Graph-based chain mapping for comparing protein complexes

Pierre BERRIET[1], Bastien CAZAUX[1], Marc F. LENSINK[2] and Jean-Stéphane VARRÉ[1]

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000, Lille, France

[2] Univ. Lille, CNRS, UMR 8576 Unité de Glycobiologie Structurale et Fonctionnelle (UGSF), F-59000, Lille, France

Corresponding author: pierre.berriet@univ-lille.fr

Background

The rise of protein conformation prediction models, such as AlphaFold [1], based on deep neural networks, has led to a major transformation in the field of structural biology. Tools like MassiveFold [2] enable the modeling of protein complexes by massively sampling AlphaFold-based models and then selecting the most suitable of them, which requires to automatically compare thousands of models.

The comparison of two complexes requires the mapping of each chain in one model to the ones in the other, taking into account their respective positions within the complex. Actual scoring methods require an enumeration of all potential mappings [3], which may prove impossible in many situations because the number of possible mappings is factorial in the number of chains, so several millions for only ten chains. Some tools attempt to reduce the number of possible mappings by taking complex symmetries into account, but this strategy is sensitive to symmetry inaccuracies [4].

Results

The proposed methodology involves the clustering of chains in advance in order to regroup those with similar functions in the complex and thus reduce the number of potential mappings. This clustering is based on both the chain sequence and the spatial information. In order to avoid overinterpretation of a given feature, the clustering process is executed across multiple models. Subsequently, the various possible mappings are enumerated with these constraints using a graph-based representation of the models. This method can be generalized to several models, again to retain only the main topological structure of the complex. This method is tested on MassiveFold predictions obtained during the CASP16 experiment.

Conclusion

This method reduces the number of possible mappings by using the clustering of the sequences as constraints in the enumeration.

References

[1] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583-9.

[2] Raouraoua N, Mirabello C, Véry T, Blanchet C, Wallner B, Lensink MF, et al. MassiveFold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling. Nature Computational Science. 2024 Nov;4(11):824-8. Publisher: Nature Publishing Group.

[3] Mirabello C, Wallner B. DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. Bioinformatics. 2024 Oct;40(10):btae586.

[4] Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Scientific Reports. 2017 Sep;7(1):10480. Publisher: Nature Publishing Group.