



Contribution ID: 19

Type: Talk

MassiveFold: optimized massive sampling with AlphaFold2

Wednesday, 14 February 2024 17:15 (30 minutes)

Massive sampling with AlphaFold-multimer(1,2) showed impressive results for structural prediction of macromolecular assemblies at CASP15-CAPRI(3). Generating a very large number of predictions (>1000) and pushing their diversity by playing with the neural network model versions, the number of recycle steps, the use of templates or not and the activation of the dropout in the Evoformer and in the structure module, ranked this method first for the prediction of complexes(4). Subsequently named AFsample(5), the method allows to run massive sampling with AlphaFold's neural network models v1 and v2. We now created MassiveFold, which is based on AFsample and integrates all these diversity parameters, including all the neural network models provided by all the versions of AlphaFold-multimer (v1 to v3). Our tool is optimized to run on a parallel computing CPU/GPU infrastructure as it automatically performs the multiple sequence alignments on CPU and then sends individual structure prediction runs in batches to GPU servers, afterwards gathering all the prediction results to produce a combined ranking. The final results contain many plots including the well-known pLDDT and Predicted Aligned Error plots, but also diagrams and box plots that show the diversity in predictions. MassiveFold allows thus to take full advantage of a CPU/GPU computing infrastructure and to save up to months of calculation with its optimized parallelization feature.

For CAPRI Round 55, we used MassiveFold to compute 6 runs for each target, generating 1005 structures per run. Each run was parameterized with the 3 versions of neural network models, 21 recycles, 0.5 threshold for early stop tolerance on recycling, and of variations in the activation (true/false) of the following parameters: dropout in the Evoformer, dropout in the structure module and/or template activation, totaling 6030 predicted structures for each target. All predictions were ranked following the ipTM+pTM AlphaFold confidence measure. To ensure diversity in the top 5 submitted predictions, the TM-score between the first ranked model and each model was computed, and a K-means clustering was performed on these TM-scores to create 5 clusters. The model with the highest AlphaFold confidence score in each cluster was kept for submission. The last 95 predictions for each target were chosen randomly among the remaining 6025 structures.

1. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).
2. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. 2021.10.04.463034 <https://www.biorxiv.org/content/10.1101/2021.10.04.463034>.
3. Lensink, M. F. et al. Impact of AlphaFold on structure prediction of protein complexes: The CASP15-CAPRI experiment. *Proteins* 91, 1658–1683 (2023).
4. Wallner, B. Improved multimer prediction using massive sampling with AlphaFold in CASP15. *Proteins* (2023) doi:10.1002/prot.26562.
5. Wallner, B. AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinforma. Oxf. Engl.* 39, btad573 (2023).

Submitting to:

8th CAPRI assessment meeting

Primary authors: RAOURAOUA, Nessim (CNRS University of Lille); LENSINK, Marc (CNRS); BRYSSBAERT, Guillaume (CNRS University of Lille)

Presenter: BRYSSBAERT, Guillaume (CNRS University of Lille)

Session Classification: CAPRI