TECHNISCHE UNIVERSITÄT DARMSTADT

Thesis (B.Sc. / M.Sc.)
Decoding the Genetic Code: Large-Scale NLP for Codon Optimization and Enhanced Protein Synthesis.

Department 18
Electrical Engineering and Information Technology
Self-Organizing Systems Lab

Prof. Dr. Heinz Koeppl
Head of lab

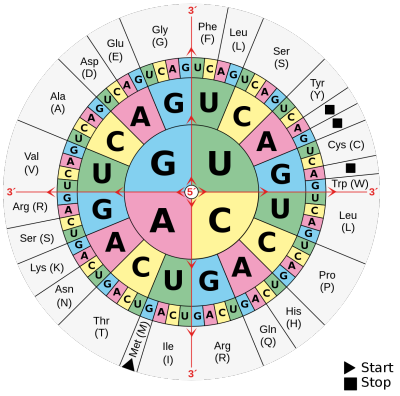
Philipp Froehlich
Project supervisor

S3|06 206
Merckstrasse 25
64283 Darmstadt

philipp.froehlich@tu-darmstadt.de
https://www.bcs.tu-darmstadt.de

March 24, 2025

Background: Microorganism-based protein production is a vital part of industrial biotechnology, spanning from bioethanol-producing enzymes to therapeutic antibodies. A central challenge is the efficient synthesis of heterologous (i.e., not native to the producing organism) proteins. One method to enhance production is codon optimization, where the DNA sequence is strategically modified to match the host organism's preferred sequence patterns, without changing the resulting protein. Decoding intricate patterns to achieve a desired output is analogous to natural language processing (NLP) methods. This topic presents an intersection where information processing techniques meet and address biological complexities.



Codon table. It emerges that 18 of 20 amino acids are encoded by multiple synonymous codons, making the genetic code redundant.

Objective: The aim of this project is to leverage deep learning, specifically using a Large Language Model, to identify patterns in homologous protein gene sequences that indicate high expressibility. Using these identified patterns, we want to predict the producibility of heterologous proteins from their DNA sequences and validate these predictions experimentally. Additionally, we are laying the groundwork for a new codon optimization framework developed collaboratively by our interdisciplinary team, though its specific design and implementation remain flexible at this stage.

Research Opportunities: We've recently developed a Large Language Model that operates on amino acid sequences for predicting protein synthesis capability. This model is currently under rigorous evaluation and experimental testing. There are two possibilities for a thesis, one with a computational focus and one with a biological focus.

The Self-Organizing-Systems Lab provides an interdisciplinary team environment. We are actively working towards a publication, offering students a chance to be part of this academic journey. Our approach is to involve students deeply in our research, encouraging them to introduce and realize their own innovative ideas. Joining our team means gaining insights into cutting-edge topics such as deep learning, NLP, and synthetic biology.

For further information, please contact Philipp Froehlich.