Domenico et al. (2023)

Thursday, January 25, 2024 10:59 AM

Article

A Deep Learning Approach to Optimize Recombinant Protein Production in *Escherichia coli* Fermentations

Domenico Bonanni [1], Mattia Litrico [2], Waqar Ahmed [2], Pietro Morerio [2], Tiziano Cazzorla [3], Elisa Spaccapaniccia [4], Franca Cattani [4], Marcello Allegretti [4], Andrea Rosario Beccari [1], Alessio Del Bue [2] and Franck Martin [4,*]

1 Dompé Farmaceutici SpA, EXSCALATE, Via Tommaso De Amicis, 95, 80131 Napoli, Italy; domenico.bonanni@univaq.it (D.B.); andrea.beccari@dompe.com (A.R.B.)
2 Pattern Analysis and Computer Vision, Fondazione Istituto Italiano di Tecnologia, Via Morego, 30, 16163 Genova, Italy; mattia.litrico@phd.unict.it (M.L.); waqar.ahmed@iit.it (W.A.); pietro.morerio@iit.it (P.M.); alessio.delbue@iit.it (A.D.B.)
3 M-Squared, Strada per Cernusco 1, 20060 Bussero, Italy; tiziano.cazzorla@gmail.com
4 Dompé Farmaceutici SpA, Via Campo di Pile, Nucleo Industriale Pile, 67100 L'Aquila, Italy; elisa.spaccapaniccia@dompe.com (E.S.); franca.cattani@dompe.com (F.C.); marcello.allegretti@dompe.com (M.A.)
* Correspondence: franck.martin@dompe.com

- **Introduction:**

Within this framework, we tracked the optical density at 600 nanometers (OD600nm), a classical fermentation parameter used as a proxy variable for produced biomass. From this absorbance, it is indeed possible to estimate the bacterial concentration in the culture medium. As all fermentations
have been carried out in the same culture medium using the same expression vector and the same Escherichia coli strain, it is reasonable to think that OD600nm is the most relevant proxy variable for the produced biomass, reflecting the production of inclusion bodies in which the recombinant protein accumulates

In this work, we developed a ML model based on LSTM networks and fed by ten culture critical process parameters (CPP) to accurately predict real-time
and final OD600nm values. Historical series for the evolution of such ten-dimensional state vectors were derived and used as inputs for the network, and the OD600nm values were used as labels. Furthermore, such online descriptors have been complemented by further global variables obtained off-line post-fermentation, such as recombinant protein dosage, induction time, and inclusion body weight. Those extra parameters have been used to confirm the different fermentations trends and select the best ones to train the system.

- **Materials and Methods:**

a. Fermentations

Summary of Fermentation Process:

1. Scale and Preparation:
   - Fermentations are conducted at a 1-liter scale, ideal for parallel cultures and sufficient downstream material production.
   - Overnight cultures in shaking flasks are used to inoculate fermenters with 20 mL of exponentially growing culture.
     (to inoculate the fermenters with 20 ml means that at the end of the overnight period when the bacteria are in a phase of rapid growth, known as the "exponential" or "log" phase, they take 20 milliliters (about 4 teaspoons) of this rapidly growing bacterial culture from the shaking flasks and add it to the larger fermentation vessels, which are set up for the main part of the experiment)
   NB: Inoculate" means to introduce microorganisms into a new environment where they can grow.

2. Strain and Vector:
   - Strain: Escherichia coli BL21 (DE3), chosen for its protein expression capabilities under the T7 promoter.
   - Vector: A kanamycin-selectable expression vector houses the recombinant protein gene, allowing for antibiotic selection.
     (A promoter is a DNA sequence that acts like a "start" signal for gene expression. The T7 promoter is recognized by a special enzyme (T7 RNA polymerase) that starts the process of reading the gene and making the protein. This setup ensures that once the promoter is activated, the E. coli will start producing the protein encoded by the recombinant gene.)
     (Expression vector is a piece of DNA (like a mini-chromosome) that carries the recombinant protein gene and the T7 promoter. It's designed to easily insert into bacteria and make them produce the desired protein).
     (Kanamycin-selectable expression vector means the vector also includes a gene that makes the E. coli resistant to kanamycin, an antibiotic)

3. Media Composition (two types of nutrient solutions were used) - "culture medium":
   - Original Terrific Broth (TB): Contains yeast extract, soy-peptone, potassium di-hydrogen orthophosphate, di-potassium hydrogen ortho-phosphate, and glycerol.
   - Fed-batch Medium: Modified with increased glycerol concentration (or substituting glycerol as carbon source with glucose) and reduced yeast extract, with kanamycin for selection.
     (In **batch fermentation** i.e. TB medium), all the nutrients are added at the beginning, and the culture grows without any further addition of nutrients or removal of culture volume until the process ends. The volume of the culture remains constant throughout the fermentation process, and there's no removal of culture volume in this method either.)
     (**Fed-batch fermentation** is a modified version of batch fermentation where additional nutrients are added ("fed") to the culture during the fermentation process without removing any culture volume. This allows for extended growth and protein production phases.)
4. Inoculum Development:
   Inoculate" means to introduce microorganisms into a new environment (Bioreactor) where they can grow.

Table: Inoculum Development Process

Parameter	Inoculum Preparation	Transfer to Bioreactors
Starting Material	25 µL from cell bank	-
Culture Medium	25 mL TB	700 mL TB
Antibiotic Concentration	50 µg/mL kanamycin	50 mg/L kanamycin
Container	125 mL baffled shake flask	Bioreactors
Incubation Temperature	28°C	-
Incubation Time	16 hours	-
Shaker Speed	180 rpm	-
Volume for Inoculation	-	20 mL overnight culture

5. Fermentation Setup (Bioreactor):
   - Utilizes eight 1L autoclavable stirred fermenters with controlled temperature, pH, and dissolved oxygen, managed by Lucullus software.

6. Fermentation Conditions:
   - Temperature: 30°C for batch phase, adjusted for fed-batch.
   - pH: Maintained at 7.0.
   - Agitation and Aeration: Managed to maintain 50% air saturation.

b. Inclusion Body Recovery:
   - Post-fermentation, biomass is centrifuged, and inclusion bodies are extracted and stored for further use.

(Inclusion Body Recovery involves isolating protein aggregates from bacteria after fermentation. The process includes cell harvesting, lysis to release inclusion bodies, washing to purify, solubilizing the proteins, refolding them into their functional forms, and final purification to obtain the desired p rotein for further use.)

Tables for Clarity:

Table 1: Fermentation Steps

	Step	Location	Process Description	Summary of Actions
Step 1	Fermentation in Shake Flask	Shake Flask	Initial growth of E. coli in a controlled small-scale environment.	Grew E. coli from a cell bank in 25 mL TB medium with 50 µg/mL kanamycin at 28°C for 16 hours, shaking at 180 rpm.
Step 2	Inoculation of Bioreactors	Bioreactor	Transferring the culture to a larger vessel to scale up production.	Transferred 20 mL of the overnight shake flask culture into each bioreactor containing 700 mL TB medium with 50 mg/L kanamycin.
Step 3	Fermentation Monitoring and Control	Bioreactor	Maintaining optimal growth conditions in the bioreactor.	Set fermenters at 30°C, controlled pH to 7.0, and adjusted oxygen levels with agitation and aeration.
Step 4	Fed-Batch Phase	Bioreactor	Adding nutrients in stages to sustain longer growth and production.	Implemented a nutrient feeding strategy in the bioreactor, starting at 0.3 mL/min, increasing to 0.9 mL/min for nutrient addition.
Step 5	Inclusion Body Recovery	-	Processing the culture to extract protein aggregates after fermentation.	Post-fermentation, centrifuged the culture to collect biomass and isolated inclusion bodies, stored at -70° C.

Table 2: Media Composition

Component	TB Medium (per L)	Fed-Batch Medium (per L)
Yeast Extract	24 g	300 g
Soy-Peptone	12 g	-
Potassium Di-Hydrogen Phosphate	4.8 g	-
Di-Potassium Hydrogen Ortho-Phosphate	2.2 g	-
Glycerol	5 g	300 g
Kanamycin	-	50 mg

Table 3: Potential Manipulated Parameters in Fermentation Runs

Parameter	Description
Temperature	The operating temperature, which could vary to optimize protein production.
Feeding Rate	The rate at which nutrients are added during the fed-batch phase.
Nutrient Composition	Changes in the composition of the fed-batch medium, such as glycerol or glucose concentrations.
Induction Strategy	The method used to induce protein expression, possibly involving changes in chemical inducers or their concentrations.
Agitation Speed	The speed of the stirrer in the bioreactor, affecting oxygen distribution and mixing.
Aeration Rate	The volume of air or oxygen supplied to the culture, influencing oxygen availability.
pH Control	Adjustments to the acidity or alkalinity of the medium, which can impact microbial growth and protein production.

Table 4: Fermentation Conditions Summary

Parameter	TB Batch Phase	Fed-Batch Phase
Temperature	30°C	Adjusted as needed
pH	7.0	7.0
Agitation	500-1500 rpm	Adjusted as needed
Aeration	0.5-1.5 L/min air	Up to 1 L/min O2
Antifoam	Automatic addition	-
Feeding Strategy	-	Near-exponential

Table 5: Fermentation Results

Run ID (selected Batches)	Final OD600nm	Biomass (g)	IB (g)	OD600nm/IB Ratio
8	63.4	66.6	13.5	4.7
11	55	54.4	12	4.6
12	62.7	59.3	12	5.2
16	56	51	10.8	5.2
22	51.1	53.6	10	5.1
23	50.5	58.4	12.4	4.1
24	52	53.9	10.1	5.1
25	50.3	302	NA	NA
26	55.6	301.1	NA	NA
27	44.6	66.3	NA	NA
28	32.4	69.3	NA	NA

- Final OD600nm: Optical Density at the end of the experiment. OD600nm stands for Optical Density measured at 600 nanometers, a wavelength in the visible light spectrum. It's a common measure used in microbiology to estimate the concentration of microorganisms (like bacteria or yeast) in a liquid culture.

- Biomass (g): refers to the total mass of microorganisms present in the culture. It's often expressed in grams (g).
- IB (g): stands for Inclusion Bodies, which are aggregates of proteins that accumulate inside bacterial cells, often as a result of overexpression of recombinant proteins in biotechnological applications.

Table 6: Inclusion Body Recovery Process

Step	Description
1. Centrifugation	Separate biomass at 8000 rpm
2. Resuspension	In 0.1 M Tris with 0.01 M EDTA, pH 8.0
3. Homogenization	800 ± 50 bars, 4 cycles
4. Washing	Centrifuge and wash twice with buffer
5. Storage	Store inclusion bodies at -70°C

- **Machine learning Pipeline:**

a. Data Preparation (data selection, pre-processing, normalization, and sequencing):

1. Parameter Selection:
   - Objective: Analyze Critical Process Parameters (CPPs) to select relevant ones for the algorithm.
   - Outcome: Table of selected CPPs with their nomenclature.

   Table: Selected Critical Process Parameters (CPPs)

Table: Selected Critical Process Parameters (CPPs)

CPP	Unit	Nomenclature
pH	unit	m_pH
Dissolved Oxygen	%	m_ls_opt_do
Temperature	°C	m_temp
Stirrer	rpm	m_stirrer
Pure Oxygen	L	dm_o2
Compressed Air	L	dm_air
Pump 1 (Base)	rpm	dm_spump1
Pump 2 (Acid)	rpm	dm_spump2
Pump 3 (Antifoam)	rpm	dm_spump3
Pump 4 (Feed)	rpm	dm_spump4
Induction	Binary	induction

2. Fermentation Selection (They selected some batches to be used for the ML):
   - Criteria: Quality, quantity, and as well as by analyzing whether the fermentation had a standard and canonical progress.
   - Selected Batches: 8, 11, 12, 16, 22, 23, 24, 25, 26, 27, 28.

3. Data Pre-processing Steps:
   - Cumulative to Non-Cumulative Data: - Convert cumulative data (e.g., spumps) to non-cumulative for better algorithm performance.
   - Data Normalization: - Utilize z-score normalization to ensure data distribution with mean 0 and standard deviation 1.
   - Data Sequencing: - Group data using a sliding window approach to format it for machine learning analysis.
   - OD600nm Interpolation: - Address temporal inconsistencies between CPP and OD600nm data by interpolating OD600nm values.

b. Deep Learning Model:

Fermentation processes are dynamic, with critical process parameters (CPPs) and optical density at 600 nm (OD600nm) values evolving over time. This temporal evolution is well captured by models designed for sequential data, like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which can uncover hidden correlations across different time points.

1. Model Design:
   - Initial Layer: A fully connected layer that maps input data into a high-dimensional latent space, preparing it for sequential processing.
   - Sequential Processing: Utilizes RNN/LSTM modules to handle data across time, taking into account the influence of past data points within a defined window of time.
   - Output Layer: Another fully connected layer that produces the final predictions for OD600nm values.

2. Application to Sequential Data:
   - The model processes sequences of CPPs and OD600nm values, reflecting the dynamic nature of fermentation.
   - Input data is structured as sequences with dimensions based on the window size and the number of features (CPPs).
   - The RNN/LSTM layers allow the model to consider historical data within each sequence, leading to more accurate predictions.

3. Hyperparameters:

The model's performance is fine-tuned using a grid search method to optimize hyperparameters such as input size, time sequence length, stride, latent space size, and the number of recurrent layers.

Table: Hyperparameters for Network Architecture

Hyperparameter	Value
Input size	11
Time sequence length	20
Stride	5
Latent space size	16
Number of recurrent layers	2

- **Experiments Overview**

The experiments were designed to validate the effectiveness of the described deep learning architecture in predicting fermentation outcomes. Key aspects of the experiments included data preparation, model training

and testing, and performance evaluation.

a. Setup:
   1. Data Pre-processing:
      - A subset of fermentation batches was selected based on data quality and consistency.
      - "Canonical" settings were defined, excluding preliminary and non-standard fermentations.
      - Batches with process failures or absent recombinant protein production were also excluded.
      - Data points, especially from batch ends lacking periodic OD600nm readings, were trimmed.

   2. Training Parameters:
      - Learning Rate: 0.001
      - Batch Size: 256
      - Optimizer: Stochastic Gradient Descent (SGD)

b. Evaluation Criteria:
   1. Leave-One-Out Cross-Validation (LOOCV):
      - The LOOCV method was used, treating each fermentation as a separate test case with the remaining data as the training set.
      - This approach helps evaluate the model's generalization ability across different fermentation scenarios and provides a reliable performance estimate.

c. Evaluation Metrics:
   1. Root Mean Squared Error (RMSE):
      - Measures the model's accuracy by calculating the square root of the average squared differences between predicted and actual OD600nm values.

   2. Relative Error on Final Yield (REFY):
      - Focuses on the absolute error at the fermentation's last timestamp, emphasizing the prediction's accuracy at the most critical point.

- **Results**

The experiments compared the performance of LSTM and RNN networks in predicting fermentation outcomes, particularly focusing on root mean squared error (RMSE) and relative error on final yield (REFY). Both networks showed promising results, with RNN slightly outperforming LSTM in general.

a. Key Observations:
   i. Initial Yield Estimation: Both models exhibited higher RMSE at the start of the fermentation process, likely due to limited historical data available for initial predictions.

   ii. Generalization Over Time: Model accuracy improved over successive timestamps, indicating better generalization with more context.

   iii. Final Yield Estimation: Some batches (notably 8, 12, and 25) showed a plateau in final yield estimation accuracy, with batches 8 and 25 having higher REFY values.

   iv. Overall Performance: While certain batches showed higher errors, the majority of the batches (16, 22, 23, 24, 26, 27, and 28) demonstrated accurate yield predictions, underscoring the potential of machine learning in fermentation process modeling.

b. Quantitative Results:

The table below encapsulates the performance metrics of LSTM and RNN networks across various batches, highlighting RMSE, REFY, and the comparison with ground truth final yields.

Table: LSTM and RNN Network Results

Batch	RMSE (LSTM)	RMSE (RNN)	REFY (LSTM)	REFY (RNN)	Final Yield (Ground Truth)
8	3.26	3.50	13.68%	15.74%	62.83
11	7.08	7.22	7.85%	1.24%	54.84
12	9.18	8.20	18.15%	15.03%	61.69
16	3.53	3.16	3.07%	1.62%	54.72
22	4.08	2.89	2.39%	3.16%	50.83
23	1.63	2.34	0.47%	2.41%	50.31
24	2.47	2.52	1.87%	0.18%	51.51
25	3.85	3.84	20.10%	11.58%	44.72
26	2.13	2.17	10.56%	6.07%	43.57
27	3.01	4.36	4.31%	12.63%	44.75
28	4.90	3.85	0.31%	9.61%	44.58
Avg. (Std.)	**4.10 (2.24)**	**4.00 (1.97)**	**7.52 (6.80)%**	**7.21 (5.60)%**	**51.30 (6.76)**

*Cases where the prediction exceeded the actual final yield are marked with an asterisk.

- **Discussion**

a. *Production Insights:*

- Fermentation Setup: Parallel fermentations were conducted in 1L fermenters, facilitated by a software interface for real-time parameter monitoring. This setup proved beneficial for simultaneous experimentation and immediate assessment of parameter impacts.

- Volume Efficiency: The chosen volume strikes a balance between operational convenience and the generation of sufficient recombinant protein quantities for downstream processing and preclinical studies, with yields ranging from 10 to 13.5 grams of inclusion bodies per batch.

- Culture Conditions: High yields were attributed to a nutrient-rich culture medium and a strategic approach to experimental design, contributing to efficient E. coli growth.

- Time Efficiency: Fermentations were notably brief, not exceeding 10 hours, indicating a streamlined process conducive to scaling and industrial application.

- Adaptability: The system maintained predictive accuracy even with modifications to the expression system, underscoring the robustness of the ML model across different conditions and expression forms of the target protein.

b. *ML Prediction Performance:*

- Challenges in Unusual Conditions: Batches with atypical final yields (notably batches 8 and 12) presented challenges for the model, leading to higher REFY values due to these outliers.

- Model Behavior: Initial estimations were less accurate, especially in batches with distinctive yield trends, impacting RMSE. Specific CPP trends, like unusual pump activity, also influenced prediction accuracy.
- Generalization and Improvement: The results suggest a need for broader training data to enhance model generalization, particu larly for conditions underrepresented in the current dataset.