[BIOINFO 2024]

Codon optimization using Transformer and contrastive learning to enhance protein expression

Juseong Kim[1], Jeongmu Kim[1], Seungyun Shin[1], and Giltae Song[1,2,3,*]

[1]*Division of Arificial Intelligence, Department of Information Convergence Engineering, Pusan National University*

[2]*School of Computer Science and Engineering, Pusan National University*

[3]*Center for Artificial Intelligence Research, Pusan National University*

**Corresponding author: gsong@pusan.ac.kr*

Codon optimization is a vital technique in molecular biology and genetic engineering, aimed at enhancing recombinant protein expression. By modifying gene sequences to match the host organism's preferred codon usage, researchers can improve protein production efficiency and ensure smooth biological function. Traditional methods, while effective, are resource-intensive and time-consuming, making them less practical for large-scale or high-throughput applications. Recent advances in computational biology, particularly machine learning, have introduced new techniques to improve traditional codon optimization. However, deep learning approaches face challenges like overfitting to original sequences, which can hinder protein expression improvements. We introduce a hybrid deep learning framework that combines convolutional layers and transformer encoders to capture both local and global features essential for codon optimization. A key innovation is the use of a contrastive loss function between amino acid and codon embedding vectors, allowing the model to precisely map relationships and effectively identify subtle differences in codon sequences that impact protein expression. Additionally, by using pre-optimized sequences from three established tools, we created a robust training set, enabling the model to recognize more general patterns in codon usage. Empirical results showed that our optimized sequences exhibited higher protein expression levels than wildtype and sequences from traditional codon optimization tools, achieving higher Codon Adaptation Index (CAI) and GC content scores along with lower Minimum Free Energy (MFE) scores. In protein expression experiments, these sequences also demonstrated superior performance, resulting in significantly enhanced protein yield and efficiency. Although this study focuses on Homo sapiens, the model's adaptable design suggests potential applications across various species, making it a versatile tool for codon optimization.