Metagenome-scale structural homology detection with Foldseek-ProstT5

Deep-learning based methods like AlphaFold2 have revolutionized structural biology, enabling sequence analysis based on rich structural information and not only amino acids. Foldseek, for instance, enables fast and sensitive homology detection of protein structures. Together with predicted structures from ColabFold—a ~100x accelerated version of AlphaFold2—Foldseek facilitated sensitive annotation of dark parts of a sponge proteome, identifying an additional 50% beyond sequence-based methods. However, structure prediction of entire metagenomes remains cost-prohibitive, despite ColabFold's acceleration.

To pave the way for structural metagenomics and sensitive annotation of previously dark proteins, we present Foldseek-ProstT5. Utilizing the ProstT5 protein language model, we replace costly structure prediction with >3500x accelerated translation of amino-acid sequences directly to structural interaction tokens (3Di). On the Foldseek sensitivity benchmark, ProstT5's 3Di sequences improve sensitivity for Fold, Superfamily, and Family recognition by 4.3%, 12.8%, and 23.1% respectively without backbone coordinates, and by 3.4%, 10.5%, and 18.2% respectively with backbone information. We have further improved ProstT5 by shifting to a newer ModernBERT architecture, allowing for a 20-times smaller model that maintains similar searching sensitivity.

Foldseek-ProstT5 is free and open source software available at https://foldseek.com and its webserver at https://search.foldseek.com.

van Kempen, Kim, et al. Nat. Biotechnol., 2023
Heinzinger, Weissenow, et al. bioRxiv, 2023.07.23.550085, 2023
Ruperti, Papadopoulos, et al. Genome Biology, 2023

Authors: Milot Mirdita, Victor Mihaila, George Bouras, Michael Heinzinger, Martin Steinegger
Keyword (3 or more): protein structure analysis, protein language model, homology detection, metagenomics