TECHNISCHE UNIVERSITÄT DARMSTADT

Thesis Topic (B.Sc / M.Sc)

Deep Learning of Next-Generation-Sequencing Biological Data

The goal of this project is to set up a deep neural network in Pytorch to predict the efficiency of biological components. The main source of data is given in [1]. In the course of this project, additional sources of data might be used to improve the accuracy of the model.

The main data consists of 244000 DNA sequences and their corresponding protein production rate. A feature vector can be derived from each DNA sequence using bioinformatics tools. The tentative analysis in [1] concluded that biophysical features, such as the minimum free energy and secondary structures, can explain only approximately 50% of the observed variance. Further features are introduced in [2]. Both papers utilized simple models such as linear regression, random forests and support vector machines.

This project aims to use a more complex model and apply methods as in [3] to discover more features. The model should finally be used to design new sequences with improved production rate, which may be experimentally verified.

Python knowledge is required. Experience with neural networks is recommended.

For further information, contact Felix Reinhardt (felix.reinhardt@bcs.tu-darmstadt.de).

Department 18
Electrical Engineering and Information Technology

Institute for Telecommunications

Bioinspired Communication Systems

Prof. Dr. Heinz Koeppl
Head of lab

Felix Reinhardt
felix.reinhardt@bcs.tu-darmstadt.de

[1] G. Cambray, J. C. Guimaraes, A. P. Arkin. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nat Biotechnol*. 2018;36(10):1005-1015. https://doi.org/10.1038/nbt.4238

[2] G. Terai, K. Asai. Improving the prediction accuracy of protein abundance in *Escherichia coli* using mRNA accessibility. *Nucleic Acids Research*. Volume 48. Issue 14. 20 August 2020. Page e81. https://doi.org/10.1093/nar/gkaa481

[3] Sequence-to-function deep learning frameworks for synthetic biology. J. Valeri, K. M. Collins, B. A. Lepe, T. K. Lu, D. M. Camacho. bioRxiv 870055; doi: https://doi.org/10.1101/870055