Deep Learning Models for Predicting Microbial Protein Expression

Contact
Dr Gizem Buldum, gizem.buldum@newcastle.ac.uk

Research project

Project Summary

This project aims to develop data-efficient, explainable deep learning (DL) models to predict and optimize microbial protein expression. The aim is to address the high costs of data acquisition and the limitations of existing models in interpretability and generalization. The project will focus on designing interpretable models that require fewer training samples to be accurate and reliable, thus making deep learning more accessible for microbial strain engineering. The end goal is to create a streamlined computational platform for microbial engineering applications in biotechnology.

Background and Motivation

Synthetic biology and microbial engineering rely on designing microbial strains to produce high-value proteins. Current advancements in high-throughput DNA synthesis and sequencing have enabled the collection of large datasets linking DNA sequence to protein expression. Although existing deep learning approaches have achieved high predictive accuracy, they demand extensive, costly training datasets, limiting their accessibility. Additionally, these models act as "black boxes," making it difficult for scientists to understand the underlying biological insights guiding the predictions. By combining data-efficient deep learning techniques and Explainable AI, this project aims to address the current challenges.

Applicant skills/background

This project requires strong computer science skills. Experience in applying deep learning models to biological data is preferred. Professionalism in collaborative work, planning, and strong interest in innovation are essential. A master's degree in bioinformatics, machine learning or a related field will be valued.

References

1. Nikolados, EM., Wongprommoon, A., Aodha, O.M. *et al.* Accuracy and data efficiency in deep learning models of protein expression. *Nat Commun* 13, 7755 (2022). https://doi.org/10.1038/s41467-022-34902-5
2. Avsec, Ž., Agarwal, V., Visentin, D. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* 18, 1196–1203 (2021). https://doi.org/10.1038/s41592-021-01252-x
3. Pierre-Aurélien Gilliot, Thomas E Gorochowski, Transfer learning for cross-context prediction of protein expression from 5'UTR sequence. *Nucleic Acids Research* 52, e58 (2024). https://doi.org/10.1093/nar/gkae491