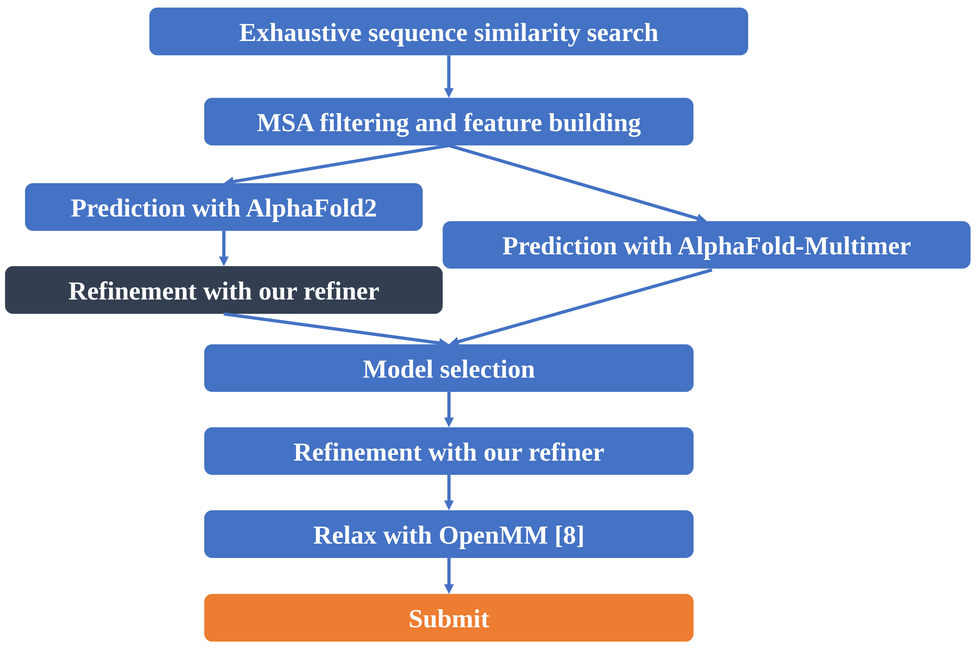
Protein tertiary and quaternary structure prediction using AlphaFold2 with various metagenomic databases

Toshiyuki Oda (oda@pezy.co.jp)

XTus Inc.

Abstract At CASP14, AlphaFold2[1], developed by DeepMind, demonstrated outstanding accuracy of monomeric structure prediction. After the competition, its derivative, AlphaFold-Multimer[2], was presented. It also showed excellent performance in predicting multimeric structures. Since their inference code and weights are publicly available under the generous license, their predictions will be the baseline for CASP15. Therefore, I set the following challenges for CASP15: (1) Collect a sufficient number of evolutionary related sequences for inputs: In addition to the tools and the databases which AlphaFold2 and AlphaFold-Multimer pipelines are using, I used PZLAST[3,4] and PSI-BLASTexB[5,6] with in-house databases constructed from metagenomic assemblies in NCBI Assembly database[7]. (2) Improve the structures generated by AlphaFold2 or AlphaFold-Multimer: I made a fine-tuned version of AlphaFold-Multimer to refine predicted structures.

Basic Pipeline

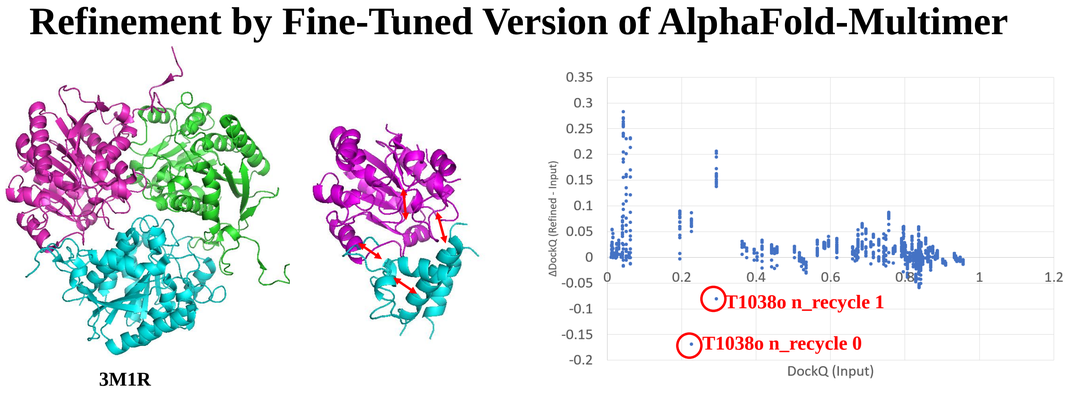


Exhaustive Sequence Similarity Search

	Tool	nThreads	nIter	DB	DB Size	Search Time
A	PZLAST	-	-	public metagenomic amino acid sequences	2.4 TB	~ XX mins
B	PSI-BLASTexB	128	~3	nr+in-house metagenomic database	370GB +680 GB	~ X hrs
C	hhblits[11]	128	3	Uniclust30[13]	220 GB	~ X mins
D	hhblits	6	2	BFD[14]	1.9 TB	~ X mins
E	jackhammer[12]	12	3	Uniprot/TrEMBL[15]	110 GB	~ X hrs
F	jackhmmer	128	3	MGnify[16]	130 GB	~ X hrs

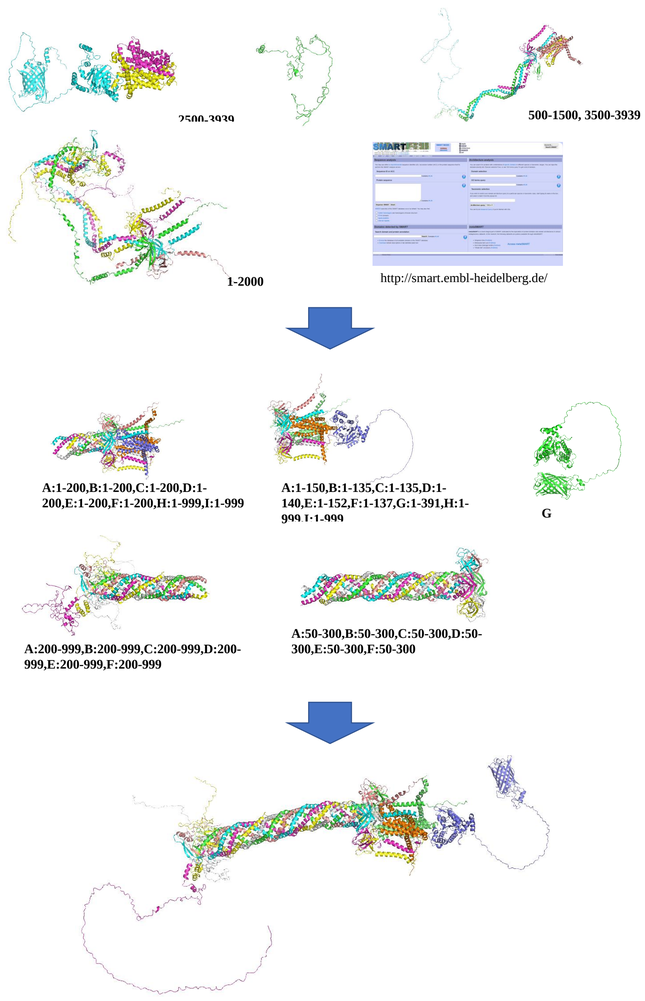
I used various tools and databases to get evolutionary related sequences as many as possible. The total processing time was several hours per sequence.

Refinement by Fine-Tuned Version of AlphaFold-Multimer



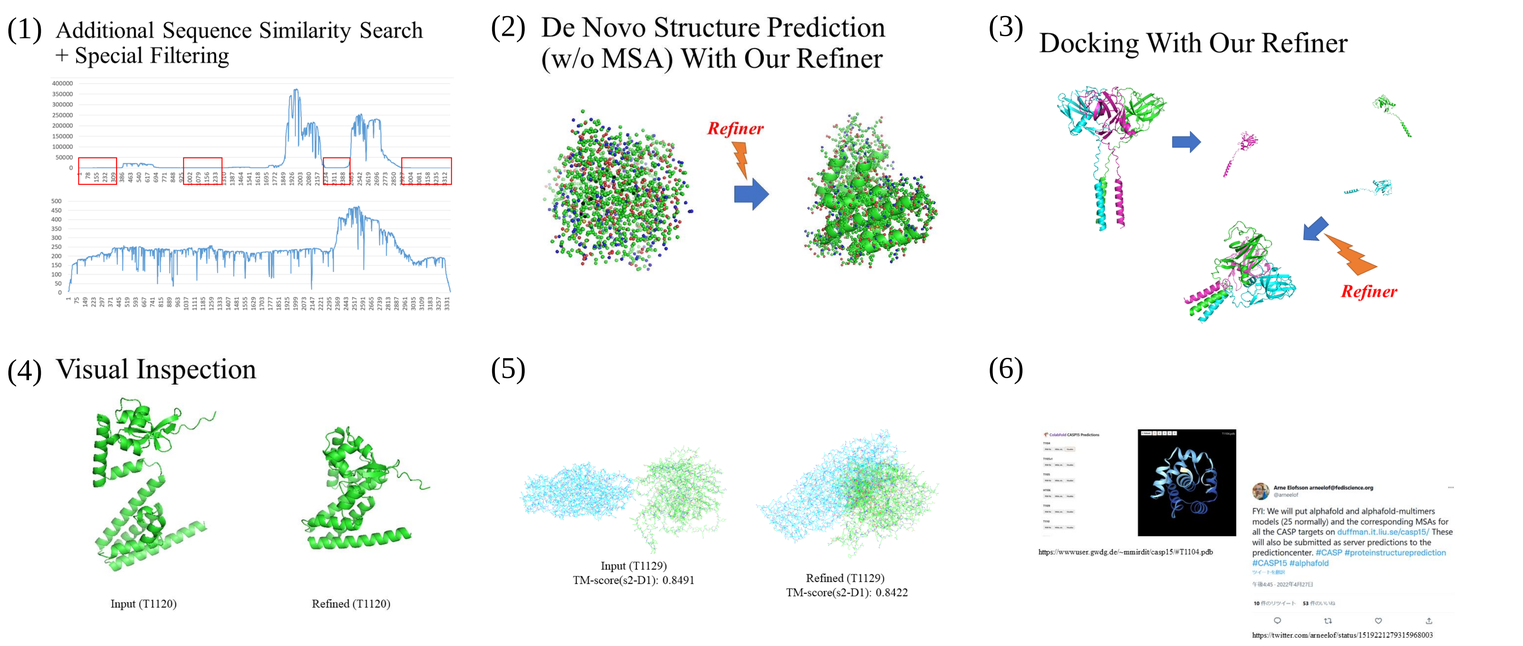
To build a refinement model, I fine-tuned the AlphaFold-Multimer model using single sequences and predicted structures as input. I intended to train the model to learn physically preferred structure; for example, I used segments with a sufficient number of contacts with other segments as ground truth. The right scatterplot indicates the DockQ[9] of the refined model subtracted by that of the input model as the function of DockQ of the input model. The dataset consists of the Benchmark2 dataset introduced in the study by Ghani et al.[10] and the CASP14 assembly target in which their experimental structure was available In most cases, the refined structures show higher accuracy than the input structures.

Domain Parsing/MSA Cropping



Basically, I fed AlphaFold2 and AlphaFold-Multimer full-length sequences of all subunits. However, if a structure was too large, our GPU could not handle it. In that case, I first split the sequence(s) into several parts, either randomly or according to the results of domain prediction. Next, I predicted the substructures from the partial sequences and investigated which domains and subunits would interact with each other, and estimated good positions to split. I then split the full-length sequence(s) according to the estimation and built the partial model again. Then I concatenated the partial models to construct the overall structure.

Visual Inspections / Manual Interventions



I performed several visual inspections and manual interventions: (1) If the depth of the MSA was highly skewed, I kept sequences with amino acids in thin regions and randomly removed other sequences to flatten the depth. (2) When I could not find any reliable evolutionary related sequence, atoms were randomly placed and fed into the refiner. (3) If a highly reliable multimeric structure was not found, chains were randomly placed and fed into the refiner. (4) If the refined model was found to became spherical, the I did not submit that model. (5) if the refined model had too many atom collisions, I did not submit it. (6) I checked the predictions of other groups to assess whether our protocol was working well.

Results I evaluated our predicted multimeric models against 8 currently available experimental structures (H1106,H1111,T1119o,T1121o,T1123o,T1124o,T1170,H1134) with MM-align[17]. TM-score of our MODEL 1 structures were higher than the average of all submitted MODEL 1 structures in 6 out of 8 cases, indicating that our protocol worked well. More reliable results will be provided by the assessors at the conference.

References

[1] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.
[2] Evans, Richard, et al. "Protein complex prediction with AlphaFold-Multimer." *BioRxiv* (2022): 2021-10.
[3] Mori, Hiroshi, et al. "PZLAST: an ultra-fast amino acid sequence similarity search server against public metagenomes." Bioinformatics 37.21 (2021): 3944-3946.
[4] Ishikawa, Hitoshi, et al. "PZLAST: an ultra-fast sequence similarity search tool implemented on a MIMD processor." *2021 Ninth International Symposium on Computing and Networking (CANDAR)*. IEEE, 2021.
[5] Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic acids research* 25.17 (1997): 3389-3402.
[6] Oda, Toshiyuki, Kyungtaek Lim, and Kentaro Tomii. "Simple adjustment of the sequence weight algorithm remarkably enhances PSI-BLAST performance." *BMC bioinformatics* 18.1 (2017): 1-8.
[7] Kitts, Paul A., et al. "Assembly: a resource for assembled genomes at NCBI." *Nucleic acids research* 44.D1 (2016): D73-D80.
[8] Eastman, Peter, et al. "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics." *PLoS computational biology* 13.7 (2017): e1005659.
[9] Basu, Sankar, and Björn Wallner. "DockQ: a quality measure for protein-protein docking models." *PloS one* 11.8 (2016): e0161879.
[10] Ghani, Usman, et al. "Improved docking of protein models by a combination of alphafold2 and cluspro." *BioRxiv* (2022): 2021-09.
[11] Remmert, Michael, et al. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment." *Nature methods* 9.2 (2012): 173-175.
[12] http://hmmer.org
[13] Mirdita, Milot, et al. "Uniclust databases of clustered and deeply annotated protein sequences and alignments." *Nucleic acids research* 45.D1 (2017): D170-D176.
[14] https://bfd.mmseqs.com/
[15] UniProt: the universal protein knowledgebase in 2021. *Nucleic acids research*, 2021, 49.D1: D480-D489.
[16] Mitchell, Alex L., et al. "MGnify: the microbiome analysis resource in 2020." *Nucleic acids research* 48.D1 (2020): D570-D578.
[17] Mukherjee, Srayanta, and Yang Zhang. "MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming." *Nucleic acids research* 37.11 (2009): e83-e83.