Learning co-evolution information with natural language processing for protein folding problem*

***Egor Zverev*[1], *Sergei Grudinin*[2], *and Ilia Igashov*[1,2]**
zverev.eo@phystech.edu; sergei.grudinin@inria.fr; igashov.is@phystech.edu
[1]Organization, address; [2]Organization, address

Many problems in bioinformatics are related to protein folding. Most of the techniques aimed to solve such problems usually begin by building protein sequence descriptors. One of the most significant parts of such descriptors is co-evolution information which computation is based on Multiple Sequence Alignments (MSA). This technique requires significant computational efforts, moreover, it does not guarantee precise results since it fully relies on finite databases. Besides, this method fails for sequences with shallow alignment. In this work, we will consider pre-trained language models (LMs) as a potential alternative to this increasingly time-consuming database search. Specifically, our main focus is fold classification problem. We will examine several state-of-the-art methods and modify them by replacing MSA-based feature-generation parts with pre-trained LMs.

Keywords: *fold classification; co-evolution, NLP; transformers; BERT;*

DOI: 10.21469/22233792

1 Introduction

Protein properties are determined by its shape which is specified by its amino acid sequence [8]. Therefore it is important to learn how to analyze this sequence. Any analysis usually begins by constructing protein sequence descriptors whose essential part is a co-evolution information. This information can be obtained by searching for homologues of the target protein in large databases and further computing multiple sequence alignment (MSA) on it.

Co-evolution-based methods assume that a mutation of one amino acid very often leads to mutations of others. Therefore given an amino acid sequence it is natural to look for other similar sequences. That is usually performed with MSA [2]. The problem arises when we're dealing with the sequences that did not evolve a lot. They will have shallow alignments, i.e. there will be only a few similar sequences. Therefore, information extracted from these alignments will be insignificant.

Besides, computing MSA is a heavy task. Moreover, alignment databases are finite. If for a given sequence there are no matches in the database, MSA-based methods fail to produce reliable results.

In this work, we will consider pre-trained language models (LMs) [5] as a potential alternative to traditional MSA approaches. It is assumed that protein sequences are not random [1], that there is logic in amino acids structure. Therefore, a set of all amino acids could be seen as a language with complicated inner rules. With the invention of BERT [3] it has become possible to learn the structure of any abstract language. It is natural to assume that by learning amino acids structure, BERT will be able to implicitly learn co-evolution information.

Our main aim is to study the efficiency of pre-trained LMs in application to the fold classification problem. To start with, we will consider the state-of-the-art method DeepSF [6] which solves the protein fold classification problem. By replacing its MSA-based feature-generation

*

part with a pre-trained LM, we will study how NLP approach affects on learning the fold-related information. The whole framework is schematically represented in Figure 1. Further, we can do the same with other fold classification algorithms [4, 9] as well as with end-to-end protein structure prediction methods [7, 10].

2 Preparing a manuscript

Manuscripts are prepared using `jmlda.sty` style package. You are recommended to use `jmlda_rus.bst` and `jmlda_eng.bst` style files for generating bibliography using BibTEX.

Visit the `http://jmlda.org/?lang=en` website for detailed submission instructions, templates and other information.

Please note that this file must be saved in `UTF-8` encoding. Where possible select `UTF-8 without BOM` encoding. To change the encoding please use `Sublime Text` or `Notepad++` text editors.

3 Structure of the article

Divide your article into clearly defined and numbered sections and paragraphs.

3.1 Paragraph

Sections and paragraphs are numbered and have a brief heading.

4 Concluding Remarks

This section should provide the summary and explore the significance of the results achieved and list problems not yet solved. Results should be clear and concise.

References

[1] Davide De Lucrezia, Debora Slanzi, Irene Poli, Fabio Polticelli, and Giovanni Minervini. Do natural proteins differ from random sequences polypeptides? natural vs. random proteins classification using an evolutionary neural network. *PLOS ONE*, 7(5):1–10, 05 2012.

[2] S. de Oliveira and C. Deane. Co-evolution techniques are reshaping the way we do structural bioinformatics. *F1000Res*, 6:1224, 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[4] W. Elhefnawy, M. Li, J. Wang, and Y. Li. DeepFrag-k: a fragment-based deep learning approach for protein fold recognition. *BMC Bioinformatics*, 21(Suppl 6):203, Nov 2020.

[5] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*, 2020.

[6] Jie Hou, Badri Adhikari, and Jianlin Cheng. DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, 34(8):1295–1303, 12 2017.

[7] Shaun M Kandathil, Joe G Greener, Andy M Lau, and David T Jones. Deep learning-based prediction of protein structure using learned representations of multiple sequence alignments. *bioRxiv*, 2020.

[8] Akif Uzman. Molecular biology of the cell (4th ed.): Alberts, b., johnson, a., lewis, j., raff, m., roberts, k., and walter, p. *Biochemistry and Molecular Biology Education*, 31(4):212–214, 2003.

[9] A. Villegas-Morcillo, A. M. Gomez, J. A. Morales Cordovilla, and V. E. Sanchez Calle. Protein fold recognition from sequences using convolutional and recurrent neural networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1, 2020.

[10] Jinbo Xu, Matthew Mcpartlon, and Jin Li. Improved protein structure prediction by deep learning irrespective of co-evolution information. *bioRxiv*, 2020.

Received January 01, 2017