Protein secondary structure prediction using transformers

Sareh Ahmadi

Protein secondary structure prediction

- Protein secondary structure prediction is helpful for structure prediction
- Accurate prediction of the secondary structure can lead to understanding how the proteins fold
- Secondary structure prediction can be beneficial for detecting protein malfunction
- Secondary protein structure can be simplified from an eight-class calcification (8-state Q8) into a three-class classification task (3-state Q3)
- Residuals are assigned to three categories of alpha-helix (H), beta-sheet (B), and coil (C).

Previous Work

- Recent methods have been proposed to leverage large proteins datasets using deep neural networks such as convolution networks [2]
- An ensemble of the Long-Short-Term Memory Cells in Bidirectional Recurrent Neural Networks (LSTM-BRNNs) and convolution networks [3]
- Ensembled of several neural networks architectures including recurrent convolutional neural network (RCNN), convolutional recurrent memory network (CRMN), FractalNet [4]

Method

- Protein language modeling is the same as language modeling but on protein sequences
- Leverage protein language modeling using pre-trained transformers ESM-1b, ESM-1v, MSA , ProteinBERT, Prottrans
- Transformers are language models which are consist of self-attention blocks[5].Powerful transformers such as BERT [6] and RoBERTa [7]
- The difference between these transformer-based protein language models is in the data sets, the pertaining language modeling objective, and some changes in the architecture of the transformers
- Trained on large and diverse proteins databases such as UniRef50 and UnifRef90, UnifRef100
- Effective for downstream tasks such as prediction of secondary structure, tertiary contacts, remote homology, mutation effect, per-protein location, membrane prediction, etc.
- Except for MSA transformers, other transforms can be used without the multiple sequence alignment (MSA) of a protein
- Most of the pre-trained transformers are available in the HuggingFace library [13] and can be used for downstream tasks

[5] “Attention is all you need,” Advances in neural information processing systems, A. Vaswani, , 2017

[6] “BERT: Pre-training of deep bidirectional transformers for language understanding,” J. Devlin, 2019

Method

Dataset:

- This dataset is available on the Kaggle website
- A sequence of the peptide with a variant length of 3-5037.

Evaluation:

- Accuracy is widely applied across the literature for protein secondary structure prediction.