Vilnius University

Mathematics and Informatics Faculty

Institute of Informatics

Bioinformatics study program

Protein thermostability prediction using sequence representations from protein language models

Author: Ieva Pudžiuvelytė

Supervisor: Kliment Olechnovič, PhD

Course work project

Vilnius, 2022

Contents

1 Introduction

This work is a prolongation of the previous work - the model that performed binary classification into thermostability classes. The model took ESM-1b protein embeddings as input and provided prediction for each protein, how likely it belongs to the thermostable class.

2 Abstract in Lithuanian (Santrauka)

3 Theory

3.1 ESM-1b embeddings

3.2 ProtTrans embeddings

4 Methods

4.1 Objective of this work

The main objective of this work is to analyse which numerical representation of proteins is the most suitable to use as input for the neural network model to make binary protein classification into thermostability classes. Additionally, it was decided to try model architectures with one or two hidden layers and evaluate whether the different architecture improves the performance.

4.2 Data set

The data set that was used for the research of this work was used the same as for the training, validation, and testing of the single-layer perceptron (SLP) with mean ESM-1b representations in the previous work, although for this research the data set was filtered out of identical sequences to get more accurate evaluations.

Table 1: Number of sequences with embeddings before and after filtering the data set

Subset	Original	After filtering
Training	284309	283360
Validation	65156	63158
Testing	73662	73308

Table 2: Number of sequences with embeddings in each class before and after filtering the data set

Class	Original	After filtering
0	216595	212129
1	212729	207697

4.3 Correlation analysis of embeddings' components

4.4 Analysed representations

Both protein language models - ESM-1b and ProtTrans - provide per token or per residue representations - each amino acid of the protein gets 1280 or 1024-dimensional vector from

ESM-1b or ProtTrans model respectively. Therefore, each protein is originally represented by $m \times n$ matrix, where m is the number of dimensions of the chosen type of embedding and n is the number of amino acids that compose the protein. These representations were processed to get vectors of the same dimension for each protein in the data set. The representations that were included in the analysis:

1. Mean ESM-1b and ProtTrans
2. Joined mean ESM-1b and ProtTrans
3. Normalised mean ESM-1b and ProtTrans
4. Joined normalised mean ESM-1b and ProtTrans
5. Median ESM-1b and ProtTrans
6. Minimum, median, and maximum ESM-1b and ProtTrans
7. Quantiles (including minimum and maximum) ESM-1b and ProtTrans
8. Quantiles (including minimum and maximum) and mean ESM-1b and ProtTrans
9. Octiles (including minimum and maximum) ESM-1b and ProtTrans

Table 3: Sizes of the analysed representations' vectors

Representation	ESM-1b	ProtTrans
Mean	1280	1024
Joined mean	2304	
Median	1280	1024
Minimum, median, maximum	3840	3072
Quantiles	6400	5120
Quantiles and mean	7680	6144
Octiles	11520	9216

4.5 Analysed architectures

Table 4: Sizes of the analysed representations' vectors

Model	Number of hidden layers	Size of hidden layers
SLP	0	-
C2H1_h128	1	128
C2H1_h256	1	256
C2H1_h512	1	512
C2H2_h256-128	2	256, 128
C2H2_h512-256	2	512, 256

5 Results

6 Conclusions

7 Availability

The code that was used to receive the results of this work can be found in the designated Github repository: https://github.com/ievapudz/Course_Work_Project.

References