Synthetic DNA Engineering With ICOR: Improving Codon Optimization With Recurrent Neural Networks Towards Efficient, Low-Cost, High-Efficacy Recombinant Vaccine and Pharmaceutical Manufacturing

Jain, Rishab (School: Westview High School)

In protein sequences—as there are 61 sense codons but only 20 standard amino acids—most amino acids are encoded by more than one codon. Although such synonymous codons do not alter the encoded amino acid sequence, their selection dramatically affects the expression of the resulting protein. Today, many recombinant vaccines struggle with efficacy due to low expression efficiency. Codon optimization of synthetic DNA sequences is paramount for improving heterologous expression. However, industry-standard codon optimization techniques based on biological indexes result in an imbalanced tRNA pool and metabolic stress imposed on the cell, leading to cell toxicity and reduced expression. In this research, a novel recurrent-neural-network (RNN) based codon optimization tool is developed on a genomic dataset of Escherichia coli, a popular cell factory. Over 7,000 non-redundant, high-expression, robust E. coli genes are used for deep learning. The custom bidirectional long short-term memory-based architecture, allows for the sequential context of E. coli codon usage to be learned. ICOR is evaluated on 1,481 E. coli genes and a benchmark set of 40 DNA sequences whose heterologous expression has been previously studied. ICOR's performance across codon adaptation index, codon frequency distribution, GC-content, negative repeat elements, and negative cis-regulatory elements is compared to that of five industry techniques. The results indicate that ICOR's statistically significant improvements on metrics yield a 236% improvement in real-world expression. This research demonstrates that sequential context achieved via RNN yields codon selection that is more similar to host genomes, therefore improving heterologous expression towards efficient production of recombinant vaccines.

Awards Won:
First Award of $5,000
Regeneron Young Scientist Award