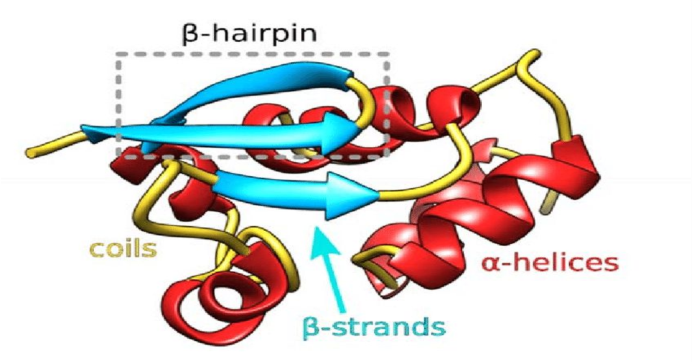
Protein secondary structure prediction using transformers

Sareh Ahmadi

Protein Secondary Structure Prediction

ss8 eight states (Q8): assigns one of the following secondary structure types to every amino acid in a protein

1. C: Loops and irregular elements
2. E: β-strand
3. H: α-helix
4. B: β-bridge
5. G: 3-helix
6. I: π-helix
7. T: Turn
8. S: Bend



sst3 three-state (Q3): "It is common to simplify the eight states (Q8) into three (Q3) by merging (E, B) into E, (H, G, I) into H, and (C, S, T) into C"

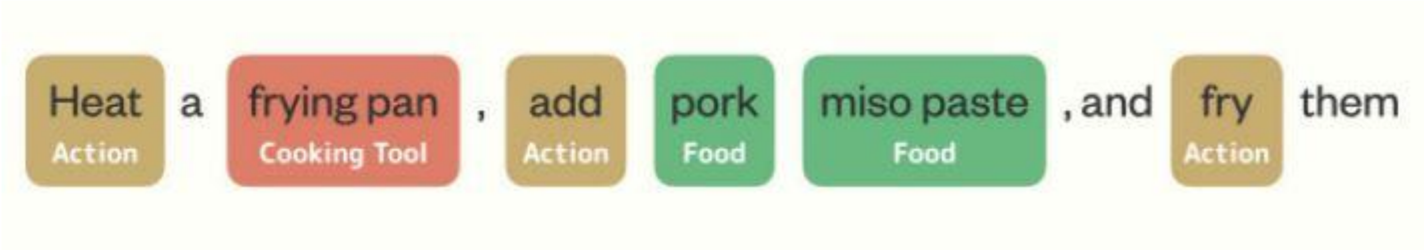
Secondary structure prediction

Given the input of amino acids:

Seq: G I V E Q C C T S I C S L Y Q L E N Y C N

Labels: C C C C C C C C C C C C H H H H H C C E C

Name Entity recognition (NER) task in NLP:



Heat a frying pan , add pork miso paste , and fry them

Dataset:

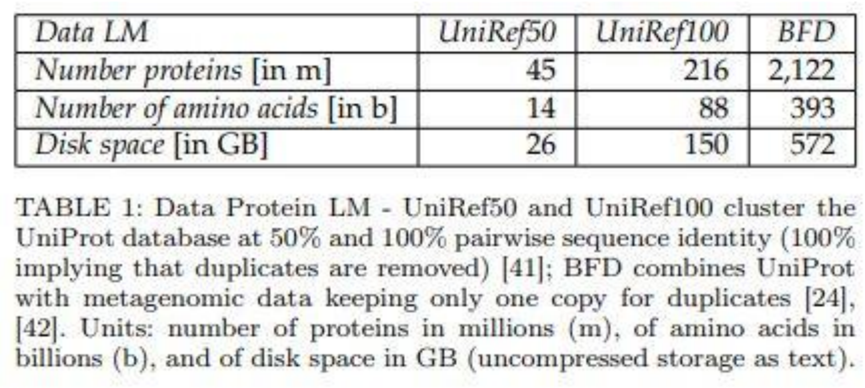
- Dataset is available on the Kaggle website
- Contains: a sequences with a variant length of 3-5037.
- For the class project, sequences with a length greater than 50 are kept.
- Overall, there are 8687 sequences
  - 20% for test
  - 10% for validation

Method

ProtTrans[1]:

- Pre-trained transformer model which has the similar artcitecure as BERT language model
- It is trained in self-supervised fashion using masked language modeling (MLM) objective

Training Data:



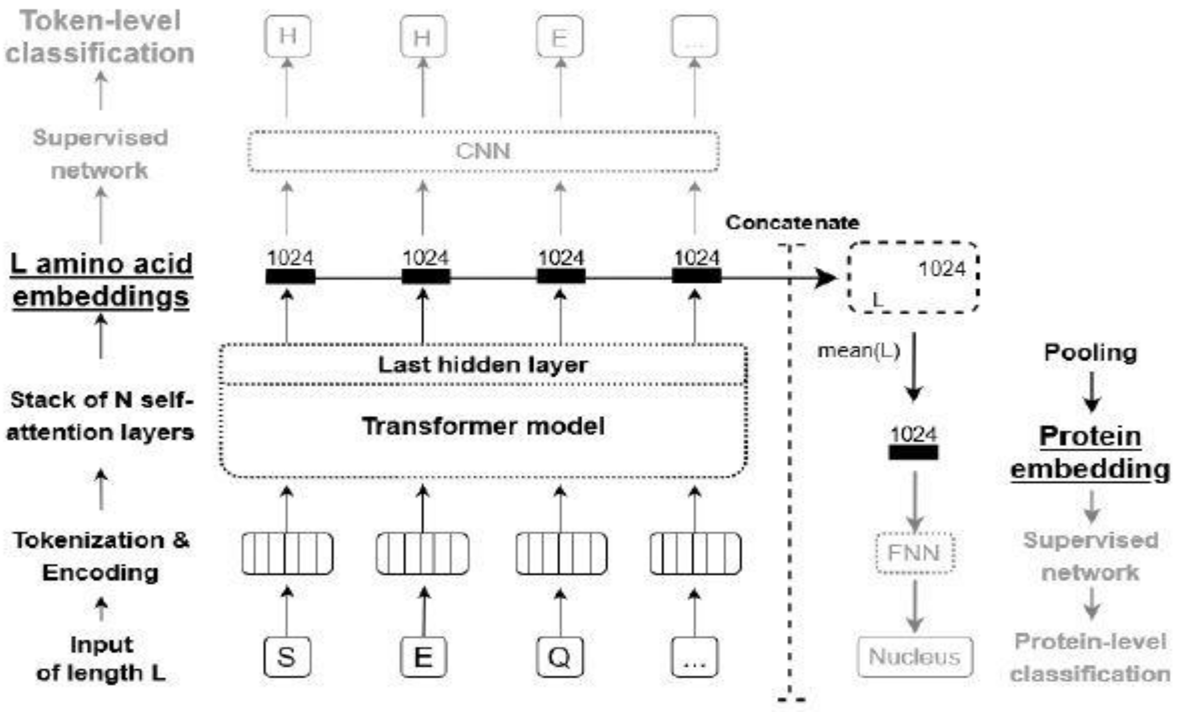
Data LM	*UniRef50*	*UniRef100*	*BFD*
Number proteins [in m]	45	216	2,122
Number of amino acids [in b]	14	88	393
Disk space [in GB]	26	150	572

TABLE 1: Data Protein LM - UniRef50 and UniRef100 cluster the UniProt database at 50% and 100% pairwise sequence identity (100% implying that duplicates are removed) [41]; BFD combines UniProt with metagenomic data keeping only one copy for duplicates [24], [42]. Units: number of proteins in millions (m), of amino acids in billions (b), and of disk space in GB (uncompressed storage as text).

[1] ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning

Method

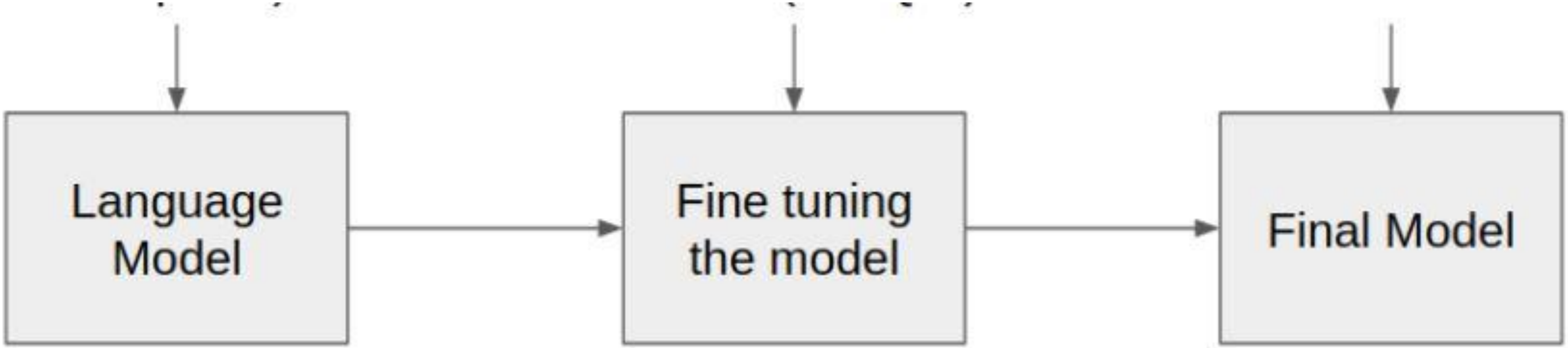
- One way is to use the model as feature extractor
- Use the features for downstream tasks such a classification



Method

“ Could gain more accuracy by fine-tuning the model rather than using it as a feature extractor.”

Pretrain on large dataset (Protein Databases) **Task specific dataset(Kaggle)** **Test Dataset**



Number of Prottrans model parameters is: 419 million

Results

accuracy	precision	recall	f1
0.8967	0.7990	0.7506	0.7741

Sequence : G S H N A D L S E A L R E L R R E L M K E T G Y S A F V V F T N A T L E A L A A R Q P R T L A E L A E V P G L G E K R I E A Y G E R I L D A I N T V L D G

Ground Truth is: C C C H H H H H H H H H H H H H H H H H H H C C C H H H H C C H H H H H H H H H H C C C C H H H H C C C C C C C H H H H H H H H H H H H H H H H H H H H C

prediction is : C C C C H H H H H H H H H H H H H H H H H H C C C H H H H C C H H H H H H H H H H C C C C H H H H H C C C C C C H H H H H H H C H H H H H H H H H H H C C