[BIOINFO 2024]

Improving Protein Embeddings through Sequence-Structure Contrastive Learning

Min Su Yoon[1], Minkyung Baek[1,*]

[1] *School of Biological Sciences, Seoul National University*
**Corresponding author: minkbaek@snu.ac.kr*

Protein sequence representation learning has emerged as a prominent technique in the field of computational biology with the recent success of protein language models. While protein language models are attractive due to their powerful feature extraction capabilities without the need for multiple sequence alignments, it is widely known that their dependence on evolutionary information is still present. We implement a contrastive learning framework motivated by CLIP in order to overcome this limitation, where the cosine similarity between the sequence and structure embeddings of a protein are maximized. Methods empirically shown to improve representation learning performance, such as the log-sigmoid loss and locked encoder tuning, have also been applied. Results indicate that while the protein sequence and structure embeddings are aligned within the hyperspherical latent space, the performance on downstream tasks, such as Enzyme Commission Number Classification, does not outperform the baseline. We will discuss these findings in detail in this poster.

Acknowledgement: This work has been supported by IITP/MSIT (RS-2023-00220628), NRF/MSIT (RS-2023-00210147).