Periscope Opt. ML Research

Monday, January 22, 2024 6:51 AM

ELSEVIER

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

PERISCOPE-Opt: Machine learning-based prediction of optimal fermentation conditions and yields of recombinant periplasmic protein expressed in *Escherichia coli*

Kulandai Arockia Rajesh Packiam [a,1], Chien Wei Ooi [a,b], Fuyi Li [c], Shutao Mei [d], Beng Ti Tey [a,b], Huey Fang Ong [e], Jiangning Song [d,f,*], Ramakrishnan Nagasundara Ramanan [a,*]

[a] Chemical Engineering Discipline, School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Bandar Sunway, Malaysia
[b] Advanced Engineering Platform, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Bandar Sunway, Selangor, Malaysia
[c] Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Victoria 3010, Australia
[d] Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Victoria 3800, Australia
[e] School of Information Technology, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Bandar Sunway, Malaysia
[f] Monash Centre for Data Science, Faculty of Information Technoology, Monash University, Victoria 3800, Australia

PERISCOPE-Opt: Machine learning-based prediction of optimal fermentation conditions and yields of recombinant periplasmic protein expressed in Escherichia coli

Introduction:

The research paper presents a machine learning (ML) model to optimize the fermentation process for recombinant protein production (RPP) in E. coli. The study integrates fermentation process conditions with amino acid sequences to predict maximum protein yields and corresponding conditions. The model is designed to reduce the need for extensive trial-and-error experiments in determining optimal fermentation conditions for RPP.

Key Terminology:

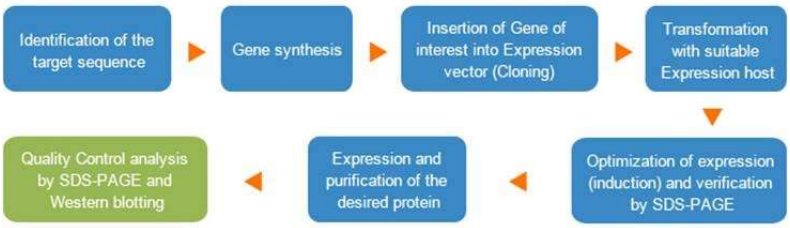
1. Recombinant Proteins
   This refers to a protein that is artificially made through the process of genetic recombination. This involves manipulating DNA sequences to produce a specific protein.
2. Escherichia coli (E. coli)
   This type of bacteria is commonly found in the intestines of humans and animals. They are versatile, fast-growing organisms with the ability to rapidly multiply, making it ideal for efficiently producing proteins.
3. Recombinant Protein Production
   Recombinant protein production using E. coli involves inserting a gene coding for the desired protein into the bacteria. These genetically modified E. coli are then cultivated in controlled conditions, where factors like temperature and nutrients are optimized for growth and protein production. The protein is either naturally produced or induced in the bacteria. After sufficient growth, the protein is harvested and purified from the E. coli cells or the surrounding medium
4. Genetic Recombination
   Genetic recombination is a natural process in which genetic material is exchanged between two different DNA molecules or rearranged within a single molecule
5. Recombinant DNA (rDNA):
   rDNA, or recombinant DNA, is a form of DNA that has been created artificially by combining genetic material from multiple sources. In the context of producing recombinant proteins using E. coli, rDNA is the genetically engineered DNA that is introduced into the E. coli bacteria (the host organism). This rDNA contains the gene that codes for the protein you want to produce. Once the rDNA is inside the E. coli cells, these cells use the genetic instructions in the rDNA to produce the recombinant protein.
6. Plasmid:
   A small DNA molecule within a cell that is physically separated from chromosomal DNA and can replicate independently. Used as a vector to transfer rDNA into E. coli bacteria. NB: the gene of interest (rDNA) is inserted into a plasmid. The plasmid then serves as a vehicle to introduce this recombinant DNA into a host organism, such as E. coli, for protein production
7. Transformation:
   The process of introducing foreign DNA into a bacterial cell, such as E. coli.
8. Expression Vector:
   The plasmid or vector that has been engineered or chosen to introduce and express the recombinant DNA (rDNA) in the E. coli cells.
9. Host Cell:

The cell into which the recombinant DNA is introduced, which in this case is E. coli.

10. Protein Expression:
The process by which proteins are synthesized, modified, and regulated in living organisms. In recombinant DNA technology, it refers to the production of the protein coded by the introduced gene.

11. Culture Medium:
The nutrient solution used to grow bacteria like E. coli in the lab. Nutrient solutions typically contain sources of carbon (such as glucose), nitrogen, minerals, vitamins, and other essential nutrients.

Procedure for Recombinant protein production/expression/purification



Literature Review:

ML-based tools primarily focus on features with respect to amino-acid-sequence, ruling out the influence of fermentation process conditions.

- Amino Acid Sequence:

This is like a recipe for the protein, directly determines the structure and function of the protein being produced. In recombinant protein production, the amino acid sequence refers to the specific order of amino acids that make up the protein being produced. This sequence is determined by the gene inserted into the host organism (like E. coli). The gene's DNA sequence encodes the information needed to assemble amino acids in the correct order to form the desired protein. The accuracy of this amino acid sequence is crucial, as it determines the structure and function of the resulting recombinant protein.

Several research that focuses on Amino acid sequence including the prediction of protein solubility, protein folding rates, and protein. This is because the amino acid sequence primarily influences the protein solubility and folding rates, which in turn affects many facets of RPP.

a. Predicting of protein solubility: Models such as PROSO II, ccSOL Omics, Protein–Sol, DeepSol, PaRSnIP, SoDoPE, and SolTranNet
b. Prediction of protein folding rates: Models such as K-Fold, Pred-PFR, PRORATE, and SeqRate. Additionally,
c. Prediction of expression yields yields in the cytoplasm and periplasm of E. coli: Models such as ESPRESSO and Periscope

- Fermentation Process Conditions:

These include various environmental and procedural factors involved in growing microorganisms (like E. coli) in a controlled setting to produce proteins. Factors like temperature, pH, nutrient availability, and oxygen levels are all part of the fermentation process conditions. Optimizing these conditions is crucial for maximizing the yield and quality of the produced protein.

Main Research:

The present study aimed to derive a global ML-based model capable of predicting the optimal protein yield and fermentation process conditions for a target recombinant protein to be expressed in the periplasmic space of E. coli.

a. **Classify Protein expression levels**: Two sets of XGBoost (XGB) classifiers were employed in the first stage to classify the target protein into high (>50 mg/L), medium (between 0.5 and 50 mg/L) or low (<0.5 mg/L) expression levels.

b. **Prediction of Optimal Fermentation conditions (to quantify the protein yield for each class)**: Three sets of regression models based on support vector machine (SVM) or random forest (RF) to quantify the protein yield with respect to each class.

Dataset:

461 datasets (collection of data) for 84 proteins and 103 SP-protein combinations. The sequence redundancy of the 103 SP-protein combinations was removed using the CD-HIT suite at 90% of the sequence similarity threshold. 11,985 features were initially extracted by considering all important.

a. the amino acid sequence of the recombinant protein expressed in the periplasm of E. coli,
b. the corresponding protein expression yield measurable in milligrams per litre (mg/L),
c. the parameters of the fermentation process conditions.

Criteria to consider before including the dataset in the research:

a. E. coli strain as the host and lac promoters for expression;
b. Heterologous protein expression in the periplasm;
c. SP at the N-terminus;
d. batch fermentation at shake flask scale, and finally,
e. neither involving any genetic modification of the host strain nor including any co-expression vectors.

Conditions to consider for the **independent test dataset**.

a. data from 10 proteins in which their amino acid sequences are completely unknown/unseen to the model, and
b. data from 18 proteins in which their amino acid sequences are known but the optimal fermentation conditions are unseen to the model.

Feature Extraction:

A total of 11,985 initial features were extracted and further classified into four major categories of features. Since RPP is primarily regulated by the amino acid sequence of a protein, we focused on the features that can be extracted directly from the amino acid sequences, as well as features like physicochemical and structural properties derived indirectly using these amino acid sequences.

a. Feature Category 1 (FC1) → **149 features**:
   Constitutes general features such as (refer to Tables S3-1 to S3-3 for details).:
   a. length of the protein,
   b. occurrences of each type of the 20 amino acid residues,
   c. maximum number of consecutive identical amino acid residues.
   d. occurrences and the maximum number of consecutive amino acid residues with similar physico-chemical properties
   e. molecular weight,
   f. isoelectric point (pI),
   g. net charge,
   h. solubility,
   i. protein folding rate,
   j. helix/sheet propensity

b. Feature Category 2 (FC2) → **800 features**:
   Feature Category 2 (FC2) deals with the occurrences of dipeptides in protein sequences. Given the high dimensionality of data, dipeptides (combinations of two amino acids) are grouped separately under this category. FC2 is specifically focused on capturing the frequency and pattern of these dipeptide occurrences in the protein sequences. By analyzing these features, the study aims to understand how the combinations of amino acids in dipeptides influence protein production in Escherichia coli. (see Table S3-4)

c. Feature Category 3 (FC3) → **11,026 features**:
   Feature Category 3 (FC3) in the study is derived from the interactions between the features in Feature Category 1 (FC1). FC3 consists of 11,026 interactive features, which result from considering all possible interactions between the two types of features in FC1 (occurrences and occurrence frequencies of amino acids and their standardized values by protein length).

d. Feature Category 4 (FC4) → **10 features**:
   Feature Category 4 (FC4) in the study focuses on features derived from the fermentation process conditions, encompassing 10 different features. These include factors like (see Table S3-5):

1. cell density (measured in optical density OD600 nm) – normalized separately,
2. inducer concentration at the time of protein expression induction (measured in molarity (M), indicating the concentration of the inducer (like isopropyl b-D-1-thiogalactopyranoside (IPTG) or lactose) added to the culture.
3. post-induction temperature (°Cor °F)
4. Post-induction time (hours (h) or minutes (min))
5. interactions between these elements.

All features except those in Feature Category 3 (FC3) and interactive features from Feature Category 4 (FC4) are normalized using the min-max normalization to avoid biases. This includes the features in other categories such as FC1 and FC2.

Feature Selection:

The important features were selected for both classification and regression tasks based on a stepwise feature selection strategy. The stepwise feature selection strategy is applied to both the classification and regression tasks to reduce the bias caused by the high dimensionality.

a. **Step 1** (method = CfsSubsetEval):
   Features were selected from both FC2 and FC3 using the **CfsSubsetEval** method along with the search method, Best First
b. **Step 2** (method used = CfsSubsetEval, and Greedy Stepwise) **:**
   The numbers of the selected FC3 features were further reduced using **CfsSubsetEval** and **Greedy Stepwise** methods. 'generate ranking" option set to True, and no of features to be retained == 10
c. **Step 3** (method used = ClassifierSubsetEval)**:**
   key optimal features were selected from all the features from FC1 and FC4 along with previously selected features from FC2 (Step 1) and FC3 (Step 2) using ClassifierSubsetEval

NB:

CfsSubsetEval: This method evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. It prefers subsets of features that are highly correlated with the class while having low intercorrelation.

Greedy stepwise method: This is a search method that either starts with no/all attributes and searches through them by adding or removing one at a time. It's a form of forward or backward selection that makes the decision based on the evaluation method used.

ClassifierSubsetEval: This method evaluates attribute subsets by using a classifier. The worth of a subset of attributes is directly related to the classification accuracy of the selected classifier. This approach uses the performance of a specific classifier to determine the usefulness of feature subsets.

Model Training and Evaluation:

- Three ML algorithms, SVM, XGB, and RF, were employed in both classification and regression tasks.
- Fine-tuning of the hyperparameters was not performed because the default parameter values gave a good result.
- The following performance metrics were used:
  Classification:
  a. Accuracy.
  b. error rate.
  c. precision.
  d. recall.
  e. F-measure.
  f. Mathew correlation coefficient (MCC).
  g. area under the curve (AUC).
  Regression:
  h. Pearson correlation coefficient (PCC).

i. mean absolute error (MAE).
j. root mean squared error (RMSE).

- **Cross-validation**: Leave one out cross-validation (LOOCV)
- **Handling Imbalance**: Two strategies were used to address data imbalance.
  Firstly, two binary classifiers were utilized in a way that the first classifier separates the data into the majority class (medium expression levels) and a combined group of minority classes (High and Low expression levels). The second classifier then categorizes the combined minority classes into their respective classes.
  Secondly, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to generate additional dummy data in the minority classes during both classification tasks.

- **Model Architecture**:

d. **Classification of Protein Expression Levels**: This stage utilized two sets of XGBoost classifiers. The first classifier categorized proteins into medium-level expression or non-medium-level (high and low) expression. The second classifier further differentiated non-medium-level expressions into high and low expressions.
e. **Prediction of Optimal Fermentation Conditions**: This involved three sets of regression models (Support Vector Machine and Random Forest) to quantify protein yield for each classified expression level. Depending on the expression level determined in the first stage (low, medium, or high), a corresponding regression model (SVR for low, RFR for medium or high) was applied to predict the expression yield.

- **Additional Training:**
  The study created 180 additional combinations of fermentation process conditions, each varying in optical density at 600nm (OD) (0.4, 0.7, 1.0), IPTG concentration (0.1, 0.5, 1.0 mM), temperature (20, 25, 30, 37 ^{O}C), and time (4, 8, 12, 16, 24h). These combinations provide a diverse range of scenarios for analyzing the impact of different fermentation conditions on protein expression. The results allowed for the identification of the top ten combinations for optimal expression yields of recombinant proteins under specific fermentation conditions. The "top 10 combinations" refers to the ten best sets of fermentation conditions identified by the model, which resulted in the highest periplasmic expression yields of recombinant proteins.

Features Selected:

Having performed feature selection using the steps above, the selected features for both classification and regression tasks consist of features mostly from FC1 and FC4, a few within FC3, and none from FC2.

The features selected from the study can be grouped as follows:

a. Amino Acid Sequences:
   - Occurrences and occurrence frequencies of amino acids such as glutamic acid (Occ_E, OF_E), valine (Occ_V), and methionine (OF_M), valine (Occ_V), sulfur (OF_S), phenylalanine (OF_F) and methionine (OF_M)
   - Maximum consecutive sequences like alanine (MNC_A, OF_MNC_A) and cysteine (OF_MNC_C).

b. Physico-Chemical Properties:
   - Occurrence frequency of specific residue types like aromatic (OF_Aromatic), aliphatic (OF_Aliphatic), and hydrophilic residues (OF_Hphil_ESG, OF_Hphil_KD).

c. Structural Properties:
   - Features like the expected number of amino acids in transmembrane helices (Expno_AA_TM), ratio of helix-to-sheet propensity (Helix_to_Sheet_PHD), coil propensity (Coil_PHD), and solubility score (Pred_Sol).

d. Fermentation Process Conditions:
   - For classification, temperature and OD*Time are key.
   - For regression, nearly all process-condition features are significant.

These features, capturing various aspects of protein characteristics and fermentation conditions, are crucial for the model's predictive accuracy

Assessment of Feature Importance:

The assessment of feature importance in the study was conducted by training models while systematically removing one feature at a time.

Key findings include:

1. For XGB–Classifier 1, features like OF_Aromatic, OF_MNC_C, and temperature were crucial. Their removal significantly decreased model accuracy and MCC.
2. In XGB–Classifier 2, removing Occ_V or Helix_to_Sheet_PHD drastically affected performance, as indicated by substantial drops in accuracy and MCC.
3. For regression tasks, temperature was vital. Its absence increased MAEs and RMSEs noticeably.
4. Interaction features like OD_Temp, IPTG_Temp, and Temp_Time were significant for some regression models, with their removal impacting performance.
5. Other features like OF_Hphil_ESG and specific FC3 interactive features also showed notable impact on model performance when removed.

Model Performance:

1. Performance on training and/or Cross-validation Dataset.

   The performance of the model in classification and regression tasks showed notable results:

   - In classification tasks, XGBoost (XGB) outperformed both Random Forest (RF) and Support Vector Machine (SVM).
   - Average accuracies for the two XGB classifiers were above 75% in both internal and independent tests.
   - Performance measures like precision, recall, F-measure, and AUC were all above 0.75, with MCC around 0.5.

     For regression tasks:

   - The highest Pearson Correlation Coefficients (PCCs) were observed in SVR-Low, RFR-Medium, and RFR-High models.
   - These regression models also showed higher PCCs in independent testing.

- Mean Absolute Errors (MAEs) and Root Mean Square Errors (RMSEs) were low for SVR-Low but increased for RFR-Medium and RFR-High, in line with the expression yield ranges of each class.

2. Performance on Independent test validation
   Independent test validation refers to evaluating the model's performance using data that was not part of the model's training set.
   - Classification Accuracy:
     Achieved 75% accuracy on 28 unseen instances, with 21 instances correctly classified, including six unseen proteins.
   - Expression Yield Prediction:
     High Pearson Correlation Coefficient (PCC) of 0.91 for correctly classified instances. The PCC remained high (0.80) even with some misclassifications.
   - Error Metrics:
     Low Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for correctly classified instances.

Discussion:

The discussion section highlights the development of a robust machine learning tool for predicting maximal protein expression yields and optimal fermentation conditions. Key features from amino acid sequences and fermentation processes were combined for a precise model. The study underscores the importance of certain amino acids and process conditions in recombinant protein expression in E. coli. Temperature, valine occurrence, and specific amino acid interactions were notably significant. The model's predictions closely matched experimental data, demonstrating its reliability. Additionally, the study points out the challenges in model development due to data variability and suggests future improvements incorporating gene factors and structural properties of proteins.

Problems:

1. Sequence Data was not used
2. They did not tune the models hyperparameters.

Difference between Plasmids and Expression vectors:

Plasmids (a small, circular DNA molecule independent of a cell's chromosomal DNA, used in labs to clone and manipulate genes) serve as carriers for foreign genes, while expression vectors are specialized plasmids tailored to ensure the efficient production of proteins from these genes.

During fermentation, E. coli cells containing the expression vector are cultivated under conditions that promote the growth of the bacteria and the expression of the recombinant protein. The expression vector includes elements like a strong promoter recognized by E. coli's transcription machinery, which initiates the production of mRNA from the inserted gene, and a ribosome binding site, ensuring the mRNA is efficiently translated into the desired protein. This system allows for the large-scale production of recombinant proteins, which can be harvested and purified from the fermentation broth for various applications in medicine, research, and industry.