Continuation/Research Progression Projects Form (7)

Required for projects that are a continuation/progression in the same field of study as a previous project. This form must be accompanied by the previous year's abstract and Research Plan/Project Summary.

Student's Name(s) Rishab Jain

To be completed by Student Researcher: List all components of the current project that make it new and different from previous research. The information must be on the form; use an additional form for previous year and earlier projects.

Components	Current Research Project	Previous Research Project: Year: 2021
1. Title	Synthetic DNA Engineering with ICOR: Improving Codon Optimization with Recurrent Neural Networks Towards Efficient, Low-Cost, High-Efficacy Recombinant Vaccine and Pharmaceutical Manufacturing	CODONIFY: A Recurrent-Neural-Network-based Codon Optimization Tool to Improve Protein Expression
2. Change in goal/ purpose/objec- tive	Preserve codon usage bias and patterns of the host genome through better understanding of sequential and contextual information via deep learning. Engineering objective: build a model based on high-expression genes only, with the biological basis being these are better optimized.	Preserve important rare codons through the better understanding of sequential & contextual information of the host genome obtained via deep learning.
3. Changes in methodology	- Utilized 7,000 high-expression genes reduced by order of entropy and CAI. High-expression genes only are hypothesized to improve model's codon optimization. - Tested BERT and autoencoder natural language models. - Utilized 4 BiLSTM blocks, multiparameter encoding. - Curated 40 benchmark gene sequences from past studies on codon optimization. Tested model against these benchmark sequences in addition to E. coli gene test set for validation. - Visualize and identify secondary structural elements.	- Utilized all 42,000 genes - Utilized simple RNN architecture. - Tested BiLSTM architecture only. - Test set of E. coli genes only. - Conduct all analysis on the primary structure of the protein only.
4. Variable studied	- Added metrics for CAI, CFD, GC Content, free energy, negative repeat elements, negative CIS elements. - Hyperparameters during model building were expanded to include regularization techniques, more NLP embeddings. - Study NLFT (non-linear-fisher-transform) based on amino acid physicochemical properties.	- Only variable studied for validation was the codon adaptation index (CAI). - Model was only tested with one-hot-encoding and basic hyperparameters such as minibatch size, initial learn rate.
5. Additional changes	- Compared tool to industry-standard tools such as GenScript's GenSmart and 4 other codon optimization tools. - Used ONNX framework to convert model into Pythonic environment, improved model is open-access on the web. - Tool is being actively used, tested through licensing agreement with biotech companies.	- Compared tool to original sequences only. - Model was restricted to MATLAB environment and executable only. - Tool was not used by any real customers, only was theoretically validated.

Attached are:

☑ Abstract and Research Plan/Project Summary, Year 2021

I hereby certify that the above information is correct and that the current year Abstract & Certification and project display board properly reflect work done only in the current year.

Rishab Jain	Rishab Jain	1/31/2022
Student's Printed Name(s)	Signature	Date of Signature (mm/dd/yy)