[BIOINFO 2023]

Sequence-based Prediction of Bacterial Essential Genes using Protein Language Models

Seong-Bo Heo[1, 2], Kyungmin Park[1, 2], Dae-Hee LEE[1, 2], Seong Keun KIM[2], Jonghyeok Shin[2],

Seung-Goo LEE[1, 3], Haseong Kim[1, 2, *]

[1]*Department of Biosystems and Bioengineering, Korea National University of Science and Technology*

[2]*Synthetic Biology and Bioengineering Research Center, Korea Research Institute of Bioscience and Biotechnology*

[3]*Synthetic Biology and Bioengineering Institute, Korea Research Institute of Bioscience and Biotechnology*

**Corresponding author: haseong@kribb.re.kr*

This study presents deep learning approach for prediction of essential genes which are necessary for the survival and growth of a strain. Unlike existing models, we developed deep learning model with protein sequences as the sole input. We introduced a classification model using basic Convolutional Neural Network (CNN) structure. Also, we used feature extraction technique with pre-trained large models (ProtBERT / ProtT5) based on Transformer architecture which is latest paradigm in sequence analysis. As a result, the CNN model showed achieving almost the same prediction performance compared to the large models. In particular, using mean of confidence of the 3 models as a classification criterion could significantly improve the performance. We anticipate that the proposed approach will contribute to identifying the biological properties for essential and non-essential gene, and accelerating research requiring essential gene identification.