[BIOINFO 2024]

Generating MSA for Protein Structure Prediction via Structure Search and Sequence Design

Soohyun Jo, Hayoung Lee, Heesoo Ki[#], Min Su Yoon[#], Jeonghoon Park, and Minkyung Baek[*]

School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea
 # - equal contribution
**Corresponding author: minkbaek@snu.ac.kr*

The field of protein structure prediction has been extensively explored, leading to the development of numerous computational tools such as RoseTTAFold, ColabFold, and AlphaFold. A majority of them depend on multiple sequence alignment (MSA) which serves the role of a core feature source. However, accurate MSA construction is often challenging, especially when homologous sequences are scarce as in the cases of orphan and viral proteins. Here, driven by such limitations, an alternative approach to generate MSA is suggested, which aims to supplement sequences as well as to provide structural information. Using Foldseek and a modified version of ProteinMPNN, we expand MSA based on structurally homologous and designed sequences. The incorporation of the extended set of sequences to the original MSA generated by ColabFold using MMseqs2 improved the structure prediction of T1123 from CASP15, increasing TM-score up to 0.08. In this presentation, I will discuss current status and bottlenecks of this approach, based on the benchmark results on CASP15 protein structure prediction targets.

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