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

ENBM074

Biomedical Engineering

Rishab Kumar Jain

Westview High School, Portland, OR

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.

1. In this research project, the student directly handled, manipulated, or interacted with (check ALL that apply):

human participants	potentially hazardous biological agents		
vertebrate animals	microorganisms	rDNA	tissue

2. I/we worked or used equipment in a regulated research institution or industrial setting (Form 1C):	YES	✘ NO
3. This project is a continuation of previous research (Form 7):	✘ YES	NO
4. My display board includes non-published photographs/visual depictions of humans (other than myself):	YES	✘ NO
5. This abstract describes only procedures performed by me/us, reflects my/our own independent research, and represents one year's work only:	✘ YES	NO
6. I/we hereby certify that the abstract and responses to the above statements are correct and properly reflect my/our own work.	✘ YES	NO

The stamp or embossed seal attests that this project is in compliance with all federal and state laws and regulations and that all appropriate reviews and approvals have been obtained including the final clearance by the Scientific Review Committee.

Official Regeneron ISEF 2022 Abstract and Certification *4/29/2022 5:18:47 PM*