Hessisches Kompetenzzentrum
für Hochleistungsrechnen

Utilizing Mamba for Microorganism Protein Synthesis Prediction

Project Manager
Julie Weiß

Principal Investigator
Prof. Dr. Heinz Koeppl

Project Term
2024 - 2024

Clusters
Lichtenberg II Cluster Darmstadt

Software
Python

Additional Software
CUDA

Institute
Department of Electrical Engineering and Information Technology,
Department of Computer Science

University
Technische Universität Darmstadt

Introduction

This work aimed to apply deep learning, specifically Mamba, to the problem of codon optimisation in order to be able to efficiently synthesise heterologous (i.e. not native to the producing organism) proteins. Codon optimisation is the process of adapting the nucleotide sequence of the protein to the codon usage bias (which codon the organism uses more often to encode an amino acid) of the host organism. In this work, hyperparameter tuning was done for the Mamba model and the fine-tuned model was compared to a local attention Transformer model that was trained in the context of an already completed master thesis. Although Mamba is an efficient deep learning model, it nevertheless requires a lot of compute power, making the use of the Lichtenberg cluster indispensable. The questions to be answered were the following. How well does the Mamba model solve the task of codon optimisation and capture codon usage bias? And how does Mamba perform compared to the Transformer model in terms of accuracy and runtime?

Methods

Mamba is a sequence model based on state space models and promises to handle long sequences more efficiently than the prevalent Transformer architecture. Unlike Transformers, which are quadratic in complexity due to the use of the attention mechanism, Mamba scales linearly in sequence length and is very efficient thanks to its hardware-optimised parallel algorithm. Training was done for Escherichia coli,

Corynebacterium glutamicum and Bacillus subtilis and for each organism, the training set consisted of gene sequences downloaded from GenBank, the National Institutes of Health genetic sequence database.

Results

In summary, it was found, that Mamba is well suited for the task and outperforms the Transformer model of similar size, although it needs to be mentioned that the Transformer model was not hyperparameter tuned. Hyperparameter tuning for the Mamba model was a very heuristic and not extensive process. As expected, the hyperparameter model size had a big impact on model performance, the other hyperparameters that were tuned, for example batch size or weight decay rate, influenced the accuracy relatively little. It can be said that the Mamba model performs very well on the task of codon optimisation with an unbiased accuracy of 93.34% with the fine-tuned model on the test set after ten epochs. It outperforms the best model of the previous master thesis and is much faster since it is trained on GPU.

Discussion

For this task, Mamba seems to fulfil its claim of dealing well on long sequences. Unfortunately, the question regarding runtime could not be answered conclusively because the previous model did not train on GPU unlike the Mamba model. For future work, the optimised nucleotide sequences could be experimentally validated to see how well Mambas output sequences are synthesised in practice. Some of the previous approaches had the problem that they neglected low-frequency codons, which are important for protein folding. Since this was not studied for the Mamba model and despite its good accuracy, this could also be the case. It would also be useful to conduct hyperparameter tuning for the Transformer model and train it on a GPU for a better comparison to the Mamba model.

Last Update: 2025-02-07 12:25