DOI 10.18699/SBB-2022-163

Transfer learning for protein stability prediction

Konstantinova A.V.[1, 2, 3]*, Kardymon O.L.[2], Ivanisenko N.V.[1, 2, 3]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Artificial Intelligence Research Institute, Moscow, Russia*

[3] *Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

** konstantaalina@gmail.com*

Key words: protein stability, structure prediction

Motivation and Aim: Protein stability is an important biologically significant characteristic. It is, for example, crucial in protein design or has to be considered when studying protein variants and effects of mutations. Deep learning approaches, including AlphaFold2 [2], currently demonstrate highest accuracy in protein structure prediction. In this work we aimed to investigate the accuracy of protein stability changes prediction using amino acid residues representations extracted from pretrained deep learning models including ESM-1b [1] and AlphaFold2 [2].

Methods and Algorithms: Benchmark included publicly available datasets of experimentally studied proteins with known stability characteristics including melting temperature or free energy changes upon amino acid substitutions. The amino acid representations were extracted from the hidden layers of pretrained models, including ESM-1b and AlphaFold2. Extracted representation were used to train the model to predict stability changes using boosting algorithm.

Results: The model based on AlphaFold2 and protein language models was developed, trained and demonstrated comparable accuracy to peer-reviewed methods. The obtained model was validated to estimate the effects of known SARS-CoV-2 variants on S-protein structure and stability.

Conclusion: Transfer learning approach applied in the current work allows to efficiently predict protein stability changes and can be applied for a wide range of protein design tasks.

Acknowledgements: Work was funded by the Ministry of Science and Higher Education of the Russian Federation project “Kurchatov Center for World-Class Genomic Research” No. 075-15-2019-1662 from 2019-10-31.

References

1. Rives A., Meier J., Sercu T., Goyal S., Lin Z., Liu J., ..., Fergus R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS.* 2021;118(15):e2016239118.
2. Jumper J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589.