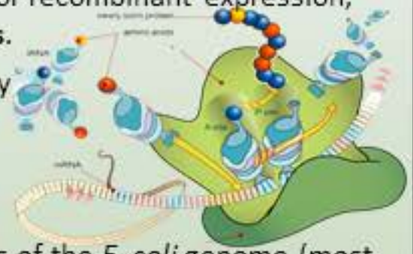
Synthetic DNA Engineering with ICOR: Improving Codon Optimization with Recurrent Neural Networks Towards Efficient, Low-Cost, High-Efficacy Recombinant Vaccine and Pharmaceutical Manufacturing

ENBM074

Rishab Jain, Westview High, Portland, Oregon, United States

Engineering Problem & Objectives

Today, recombinantly-produced pharmaceuticals ($11 billion market) have immense potential in combatting disease. When designing synthetic DNA/plasmids for recombinant expression, **codon optimization** can improve protein output up to **1000 times**.



Engineering Problem: There is a need for improving the efficiency of recombinant vaccine manufacturing, however, current approaches to codon optimization introduce cellular toxicity and metabolic stress by neglecting the context of codon usage.

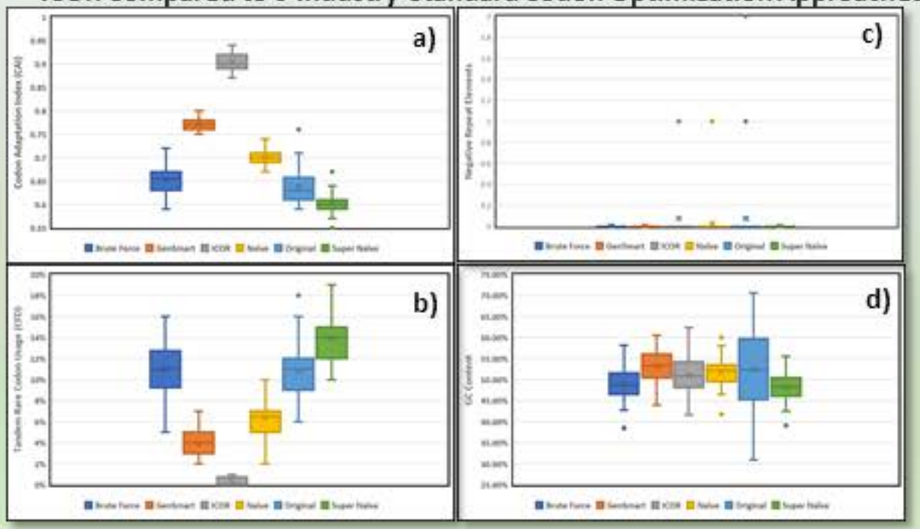
Objective: Develop a codon optimization tool that learns patterns of the *E. coli* genome (most popular cell in recombinant expression) to grasp evolutionary-instilled, optimal codon usage.

Engineering Criteria & Constraints: ***In what ways can I measure the success of my model?***

- ❑ Robustness: build model on 4.2K+ genes to be representative of the *E. coli* host genome.
- ❑ Functionality: learn deep patterns of codon usage over industry's "codon frequency/count."
- ❑ Specification: handle unique IUPAC sequence codes & have <25 hour runtime for 100 genes.

Data Analysis & Results

ICOR Compared to 5 Industry-Standard Codon Optimization Approaches on 5 Benchmark Metrics



a) Significantly improves **codon adaptation index** compared to all other codon optimization techniques (**p<0.0001**).
b) Significantly improves mean CFD/tandem rare codon usage (**p<0.0001**).
c) Insignificant difference (**p=0.1826**) in number of negative repeat & cis-regulatory elements.
d) Maintains **ideal GC content** range for ALL sequences.

Project Design & Methodologies

- Experimented with thousands of **hyperparameters** and variables to engineer the best solution.
- Made engineering **trade-offs** to overcome roadblocks (i.e. model complexity vs. batch size).
- Designed robust testing procedure to compare ICOR to alternative optimization approaches.

Curation

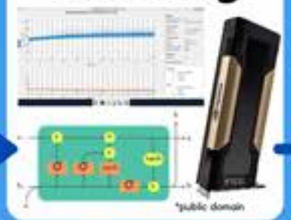
NCBI

- Retrieve genomes from NCBI
- Split into batches
- Prune hypothetical codons
- Prune redundant codons
- Merge using custom macros

Encoding

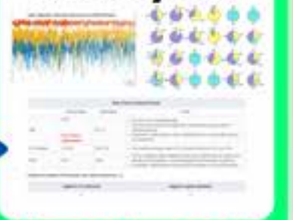
- Import FASTA data
- Natural language processing techniques (i.e. OneHot)
- Sortish sampling and codon adaptation index ranking.

Learning



- Create RNN architecture
- Hyperparameter tuning
- Engineering trade-offs
- HPC training & validation
- Combat over/under-fitting

Analysis



- Measure performance on 5 benchmark metrics
- Compare to industry-standard optimization approaches
- mRNA secondary structure analysis & simulations

Interpretation & Conclusions

Designed novel natural-language-processing and recurrent-neural-network based approach to **learn sequential information** in order to predict optimal synonymous codons.

Benchmark study reveals ICOR has a **236%** & **28.4%** increase in expression compared to original genes and GenScript's GenSmart (industry-leading tool), respectively.

ICOR is integrable into **the synthetic plasmid design pipeline** via the Windows executable application and Python command-line tool.

Research applications in improving efficient **medicine** & **drug production** for recombinant DNA technology.

Identify Synthetic Gene Sequences
exons in FASTA format

Example Targets:
- FALVAC-1
- PTP4A3
- hPDF

↓

Sequence Codon Optimization
with pre-trained ICOR deep learning model

↓

Construct Optimized Sequence Into Expression Vector