Research Week 2021

Codonify: A Recurrent-Neural-Network based Codon Optimization Tool to Improve Protein Expression

Rishab Jain
rishab.won@gmail.com
Westview High School

Keywords

codon optimization, deep learning, synthetic biology, heterologous expression

Abstract

Designing synthetic genes for heterologous expression is a keystone of synthetic biology. 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, they are not redundant. By using certain codons over others, gene expression can be improved by up to 1000 times. Industry-standard codon optimization techniques based on biological indexes replace synonymous codons with the most abundant codon found in the host organism's genome. However, this technique may result in an imbalanced tRNA pool, metabolic stress, and translational error which lead to greater cell toxicity and reduced protein expression. In this research, recurrent neural networks are used to accurately capture sequential and contextual patterns. By predicting synonymous codons based on the sequential information of the host organism, protein expression can be increased while preventing translational error and plasmid toxicity. Theoretically, deep learning should yield better codon selection because it understands sequential and contextual information of the host. The model uses a bidirectional long short-term memory-based architecture, allowing for the host genome to be taken into context. The Codon Adaptation Index (CAI) was used to measure synonymous codon usage. When tested on eGFP and FALVAC-1, the model yielded a 0% mutation rate and improved CAI from 0.72 and 0.67 to 0.91 and 0.91 respectively. On a broad test dataset of 8,000 sequences, Codonify optimized CAI by 22% which correlates with an average 236% increase in expression. This research provides evidence that sequential context may yield codon selection that is more similar to the host genome, therefore increasing protein expression and the production of recombinant vaccines.