VILNIUS UNIVERSITY

IEVA PUDŽIUVELYTĖ

PROTEIN THERMOSTABILITY PREDICTION
USING SEQUENCE REPRESENTATIONS FROM
PROTEIN LANGUAGE MODELS

Course thesis

Vilnius, 2022

Contents

Introduction

The variety of proteins is no less diverse than the variety of organisms. Just as the latter set is divided into domains, there are different attempts to classify proteins into distinct subsets. One way is to consider the heat-resistance property of biological macromolecules, which is an important trait for practical applications, for example, PCR.

Earlier studies show that protein's sequence and structural properties influence the thermostability of the macromolecule (Modarres et al. 2016). Furthermore, one of the most recent achievements in the field of deep learning are transformer architecture-based language models or, particularly, protein language models that have not yet been used to classify proteins based on their thermostability. Therefore, it was decided to apply protein representations from the protein language model to make inferences about thermostability of the biological macromolecules.

There are transformer architecture-based language models trained in an unsupervised fashion to predict probabilities of elements in sequences (Devlin et al. 2018). Simultaneously, the process of training creates sequences' embeddings – the real numbered vectors that represent semantic connections of language components. These representations can be transferred as input to specific application models trained using a supervised learning method to complete the defined task, such as the classification problem. Unsurprisingly, the transition between two types of learning has a name of 'transfer learning'. This separation is practically useful because the computationally-heavy task to train the language model can be excluded from the development of the application model.

Since proteins can be represented in aminoacid sequences assumed as a particular language, there are protein language models – models trained on protein sequences – that provide embeddings as output. The multi-dimensional vectors are transferred for the application neural network as its input to observe the results and decide whether the computed representations are suitable to solve the specific biological task.

This work presents a novel way to predict thermal stability of proteins. The solution is a feed- forward neural network (FNN). To train the FNN, the evolutionary scale model 1b (ESM-1b) (Rives et al. 2021) is used to generate

embeddings for proteins of organisms with annotated growth temperatures (Engqvist 2018). The model takes the generated embedding to predict the thermostability class of the input protein.

Theory

The main objective of the work is to create a classifier that could determine to which thermostability class the protein sequence belongs.

Thermostability classes were created by setting the threshold of 65 °C - proteins that are stable at temperatures strictly lower than the given threshold should be assigned to class '0', whereas the remaining ones compose the class '1'.

ESM-1b embeddings

Due to the novelty of embeddings, a considerably good performance of protein language models, and a recently emerged availability of embeddings, it was decided to develop a neural network model that would take protein embeddings from ESM-1b model as input and give the thermostability class label as output.

ESM-1b is one of evolutionary scale models trained by Facebook Research (Rives et al. 2021). The model has 33 layers and 650 million parameters. The model was trained in an unsupervised fashion on UniRef50 dataset (March 28, 2018). In order to ensure determinism in the validation set, authors removed protein sequences that were longer than 1024 aminoacids.

The authors made a script to extract model's embeddings available in the repository "Evolutionary Scale Modeling". The script allows to choose from which model and layer embeddings will be taken, what embeddings (mean, per aminoacid, or beginning of the sequence token) to keep. In the result of using the script, a 1280 dimensional vector for protein is generated.

The fact that sequences longer than 1024 aminoacids were removed from the validation dataset for ESM-1b model's training implies to the limitation of model's embeddings, which cannot be generated for sequences longer than 1024 aminoacids. For this reason, various methods to get the most accurate prediction for longer sequences were tried.

Methods

Generally, the workflow consisted of the following steps:

1. Collecting sets of sequences

2. Calculating ESM-1b embeddings for the sets of sequences

3. Processing the set of generated embeddings

4. Visualising ESM-1b embeddings

5. Training and validating the neural network model of the chosen architecture

6. Testing the trained neural network model

7. Presenting the results of the model

The workflow will be reviewed in terms of these points.

Collecting data

Initially, the steps prior to the neural network model construction were done using a small dataset that was composed of proteomes of two organisms: a mesophilic bacteria *Escherichia coli* (UP000000625) and a thermophilic archaeon *Sulfolobus solfataricus* (UP000001974). The growth temperature of *E. coli* is 37 °C (Jang et al. 2017) and 80 °C for *S. solfataricus* (Zaparty et al. 2010).

Results

…

Conclusions

…

Bibliography

[1] H. P. Modarres, M. Mofrad, and A. Sanati-Nezhad, "Protein thermostability engineering," *RSC advances*, vol. 6, no. 116, pp. 115252–115270, 2016.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, 2021.

[4] M. K. M. Engqvist, "Growth temperatures for 21,498 microorganisms," Feb. 2018.

[5] J. Jang, H.-G. Hur, M. J. Sadowsky, M. Byappanahalli, T. Yan, and S. Ishii, "Environmental escherichia coli: ecology and public health implications—a review," *Journal of applied microbiology*, vol. 123, no. 3, pp. 570–581, 2017.

[6] M. Zaparty, D. Esser, S. Gertig, P. Haferkamp, T. Kouril, A. Manica, T. K. Pham, J. Reimann, K. Schreiber, P. Sierocinski, *et al.*, ""hot standards" for the thermoacidophilic archaeon sulfolobus solfataricus," *Extremophiles*, vol. 14, no. 1, pp. 119–142, 2010.