Optimizing Recombinant Protein Titer in Escherichia Coli Fermentations Using Advanced Filtering Techniques and Interpretable Machine Learning.

Abstract ID: 7371

Leave space for author names and affiliations to ensure that the limit of 6 pages is not exceeded. Author names and affiliations to be added before final submission, but not during initial blind review.

Abstract

Keywords
Biomanufacturing processes, advanced filtering techniques, protein production, predictive analytics

1. Literature Review

In efforts to optimize recombinant protein titer in Escherichia coli fermentations, substantial research has been devoted to enhancing expression systems and understanding the underlying biological mechanisms that influence protein yield and solubility. Recombinant protein production in E. coli presents both opportunities and challenges due to the bacteria's rapid growth rate, cost-effectiveness, and well-understood genetics. However, issues such as protein misfolding, inclusion body formation, and inconsistent expression levels continue to impede optimal protein yield. Advanced computational approaches and machine learning techniques have increasingly been employed to tackle these challenges, aiming to predict and enhance protein expression outcomes more effectively.

Recent studies have leveraged various machine learning models to predict protein solubility and expression levels, which are crucial for optimizing titer. For instance, the use of support vector machines (SVM) and regression models has proven effective in classifying and predicting expression levels based on sequence and structural features of proteins (Packiam et al. 2022). Techniques such as these allow for the classification of proteins into high, medium, or low expression levels, providing valuable insights for biotechnological applications. The integration of interpretable machine learning models offers the potential to uncover the complex relationships between genetic sequences and protein expression outcomes, facilitating the rational design of E. coli expression systems.

The adoption of sophisticated filtering techniques such as Kalman Filters, Particle Filters, Spatial Filter Alignment, and Madgwick Filters is set to revolutionize the predictive accuracy of models in recombinant protein production. Each of these filters offers unique advantages in processing and interpreting the dynamic data associated with genetic sequences and protein expression levels. Kalman Filters and Particle Filters are particularly adept at dealing with noisy data, enabling more precise predictions of protein titer by continuously updating estimates based on observational measurements. Spatial Filter Alignment can enhance the analysis of spatial data correlations, crucial for understanding the geometric and structural constraints in protein folding. Meanwhile, Madgwick Filters, typically used in orientation estimation, could be innovatively applied to stabilize signal processing in dynamic, multi-dimensional biological systems. Integrating these advanced filtering techniques with machine learning frameworks is expected to significantly refine our ability to manipulate and optimize Escherichia coli expression systems for maximum protein yield.

Table 1. Recent Studies on Optimizing Recombinant Protein Production in Escherichia coli.

Authors (Year)	**Study Objective**	**Filtering Techniques**	**Machine Learning Methods.**	**Data Characteristics**	**Performance Metrics**	**Critical Findings**	**Limitations**
Baako et al. (2024)	Review ML and DL applications in CHO cell bioprocessing and other bioprocesses.	--	Machine learning, deep learning, multivariate statistical analysis	CHO cell biomanufacturing data, multivariate data from various bioprocessing fields	Not Specified	Highlights the integration of ML and DL in biopharmaceutical manufacturing, enhancing productivity and efficiency.	Specific limitations not discussed in the paper.
Bonanni et al. (2023)	Optimize recombinant protein production in Escherichia coli fermentations	--	RNN, LSTM	Time series of CPPs from E. coli fermentation process, OD600nm values. No mention of missing data types	RMSE, REFY	Demonstrated the potential of ML models (RNN and LSTM) for real-time yield prediction in fermentation processes. Proved that ML can optimize process control by predicting fermentation outcomes based on historical CPP data.	Limited validation of models outside the training data scope.
Kager et al. (2022)	Develop a direct control strategy for recombinant protein production in E. coli fed-batch processes.	--	Nonlinear feedback linearization	E. coli recombinant protein production data from fed-batch processes	Model fits, control errors	Demonstrated potential for stable, prolonged production with nonlinear control methods.	High model errors at elevated productivities; potential inaccuracies due to metabolic stress impacts not fully overcome.
Packiam et al. (2022)	Develop a model to predict optimal yields and fermentation conditions for recombinant protein expressed in E. coli.	--	XGBoost, SVM, Random Forest	Data combined from the bioinformatics tool Periscope, literature, and in-house experiments; 84 protein-types, 11,985 features used.	Accuracy: 75%, Pearson correlation coefficient: 0.91	Developed a two-stage ML model that integrates amino acid sequences with fermentation process conditions to predict optimal yields and conditions efficiently.	Scarcity of experimental data, missing information on process conditions, irretrievable amino acid sequences, variability in fermentation protocols, and specific host strain impacts.
Gundinger et al. (2022)	Investigate the phoA expression system for producing	--	--	Comparison of phoA (pAT) and T7lac (pET)	Fab productivity, physiologic	The study found that phosphate limitation enhances phoA-based gene expression,	Detailed performance analysis under varying

	recombinant model antigen-binding fragment (Fab) in E. coli periplasm, focusing on phosphate (PO4)-sensitive conditions and their impact on strain physiology and Fab productivity.			expression systems under varying PO4 conditions, including non-limiting and limiting scenarios, and phosphate starvation.	al changes in E. coli, efficiency of extracellular PO4 detection methods.	although leaky expression occurred even under non-limiting conditions. PO4 limitation eventually led to physiological strain changes, resulting in metabolic breakdown during PO4 starvation. Recommendations for process optimization with phoA systems were provided. The pAT system was shown to have advantages over the T7lac system under comparable conditions.	cultivation conditions, especially different phosphate concentrations, is lacking. Additionally, the paper did not address the direct impact of these conditions on the quality and functionality of the expressed Fab.
Sha et al. (2018)	Review mechanistic modeling applications in CHO cell culture.	--	Mechanistic models (stoichiometric, kinetic)	Review of literature on stoichiometric and kinetic models.	Not specified	Demonstrates how mechanistic models can improve process understanding, optimization, and control in bioprocessing.	Requires extensive parameterization; lacks general applicability across different cell lines and processes.
Khurana et al. (2018)	Develop DeepSol, a deep learning-based model for predicting protein solubility from sequence data.	--	Convolutional Neural Networks (CNNs)	Protein sequences	Accuracy, MCC, selectivity, sensitivity for soluble and insoluble proteins	DeepSol outperformed existing methods in accuracy and other metrics, enhancing protein solubility prediction.	Not specified, but implicit limitations include the dependency on sequence data quality and feature extraction.
Chang et al. (2016)	To quantitatively predict soluble protein expression in the periplasm of Escherichia coli.	--	SVM, SVR	Protein sequences, expression levels	Accuracy: 78%, PCC: 0.77	Periscope predictor successfully categorizes and quantifies soluble protein expression levels. Di-peptide composition crucial for predictions.	Data limited to specific conditions and sequences.
Liu et al. (2015)	Review strategies in metabolic engineering to improve recombinant protein	--	Not applicable (Review article)	Recombinant protein production in E. coli	Not specified	Improved protein production via workhorse selection, stress factor application, and carbon flux regulation.	Challenges like metabolic burden, physiological deterioration, and by-product formation persist.

	production in E. coli.						
Habibi et al. (2015)	Develop a model to predict recombinant protein overexpression levels in Escherichia coli based on various biological inputs.	--	Random Forest	Data from gene sequences, vectors, and hosts; includes a small, imbalanced dataset with missing data.	Accuracy: 80%	Developed RPOLP model that effectively classifies overexpression levels as low, medium, or high with promising accuracy, using gene sequence features.	Small dataset size, missing values, lack of detailed inter-feature relationships
Papaneophytou et al (2014)	Maximize recombinant protein expression in Escherichia coli using statistical approaches for optimization.	--	--	Review of recombinant protein expression challenges in E. coli	Not specified	Highlighted the importance of statistical designs over traditional one-factor-at-a-time methods to optimize protein solubility and expression. Discussed the influence of multiple factors on protein solubility and purity and advocated for the use of statistical methods throughout all stages of protein production and crystallization for improving outcomes.	Specific statistical methods not detailed; lack of consensus approach for protein expression.
Chou (2007)	Enhance recombinant protein production in E. coli through engineered cell physiology	--	--	Review of E. coli genetic, biochemical, and metabolic strategies	Not specified	Reviewed genetic and metabolic engineering strategies for improving E. coli physiology and recombinant protein productivity. Discussed limitations of gene expression levels and physiological stress.	Limited to single-gene manipulations; lacking in-depth mechanistic understanding; lacks real-time physiological monitoring techniques.
Wanner (1996)	Study phosphate-regulated gene expression in E. coli	--	Not applicable	Analysis of Pho regulon gene structure and function.	Not specified	Detailed the role of two-component regulatory systems in phosphate limitation responses in E. coli. Explored PhoR and PhoB roles in transcription regulation.	The study is primarily descriptive, lacking experimental validation of proposed regulatory mechanisms.

.

References

Baako, T.M.D., Kulkarni, S.K., McClendon, J.L., Harcum, S.W. and Gilmore, J., 2024. Machine Learning and Deep Learning Strategies for Chinese Hamster Ovary Cell Bioprocess Optimization. *Fermentation*, *10*(5), p.234.

Bonanni, D., Litrico, M., Ahmed, W., Morerio, P., Cazzorla, T., Spaccapaniccia, E., Cattani, F., Allegretti, M., Beccari, A.R., Del Bue, A. and Martin, F., 2023. A Deep Learning Approach to Optimize Recombinant Protein Production in Escherichia coli Fermentations. *Fermentation*, *9*(6), p.503.

Chang, C.C.H., Li, C., Webb, G.I., Tey, B., Song, J. and Ramanan, R.N., 2016. Periscope: quantitative prediction of soluble protein expression in the periplasm of Escherichia coli. *Scientific reports*, *6*(1), p.21844.

Chou, C.P., 2007. Engineering cell physiology to enhance recombinant protein production in Escherichia coli. *Applied microbiology and biotechnology*, *76*(3), pp.521-532.

Gundinger, T., Kittler, S., Kubicek, S., Kopp, J. and Spadiut, O., 2022. Recombinant protein production in E. coli using the phoA expression system. *Fermentation*, *8*(4), p.181.

Habibi, N., Norouzi, A., Hashim, S.Z.M., Shamsir, M.S. and Samian, R., 2015. Prediction of recombinant protein overexpression in Escherichia coli using a machine learning based model (RPOLP). *Computers in biology and medicine*, *66*, pp.330-336.

Kager, J., Bartlechner, J., Herwig, C. and Jakubek, S., 2022. Direct control of recombinant protein production rates in E. coli fed-batch processes by nonlinear feedback linearization. *Chemical Engineering Research and Design*, *182*, pp.290-304.

Khurana, S., Rawi, R., Kunji, K., Chuang, G.Y., Bensmail, H. and Mall, R., 2018. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, *34*(15), pp.2605-2613.

Liu, M., Feng, X., Ding, Y., Zhao, G., Liu, H. and Xian, M., 2015. Metabolic engineering of Escherichia coli to improve recombinant protein production. *Applied microbiology and biotechnology*, *99*, pp.10367-10377.

Packiam, K.A.R., Ooi, C.W., Li, F., Mei, S., Tey, B.T., Ong, H.F., Song, J. and Ramanan, R.N., 2022. PERISCOPE-Opt: Machine learning-based prediction of optimal fermentation conditions and yields of recombinant periplasmic protein expressed in Escherichia coli. *Computational and Structural Biotechnology Journal*, *20*, pp.2909-2920.

Papaneophytou, C.P. and Kontopidis, G., 2014. Statistical approaches to maximize recombinant protein expression in Escherichia coli: a general review. *Protein expression and purification*, *94*, pp.22-32.

Sha, S., Huang, Z., Wang, Z. and Yoon, S., 2018. Mechanistic modeling and applications for CHO cell culture development and production. *Current opinion in chemical engineering*, *22*, pp.54-61.

Wanner, B.L., 1996. Signal transduction in the control of phosphate-regulated genes of Escherichia coli. *Kidney international*, *49*(4), pp.964-967.