticides [1]. Most commercial biopesticides are of microbial origin and are primarily based on this microorganism [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 this bacterium provide effective and environmentally benign control of several insects in agricultural, forestry, and disease-vector applications [3].

Its insecticidal effect is mainly due to the production of crystalline inclusions consisting of multiples insecticidal proteins, known as δ-endotoxins or Cry proteins, which are related to the spore formation of the bacteria [1]. *B. thuringiensis* must experience certain modifications of its physiology to produce these δ-endotoxins, which makes its culture a challenging labor. One possibility to optimize the production of these δ-endotoxins is to use a model-based approach in which several simulations of mathematical

models are assessed. However, only a few models are available that can correctly emulate the dynamics of the growth and production phases of *B. thuringiensis* culture [4].

Dynamic hybrid modelling integrates both data-driven and mechanistic approaches and could offer a cost-effective solution for modelling complex biochemical processes where the underlying mechanisms are not entirely known. Although more efficient in utilizing data compared to purely data-driven methods, the construction of dynamic hybrid models still require the collection of large sets of experimental data through time-consuming experiments for new processes [5].

Hybrid modelling has been applied in fermentation-related bioprocesses, to compare the performance of kinetic models and its hybrid counterpart, to describe the production of a compound of interest, as the case study of yeast astaxanthin production under uncertainty using Gaussian processes [6]. Other algorithms used before are artificial neural network (ANN) and response surface methodology (RSM), to optimize the biohydrogen production by dark fermentation [7].

In this paper, Support Vector Machine (SVM) was chosen based on its versality [8], the type of data available and its ability to work well with small datasets [9]. This method is described in section 2.

The aim of this paper is to propose a dynamic hybrid model to represent the dynamics of the fermentation of *B. thuringiensis* with a special focus on the products: the protein and the spore. The dynamic hybrid model is described in section 2, and the results obtained are detailed in section 3.

2 Materials and Methods

2.1 Organism and culture media

Three strains of *B. thuringiensis* were studied: BLB1, HD1 and Lip, a Lebanese strain [10]. 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 Experimental setup

The model was calibrated with datasets collected from batch experiments performed at 30 °C in a 3 L Biostat B plus fermenter (Sartorius; Germany) containing 1.8 L of the SSM medium. pH was regulated continuously at 6.8 using solutions of 1 M H_2SO_4 and 3 M NaOH. Dissolved oxygen was continuously monitored by an optical oxygen sensor and maintained at 25 % of pO2-saturation with constant aeration rate (VVM = 10 with Qair =0.18 min·L−1) and variable stirring (from 250 to 1200 rpm).

2.3 Experimental data

A dataset of 9 batches was obtained based on the experimental setup: 3 batches per tested strain. Each batch contains the measurements of 4 variables: biomass, substrate, protein and spore concentration. **Table 1** presents the list of the available data by batch and strain.

Table 1. Description of the experimental data.

Strain 1 (BLB1)		Strain 2 (HD1)		Strain 3 (Lip)	
Batch	Data per variable	Batch	Data per variable	Batch	Data per variable
1	10	4	10	7	26
2	10	5	7	8	8
3	11	6	26	9	10

As the hybrid model is based on a dynamic model, additional data was generated based on these experimental measurements to train the hybrid model. In this case, 100 data were generated per each training batch. The distinction between training and validation batches is detailed in section 3.1.

2.4 Dynamic model

The dynamic model proposed by Monroy et al. [11] has been used as basis of this study. The model represents the mass balances for *B. Thuringiensis* in a batch reactor as described in Eq. (1), (2), (3) and (4)

$$\frac{dX}{dt} = (\mu - k_d) \cdot X = r_X \tag{1}$$

$$\frac{dS}{dt} = -\frac{\mu \cdot X}{Y_{XS}} = r_S \tag{2}$$

$$\frac{dPro}{dt} = X \cdot k_{pro} = r_{pro} \tag{3}$$

$$\frac{dSpo}{dt} = X \cdot k_{spo} = r_{spo} \tag{4}$$

The variables described in these equations are the concentrations of biomass (X), substrate (S), protein (Pro) and spore (Spo), expressed in $g.L^{-1}$. The parameter Y_{XS} denotes the yield coefficient between biomass and substrate in $gX.gS^{-1}$. The biomass growth rate is represented by μ and its decay constant as k_d, both in h^{-1}. In this particular case, the Contois expression was used to calculate the biomass growth rate, as given in Eq. (5).

$$\mu = \frac{\mu_{max} \cdot S}{(K_c \cdot X) + S} \tag{5}$$

The term μ_{max} is the maximal growth rate and K_c is the Contois specific constant.

The parameter k_{pro} is the specific kinetic constant for protein (gPro.gX^{-1}.h^{-1}) and k_{spo} is the specific kinetic constant for spores (CFUx10^{-5}.gX^{-1}.h^{-1}). The mass balances in Eq. (3) and (4) correspond to the production rates for protein (r_{pro}) and spore (r_{spo}).

2.5 Dynamic Hybrid Model

In this work, protein and spores were modelled by a data-driven approach. Therefore, two new equations were included in the model as follows,

$$\frac{dPro}{dt} = r^*_{pro} \tag{6}$$

$$\frac{dSpo}{dt} = r^*_{spo} \tag{7}$$

where r^*_{pro} and r^*_{spo} are calculated as the derivative of a sigmoidal function that describes the proteins and spore concentrations, respectively.

These output variables were modelled using a Support Vector Machine (SVM) approach. SVM is a non-parametric and non-linear technique, which has attracted great attention in recent years due to its stability, robustness, and generality, especially for the cases involving high-dimensional regression or classification analysis [12].

In this work, SVM is used to predict accurately the desired output y through Eq. (8).

$$y = w^T \varphi(x) + b \tag{8}$$

where φ(x) is the nonlinear mapping of the input x into a high dimensional feature space. The determination of this equation is based on an optimization problem, which aims to find parameters that improve the efficiency of the y estimation [13]. This is achieved by using a Kernel function, which can be linear, quadratic, cubic and Gaussian, according to the mapping of the data.

SVM has been used to design soft sensors to produce protein on the same process by *B. thuringiensis* where variables as pO_2, agitation and strain number were categorized as the most important features/variables to describe protein production [14].

In this case, a model for r^*_{pro} was defined based on five predictors: *B. thuringiensis* strain type used in the experiment ($strain$), the biomass growth rate (μ) from Eq. (5), and the consumption/production rates. This last three were calculated from the dynamic model, Eq. (2) to (4) (r_S, r_{pro}, r_{spo}). The SVM model can be expressed as a function of the form:

$$r^*_{pro} = f(strain, \mu, r_S, r_{pro}, r_{spo}) \tag{9}$$

For the r^*_{spo} model, r^*_{pro} is included in the predictors to increase accuracy.

$$r^*_{spo} = f(strain, \mu, , r_S, r_{pro}, r_{spo}, r^*_{pro}) \tag{10}$$

The variable $strain$ was defined as an integer from 1 to 3, assigned as 1: BLB1, 2: HD1 and 3: Lip respectively.

Experimental data were filtered and smoothed so that the numerical derivative of protein and spore of the experimental data could have a similar behavior to the production rates calculated with the dynamic model. Using this filtered data, the model could be trained to follow the expected evolution of the products. For instance, to avoid decreases in product concentration and any experimental disturbances. The filtered data is presented in the Results section.

3 Results

3.1 Training tests

The model was trained using experimental data from nine batch essays described in section 2.3. Six out of these datasets were used for model training (two per each strain) and the remaining three datasets (Batch No. 3, 6 and 9) were used for model validation (one for each strain of *B. thuringiensis*). The batches for training were selected based on the values obtained for r^*_{pro} and r^*_{spo} in order to have the validation sets inside the range of the training sets.

SVM modelling was performed using the Regression Learner tool from Matlab2020a. The hybrid model was integrated and solved in Matlab2020a.

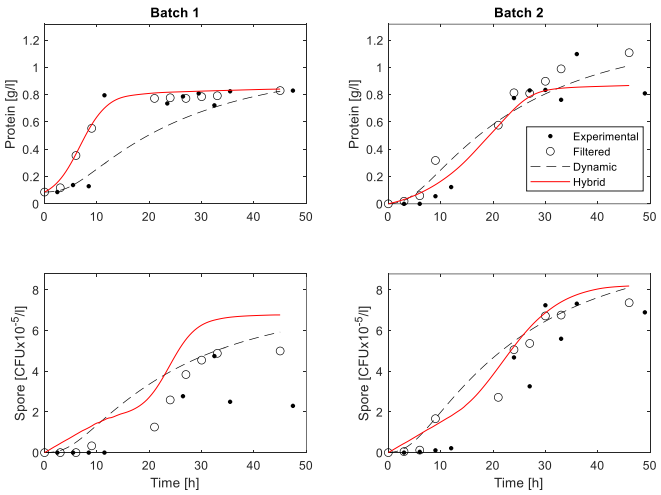

Fig. 1. Training tests for protein and spore concentration with strain 1.

The results presented in **Fig. 1** depict the performance of the dynamic and hybrid model for protein and spore production in comparison to the experimental and filtered data from 3 batches using strain 1. The dynamic hybrid model presents a similar behavior than the dynamic model for protein concentration, while there is an improvement in the spore concentration, where the hybrid model fits better the experimental data.

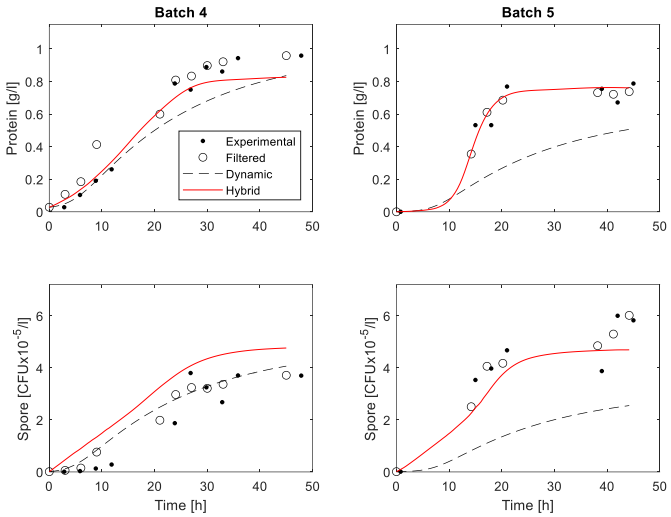

Fig. 2. Training tests for protein and spore concentration with strain 2.

Fig. 2 shows that the dynamic hybrid model is better at describing protein production than the dynamic model, but the opposite is observed for the spore in batch 4. These variations can be explained by the wide range of values that a product concentration can take at the same conditions for the same strain and the fact that there is a trade-off between the simulation of both batches.

Fig. 3 shows that both models follow the dynamics for proteins. The hybrid model fits better for spores when experimental data has higher values and starts in a low concentration point (**Fig.3 Batch 8**).

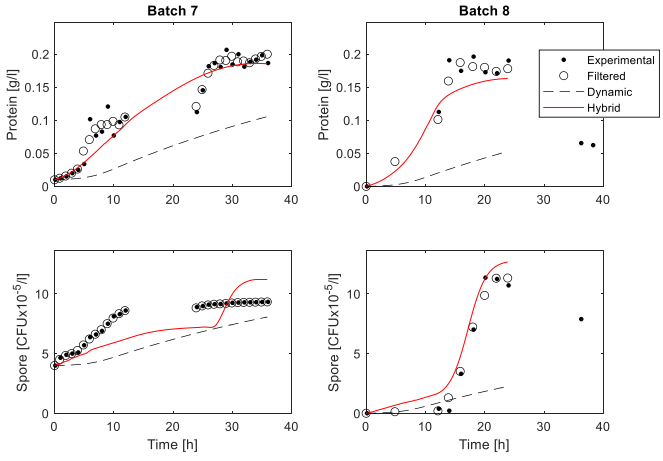

Fig. 3. Training tests for protein and spore concentration with Strain 3.

3.2 Validation tests

The batches with higher uncertainty on protein and spore concentrations were used for model validation. For this reason, significant differences are observed between models and data in **Fig.4.**

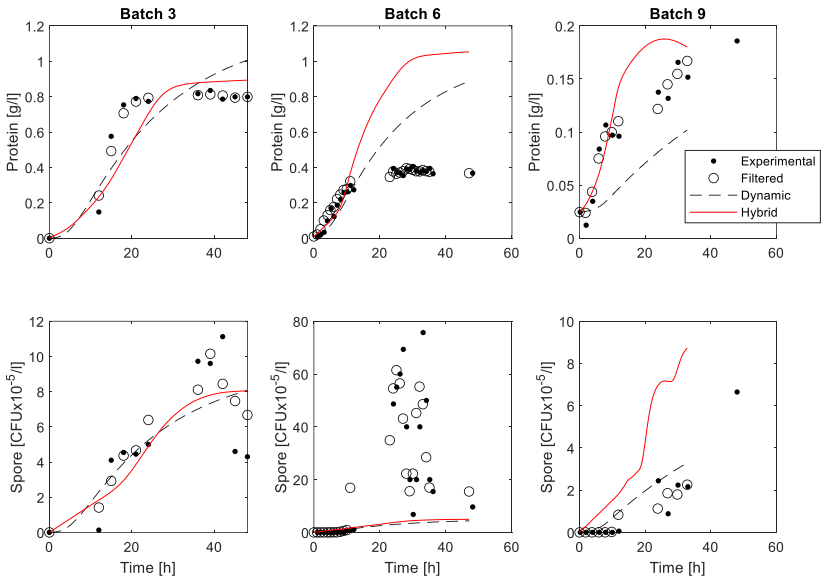

Fig. 4. Validation tests for Protein production. (Batch 3: strain 1, Batch 6: strain 2, Batch 9: strain 3).

There are some improvements in the prediction of protein production for the case of strain 1, depicted in **Fig. 4 Batch 3**. In the case of the spores, the differences are notorious, but the hybrid model is more likely to consider drastic changes in spore concentration because it takes information from its production rate, so it has an important advantage over the dynamic model.

The Normalized Root Mean Square Error (NRMSE) was calculated for each test to assess the overall performance of the models in both the training and validation stages. A distinction is made regarding the evaluation on the experimental and the filtered data. The values of NRMSE are presented in **Table 2.**

Table 2. NRSME comparison.

Variable	Data classification	Experimental data	Dynamic model	Hybrid model
Protein	Training	Original	0.4231	0.0691
		Filtered	0.4156	0.0540
	Validation	Original	0.4637	0.4049
		Filtered	0.4713	0.4144
Spore	Training	Original	0.4373	0.1143
		Filtered	0.4340	0.1065
	Validation	Original	0.3067	0.2879
		Filtered	0.3514	0.3262

The NRMSE shows that the dynamic hybrid model has better accuracy than the dynamic model in all tests, but more importantly, in the validation tests.

3.3 Shapley values

To consider the influence of each predictor in the models, a Shapley values evaluation was performed over both models. This indicator represents the deviation of the Shapley values, taking the mean as a reference. Therefore, the predictors with a higher deviation will be more determinants to the model.

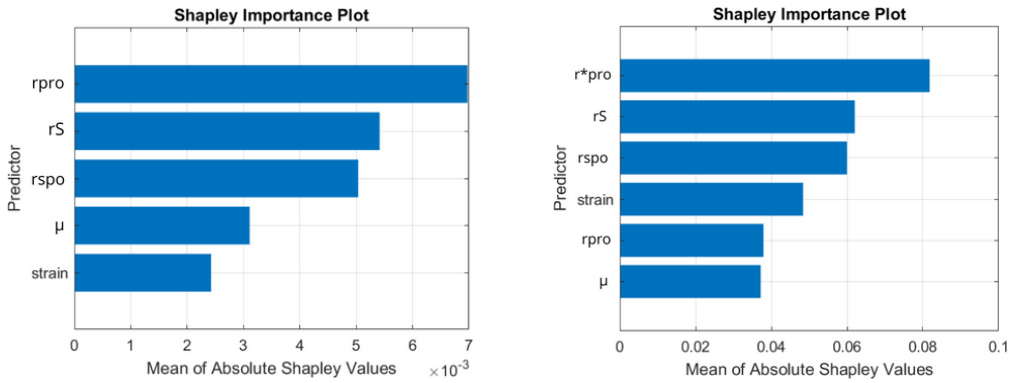

Fig. 5. Shapley importance plots for protein production rate (a) and spore production rate (b) model.

In **Fig. 5.a,** it is observed that the influence of the dynamic model predictors is very high, being the protein production rate the main predictor in the protein model, as expected. Nevertheless, in **Fig. 5.b** the output of the protein model has a significant impact over the accuracy of the spore rate model. The spore rate from the dynamic model is in third place, but its deviation score is not considerably high compared to the previous places. As the deviations from the mean value are not significant, no predictor can be neglected in the computing of both models.

4 Conclusion

A dynamic hybrid model was proposed to describe biopesticides production produced by three different strains of *B. Thuringiensis*. This model was able to fit protein and spore production. The dynamic hybrid model was compared against a dynamic kinetic model. The dynamic hybrid model has shown better performance than the kinetic dynamic model for the proteins production but for spores there is a gap in some data sets. The NRMSE values for the dynamic hybrid model present an improvement of 12% for proteins and 7% for spores over the dynamic model. The analysis of Shapley values shows all predictors are necessary for both models. This approach will be applied with different kinds of substrates as raw material in future works.

Acknowledgments. This work was supported by the 'Alternative Biopesticides For Safe Integrated Pest And Water Management Around Mediterranean' (SAFWA) project, which has received funding from the European Union's Horizon 2020 research and innovation program under the grant agreement No2022/Section 2.

References

1. E. Schnepf et al., 'Bacillus thuringiensis and Its Pesticidal Crystal Proteins', Microbiology and Molecular Biology Reviews, vol. 62, no. 3, pp. 775–806, Sep. 1998, doi: 10.1128/mmbr.62.3.775-806.1998.
2. W. Jallouli, F. Driss, L. Fillaudeau, and S. Rouis, 'Review on biopesticide production by Bacillus thuringiensis subsp. kurstaki since 1990: Focus on bioprocess parameters', Process Biochemistry, vol. 98, pp. 224–232, Nov. 2020, doi: 10.1016/j.procbio.2020.07.023.
3. G. E. Rowe and A. Margaritis, 'Bioprocess design and economic analysis for the commercial production of environmentally friendly bioinsecticides from Bacillus thuringiensis HD-1 kurstaki', Biotechnol Bioeng, vol. 86, no. 4, pp. 377–388, May 2004, doi: 10.1002/bit.20146.
4. A. K. Navarro-Mtz and F. Pérez-Guevara, 'Construction of a biodynamic model for Cry protein production studies', AMB Express, vol. 4, no. 1, p. 79, Nov. 2014, doi: 10.1186/s13568-014-0079-y.
5. S. Kay, H. Kay, A. W. Rogers, and D. Zhang, 'Integrating hybrid modelling and transfer learning for new bioprocess predictive modelling', in Computer Aided Chemical Engineering, vol. 52, A. C. Kokossis, M. C. Georgiadis, and E. Pistikopoulos, Eds., in 33 European Symposium on Computer Aided Process Engineering, vol. 52. , Elsevier, 2023, pp. 2595–2600. doi: 10.1016/B978-0-443-15274-0.50412-1.
6. F. Vega-Ramon, X. Zhu, T. R. Savage, P. Petsagkourakis, K. Jing, and D. Zhang, 'Kinetic and hybrid modeling for yeast astaxanthin production under uncertainty', Biotechnology and Bioengineering, vol. 118, no. 12, pp. 4854–4866, 2021, doi: 10.1002/bit.27950.
7. Y. Wang et al., 'Optimization of dark fermentation for biohydrogen production using a hybrid artificial neural network (ANN) and response surface methodology (RSM) approach', Environmental Progress & Sustainable Energy, vol. 40, no. 1, p. e13485, 2021, doi: 10.1002/ep.13485.

8. V. Vapnik, S. Golowich, and A. Smola, ‘Support Vector Method for Function Approximation, Regression Estimation and Signal Processing’, in Advances in Neural Information Processing Systems, MIT Press, 1996. Accessed: May 21, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1996/hash/4f284803bd0966cc24fa8683a34afc6e-Abstract.html
9. D.-C. Li and C.-W. Liu, ‘A class possibility based kernel to increase classification accuracy for small data sets using support vector machines’, Expert Systems with Applications, vol. 37, no. 4, pp. 3104–3110, Apr. 2010, doi: 10.1016/j.eswa.2009.09.019.
10. M. Khoury, H. Azzouz, A. Chavanieu, N. Abdelmalek, J. Chopineau, and M. Kallassy, ‘Isolation and characterization of a new Bacillus thuringiensis strain Lip harboring a new cry1Aa gene highly toxic to Ephestia kuehniella (Lepidoptera: Pyralidae) larvae’, Archives of microbiology, vol. 196, Apr. 2014, doi: 10.1007/s00203-014-0981-3.
11. T. S. Monroy et al., ‘Dynamic Model for Biomass and Proteins Production by Three Bacillus Thuringiensis ssp Kurstaki Strains’, Processes, vol. 9, no. 12, p. 2147, 2021, doi: 10.3390/pr9122147.
12. H. Hamedi, O. Mohammadzadeh, S. Rasouli, and S. Zendehboudi, ‘A critical review of biomass kinetics and membrane filtration models for membrane bioreactor systems’, Journal of Environmental Chemical Engineering, vol. 9, no. 6, p. 106406, Dec. 2021, doi: 10.1016/j.jece.2021.106406.
13. J. Wang, T. Yu, and C. Jin, ‘On-line Estimation of Biomass in Fermentation Process Using Support Vector Machine1’, Chinese Journal of Chemical Engineering, vol. 14, no. 3, pp. 383–388, Jun. 2006, doi: 10.1016/S1004-9541(06)60087-6.
14. C. E. Robles Rodriguez et al., ‘Soft-Sensors for Monitoring B. Thuringiensis Bioproduction’, in Distributed Computing and Artificial Intelligence, Volume 1: 18th International Conference, vol. 327, K. Matsui, S. Omatu, T. Yigitcanlar, and S. R. González, Eds., in Lecture Notes in Networks and Systems, vol. 327. , Cham: Springer International Publishing, 2022, pp. 129–136. doi: 10.1007/978-3-030-86261-9_13.