MS11 Opportunities from combining structural biology and fold prediction

MS11-04

PDB-wide model validation using deep learning-based predictions of distances and contacts

F. Sanchez Rodriguez [1], G. Chojnowski [2], R. Keegan [3], D. Rigden [1]

[1]Institute of Structural, Molecular and Integrative Biology, University of Liverpool - Liverpool (United Kingdom), [2]European Molecular Biology Laboratory, Hamburg Unit - Hamburg (Germany), [3]UKRI-STFC, Rutherford Appleton Laboratory, Research Complex at Harwell - Didcot (United Kingdom)

Abstract

Structural determination of proteins may be carried out using a range of different techniques, of which Macromolecular X-Ray Crystallography (MX) and cryogenic Electron Microscopy (cryoEM) are currently the two most popular. These experiments typically culminate into the creation of a model that satisfies the experimental observations collected for the structure of interest, which is later deposited in the Protein Data Bank (PDB) (Berman et al. 2000). However, experimental limitations can lead to unavoidable uncertainties during model building resulting in regions that require validation and potentially further refinement. Many metrics are available for model validation, but most are limited to the consideration of the physico-chemical aspects of the model or its match to the map.

Recent developments in the field of evolutionary covariance and machine learning have enabled the precise prediction of residue-residue contacts and increasingly accurate inter-residue distance predictions. Access to this accurate covariance information has played an essential role in the latest advances observed in the field of protein bioinformatics, particularly the improvement of prediction of protein folds by ab initio protein modelling, with the most notable examples being AlphaFold2 (Jumper et al. 2021) and RoseTTAFold (Baek et al. 2021).

Here we attempt to validate all PDB entries solved through cryoEM or MX at resolutions of 3.0-5.0Å, using a series of new methods for model validation based on the availability of accurate inter-residue distance predictions. These new methods include a support-vector machine classifier trained to compare the distance predictions obtained using AlphaFold2 with the distances observed in the protein model in order to detect possible modelling errors. Further analysis of possible sequence register errors is also done by performing an alignment of the predicted residue contact map and the map inferred from the contacts observed in the model. Regions of the deposited model where the maximum contact overlap is achieved through a sequence register different to that observed in the model are flagged and the optimal sequence register can then be used to fix the possible error.

Results obtained for this PDB-wide model validation suggest the presence of possible modelling and sequence register errors among the deposited models that have gone previously unnoticed and can be detected with these new methods.

References

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, et al. 2021. “Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network.” Science 373 (6557): 871–76.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. “The Protein Data Bank.” Nucleic Acids Research 28 (1): 235–42.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” Nature 596 (7873): 583–89.