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 amino acid 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 apply emebddings from protein language models 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' and the remaining ones compose the class '1'.

Protein language models

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 (accessed March 28, 2018). In order to ensure determinism in the validation set, authors removed protein sequences that were longer than 1024 amino acids.

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 amino acid, or beginning of the sequence token) to keep. In the result of using the script, a 1280 dimensional vector for each protein is generated.

The fact that sequences longer than 1024 amino acids 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 amino acids. 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). This dataset was named '001' and used only for embeddings visualisation.

The visualisation of '001' stimulated to check whether embeddings are distinguished by the thermostability property or the life domain has a significant impact to the data clusterisation. Subsequently, '002' dataset was a collection of 2 mesophilic archaea and 2 thermophilic bacteria proteomes.

Mesophilic archaea:

- *Methanobrevibacter oralis* (UP000077428)
- *Nitrosopumilus maritimus* strain SCM1 (UP000000792)

Thermophilic bacteria:

- *Aquifex aeolicus* (strain VF5) (UP000000798)
- *Thermotoga maritima* (strain ATCC 43589 / DSM 3109 / JCM 10099 / NBRC 100826 / MSB8) (UP000008183)

After receiving the results of the first neural network model, it was decided to train and test the model on the first real dataset. '003' dataset was a subset of the dataset of 21458 annotated organisms (Engqvist 2018).

Requirements for '003' dataset were: the dataset needed to be balanced and a single taxonomy identifier could be apparent only in either training, validation, or testing dataset.

To make the dataset balanced, 216595 sequences from 51 proteomes taken for the class '0' and 212729 sequences from 111 proteomes of the class '1' (Table 1, Table 2).

Proportions that were chosen to divide classes of '003' dataset were 70%, 15%, and 15% for training, validation, and testing sets respectively.

Table 1: Numbers of different proteomes that compose training, validation, and testing datasets

	Class '0'	Class '1'
Training	32	77
Validation	8	17
Testing	11	17

Table 2: Numbers of amino acid sequences that compose training, validation, and testing datasets

	Class '0'	Class '1'
Training	145128	143868
Validation	33204	32616
Testing	38263	36245

Calculating embeddings

The final layer of ESM-1b model was used to generate protein embeddings taken as input to the classification neural network. For training, validation, and initial testing stages embeddings averaged over the full sequence were chosen. These embeddings were vectors of 1280 dimensions.

Processing generated embeddings

As it was mentioned in previous sections, ESM-1b embeddings can be calculated only for sequences that are not longer than 1024 amino acids. Therefore, after the generation of embeddings, sequences without embeddings were filtered out from the initial dataset. The final dataset consisted of 423127 embeddings (212144 of class '0' and 210983 of class '1') (Table 3).

These embeddings were saved to NPZ and TSV files. NPZ files are binary files that are used to load sequence representations to the model. TSV files with embedding vectors were created for a human-readable record and analysis with other tools. The TSV files were made headerless with columns representing the following information:

Table 3: Numbers of embeddings that compose training, validation, and testing datasets

	Class '0'	Class '1'
Training	141602	142707
Validation	32793	32363
Testing	37749	35913

- #0 - Taxonomy ID of the organism, to which the sequence belongs
- #1 - Accession ID of the sequence
- #2 - Length of the sequence
- #3 - Temperature label
- #4 - #1284 Components of embeddings

Visualising embeddings

Calculated embeddings were visualised using principal component analysis (PCA) from *Scikit-Learn Python* library (version 0.24.2) (Figure 1).

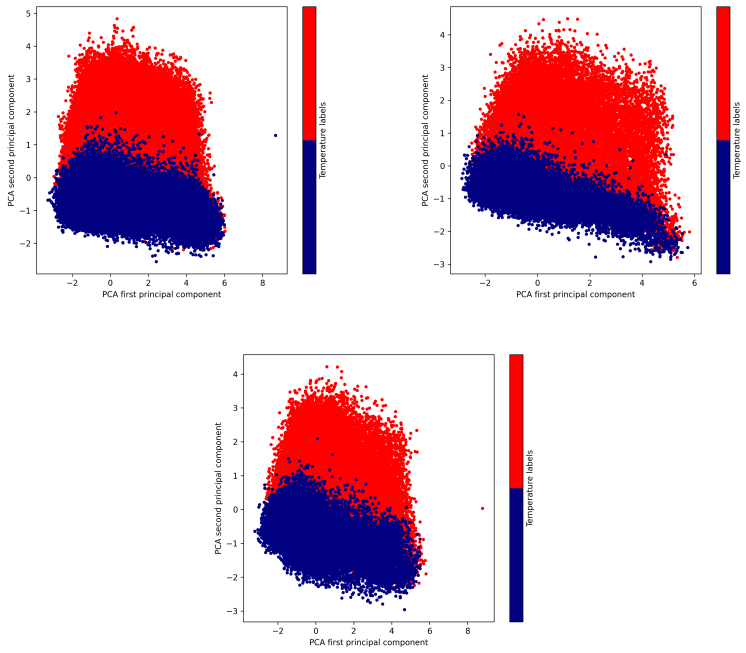

Figure 1: PCA visualisations of '003' model datasets' embeddings: training set in the upper left, validation set in the upper right, and testing set in the lower center

Additionally, *Python* minimum-distortion embedding (*PyMDE*) library (version 0.1.5) was used to visualise datasets of embeddings (Figure 2).

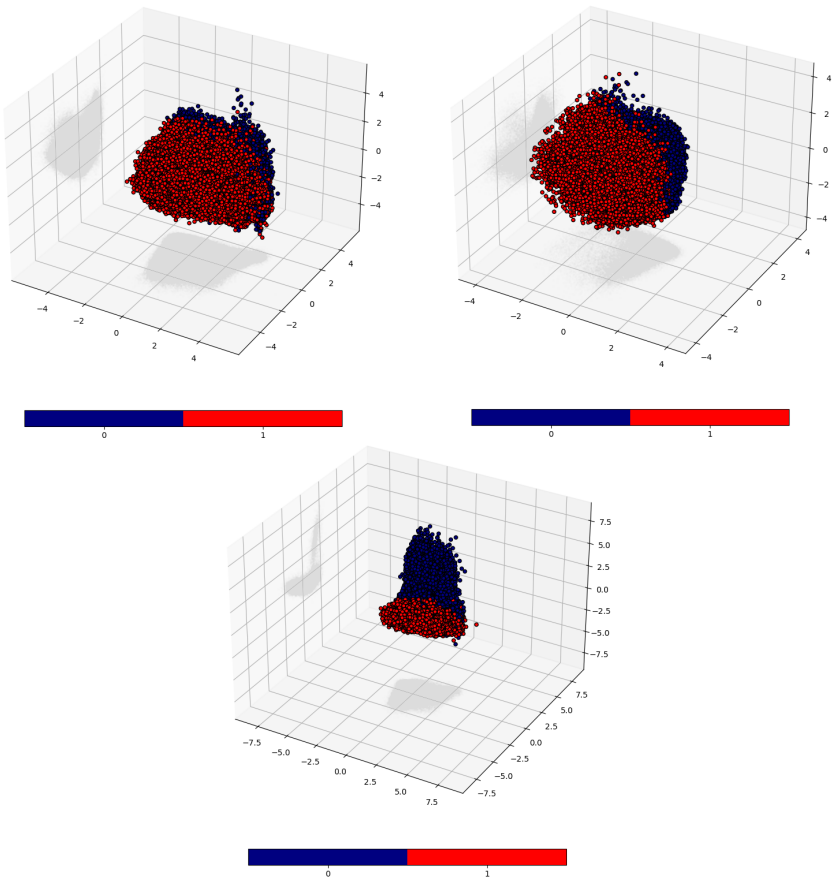

Figure 2: *PyMDE* PCA visualisations of '003' model datasets' embeddings: training set in the upper left, validation set in the upper right, and testing set in the lower center

Training and validation of the neural network

At first, it was decided to try training the neural network built of the most simple architecture - a single layer perceptron (SLP). The input of the layer is 1280-dimensional - it takes the input of protein embeddings - and the output is a binary label that represents the thermostability class. The activation function was chosen to be sigmoid and loss function binary cross entropy. Adam optimizer with learning rate of 0.0001 was chosen.

The model was trained and validated in 5 epochs taking mini-batch size of 24 embeddings.

Table 4: Confusion matrix after the fifth validation epoch

	0	1
0	32793	3636
1	1582	30770

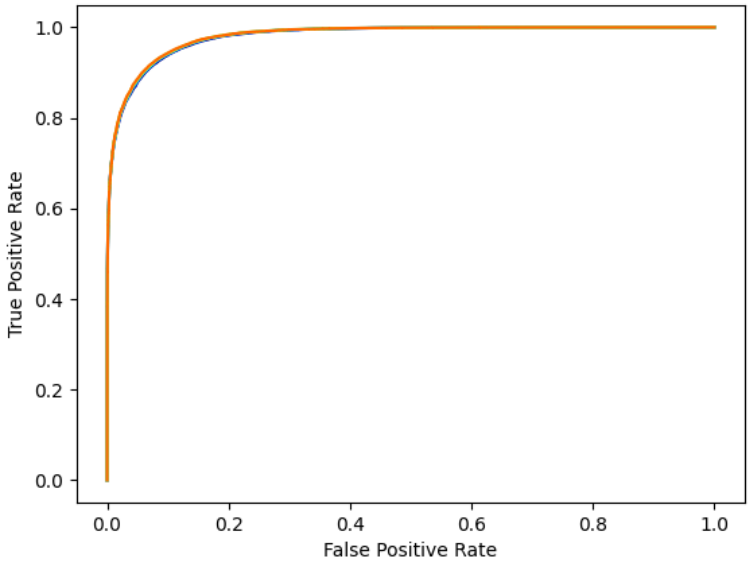

Figure 3: ROC curves for each validation epoch - the orange one denotes the curve of the fifth epoch

Table 5: Model performance metrics after the fifth validation epoch

Accuracy	0.91989
Precision	0.89432
Recall	0.9511
ROC AUC	0.92

Results

The model was tested with the testing subset of '003' dataset that was not exposed to the model before the testing stage.

Table 6: Confusion matrix in the testing phase

	0	1
0	34933	2813
1	2788	33122

Table 7: Model performance metrics in the testing phase

Accuracy	0.924
Precision	0.922
Recall	0.922
ROC AUC	0.924
MCC	0.845

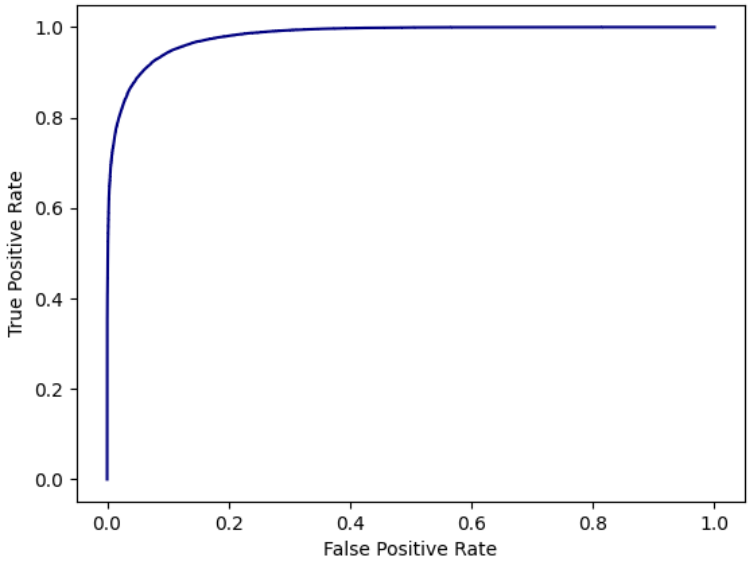

Figure 4: ROC curve after the testing stage

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.