Petabase-scale Homology Search for Structure Prediction

Sewon Lee[1,*], Gyuri Kim[1,*], Eli Levy Karin[2], Milot Mirdita[1], Sukhwan Park[3], Rayan Chikhi[4], Artem Babaian[5,6], Andriy Kryshtafovych[7] and Martin Steinegger[1,3,8,9,†]

[1]School of Biological Sciences, Seoul National University, South Korea; [2]ELKMO, Denmark; [3]Interdisciplinary Program in Bioinformatics, Seoul National University, South Korea; [4]Institut Pasteur, Université Paris Cité, France, [5]Department of Molecular Genetics, [6]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Canada; [7]Genome Center, University of California, Davis, USA; [8]Artificial Intelligence Institute, [9]Institute of Molecular Biology and Genetics, Seoul National University, South Korea
*Equal contributors; †Correspondence: martin.steinegger@snu.ac.kr

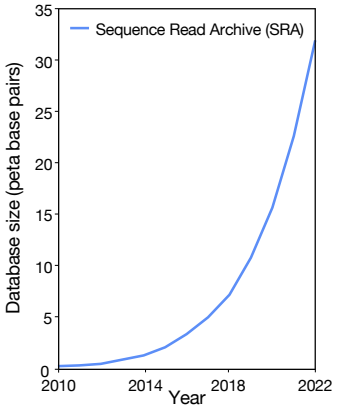
Abstract

The CASP15 competition highlighted the critical role of multiple sequence alignments (MSAs) in protein structure prediction, as demonstrated by the success of the top AlphaFold2-based prediction methods. To push the boundaries of MSA utilization, we conducted a petabase-scale search of the Sequence Read Archive (SRA), resulting in gigabytes of aligned homologs for CASP15 targets. These were merged with default MSAs produced by ColabFold-search and provided to ColabFold-predict. By using SRA data, we achieved highly accurate predictions (GDT_TS>70) for 66% of the non-easy targets, whereas using ColabFold-search default MSAs scored highly in only 52%. Next, we tested the effect of deep homology search and ColabFold's advanced features, such as more recycles, on prediction accuracy. While SRA homologs were most significant for improving ColabFold's CASP15 ranking from 11th to 3rd place, other strategies contributed too. We analyze these in the context of existing strategies to improve prediction.

SRA: the largest public sequence database

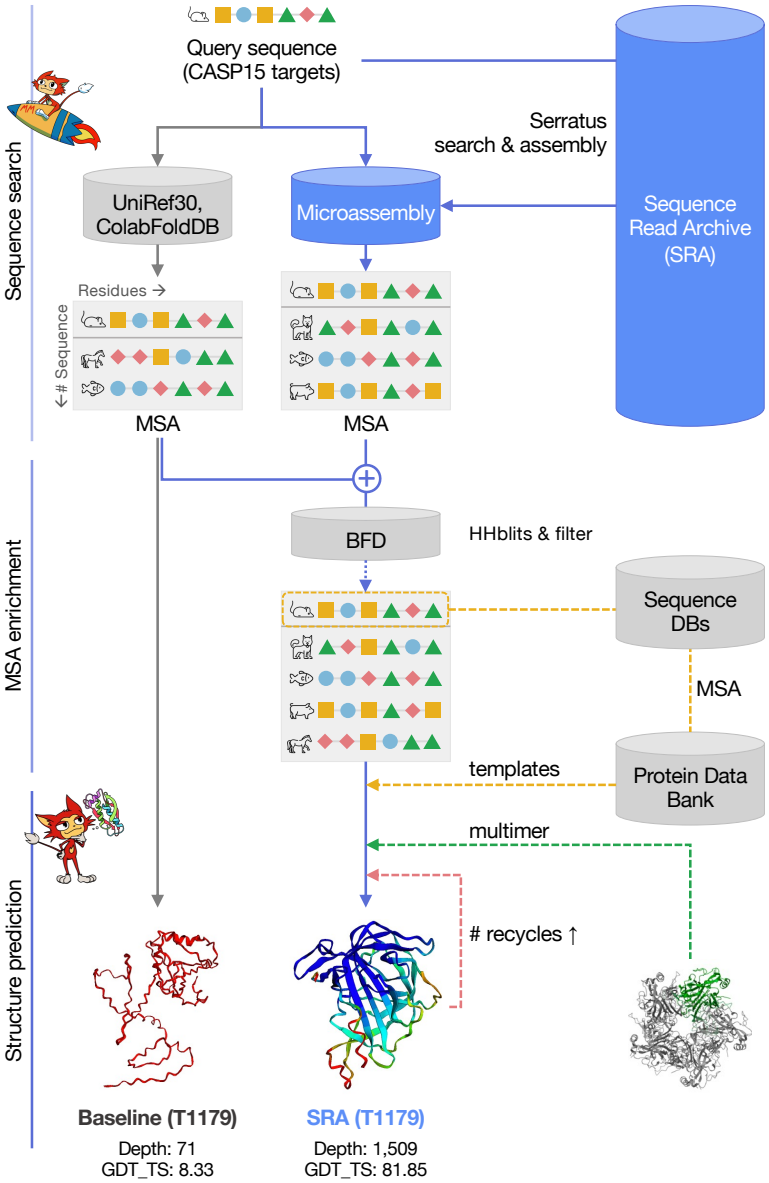
Database	Giga base pairs
SRA	32,000,000
IMG/M	27,000
MGnify*	3,500
BFD*	2,600
Metaclust*	1,700
Microassembly	**1,500**
ColabFoldDB*	800
UniRef30*	30

*Estimated based on # sequences

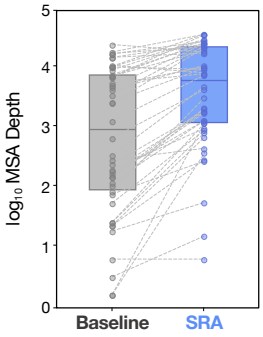


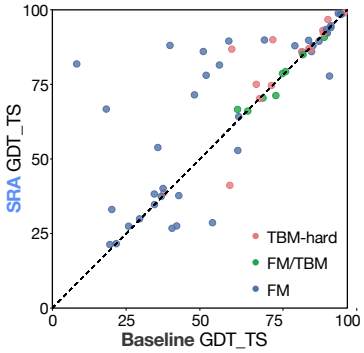
Microassembly: *search results* from SRA assembled with rnaviralSpades

Deep & wide sequence search for structure prediction



SRA-enriched MSAs lead to better predictions

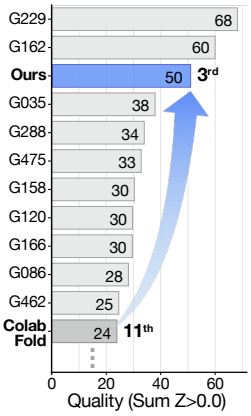




TBM: templated-based modeling; **FM:** free-modeling
GDT_TS (global distance test): prediction accuracy

Factors contributing to improved prediction

	GDT_TS	Sum Z (>0.0)
CFDB (Baseline)	65.8 ± 25.6	17.1
CFDB + HH	67.4 ± 25.2	15.8
CFDB + SRA	71.0 ± 24.4	**30.3**
CFDB + SRA + HH	**71.4 ± 23.3**	26.3
CFDB + rec	68.1 ± 25.3	17.7
CFDB + SRA + temp	70.9 ± 24.0	31.6
CFDB + SRA + mul	71.0 ± 24.1	31.0
CFDB + SRA + rec	**75.5 ± 22.2**	**40.6**
Model 1	**77.9 ± 20.6**	**50.3**



Abbr. HH: HHblits; rec: # recycles ↑; temp: templates; mul: multimer
Sum Z: sum of GDT_TS (accuracy) Z-scores among server groups over 0
Model1: predicted-best strategies for each target

References

Lee & Kim, et al. *bioRxiv*, 2023.07.10.548308 (2023). *Accepted in CSHL Perspect. Biol.*
Edgar, et al. *Nature*, *602*, 142-147 (2022).
Jumper, et al. *Nature*, *596*, 583-589 (2021).
Kryshtafovych, et al. *Proteins*, *89*, 1607-1617 (2021).
Mirdita, et al. *Nat. Methods*, *19*, 679-682 (2022).
Ovchinnikov, et al. *Science*, *355*, 294-298 (2017).
Steinegger & Söding. *Nat. Biotechnol.*, *35*, 1026-1028 (2017).

Download Poster (PDF)