Fast and Accurate Prediction of Intrinsically Disordered Protein by Protein Language Model

([1]*Graduate School of Environmental Science, Hokkaido University*, [2]*Faculty of Environmental Earth Science, Hokkaido University*) ○Shijie Xu,[1] Akira Onoda [1,2]

Keywords: Intrinsically disordered protein; Deep learning; Protein language model

Intrinsically disordered proteins (IDPs) play a crucial role in many critical biological processes and have attracted increasing attention over the past few decades. Despite its importance, recognition of protein intrinsically disordered regions (IDRs) is still laborious and expensive work and usually takes a long time because it requires accurate protein structures as references. On the other hand, predicting IDRs from protein sequences gives us fast and useful tools for protein analysis. However, the majority of existing IDP predictors require multiple sequence alignments (MSAs) as input features, which are generated from the alignments of homologous sequences by searching against the whole protein database. Since the rapid increase of protein sequences, it becomes increasingly time-consuming. Therefore, the alternative method does not rely on MSAs are needed.

In this presentation, we proposed a novel method named IDP-PLM to predict protein IDRs from sequences, based on the protein language model (PLM)[1]. The method does not utilize MSAs or any MSA-based input features but leverages only the protein sequences. It achieved state-of-the-art performance even compared with predictors using traditional MSAs features. The proposed method is composed of stacked predictors based on PLM, which are used to extract various protein features from sequences. In addition, these stacked predictors can also be regarded as independent predictors for secondary structure prediction, linker prediction, and binding predictions. All these achieved the highest accuracy on several independent test datasets compared to the existing works. The ablation experiments reveal the necessity of these stacked predictors in improving overall performance. Exemplary predictions showcase both short and long IDRs can be precisely captured by IDP-PLM, as shown in the following Fig. 1.

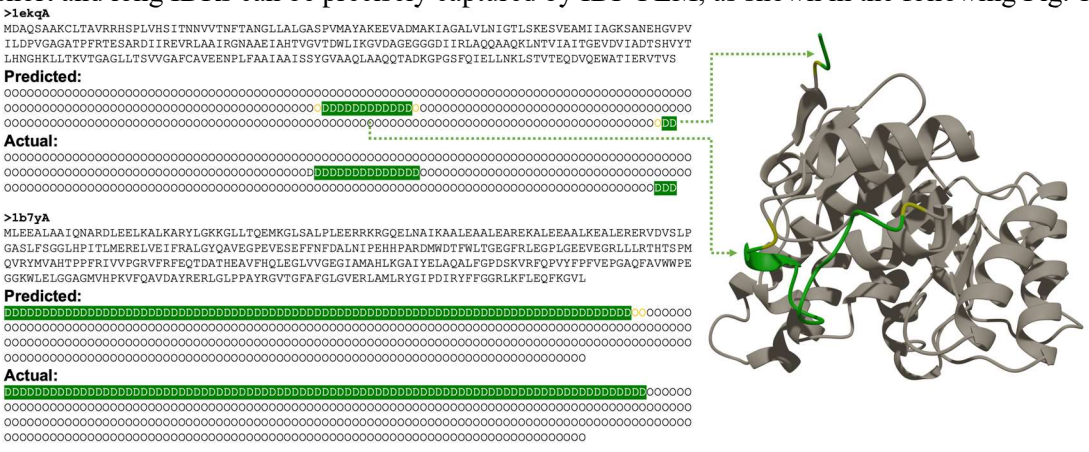

Fig. 1. (a) Exemplary prediction from IDP-PLM predictor. Here 'O' represents the ordered regions and character 'D' represents intrinsically disordered regions (colored in **green**). The predicted IDRs are compared with actual IDRs, and the wrong predicted residues are also colored **yellow**. (b) The actual IDRs (**green**) in chain A of protein 1ekq are visualized by ESMFold and ChimeraX.

1) Rives, A. et al. *Proc. Natl. Acad. Sci. U S A.*, **2021,** 118(15).