Ligand-centered assessment of SARS-CoV-2 drug target models in the Protein Data Bank

Alexander Wlodawer[1], Zbigniew Dauter[2], Ivan Shabalin[3,4], Mirosław Gilski[5,6], Dariusz Brzezinski[3,6,7], Marcin Kowiel[6], Władysław Minor[3,4], Bernhard Rupp[8,9], Mariusz Jaskólski[5,6]

[1] Protein Structure Section, Macromolecular Crystallography Laboratory, NCI, Frederick, MD, USA

[2] Synchrotron Radiation Research Section, Macromolecular Crystallography Laboratory, NCI, Argonne National Laboratory, IL, USA

[3] Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, USA

[4] Center for Structural Genomics of Infectious Diseases (CSGID), Charlottesville, VA, USA

[5] Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan, Poland

[6] Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

[7] Institute of Computing Science, Poznan University of Technology, Poznan, Poland

[8] k.-k. Hofkristallamt, San Diego, CA, USA

[9] Institute of Genetic Epidemiology, Medical University Innsbruck, Austria

The spread of the SARS-CoV-2 pandemic has triggered an immediate response of the biomedical community, working to develop treatments and vaccines for COVID-19. Rational drug design against emerging threats like this depends on accurate molecular models of the protein components of the pathogen and of their complexes with candidate drugs for further development. In the current crisis, structural biologists have reacted by presenting in quick succession structure models of SARS-CoV-2 proteins and depositing them in the Protein Data Bank (PDB) [1], which is a global repository of experimental models of biological macromolecules. Since the structures from the first-line research are often deposited before publication, there is an elevated chance of mistakes and errors. This in turn can create confusion among biologists who rely on structural models for understanding the coronavirus, and could in fact impede rather than accelerating drug development. This phenomenon has been recently labeled 'infodemic' by the World Health Organization, which in this case could be rephrased as 'datademic'.

In our ongoing project, we use model-validation metrics, data mining techniques, and expert knowledge to examine the electron density maps of SARS-CoV-2 protein models. Our goal is to help the biomedical community establish a well-validated pool of data. With every weekly update of the PDB, we use text mining to look for deposits related to SARS-CoV-2. Moreover, we track changes to existing deposits, since many authors of SARS-CoV-2 structures have submitted a second or even third version of their coordinates. We also search for raw diffraction data (experimental data upon which the electron density maps and atomic models are based) in the Integrated Resource for Reproducibility in Macromolecular Crystallography [2]. This data is then processed by a set of geometry checking (Molprobity [3]) and statistical (Twilight [4]) validation tools. Finally, expert structural biologists use the mined data and validation reports, and manually inspect the protein models. If errors are spotted, the models are re-refined and made public at our webserver: https://covid-19.bioreproducibility.org. The website also aggregates all the mined data and categorizes the analyzed proteins according to the experimental method, virus type, protein type, and ligand status. We aim to provide the biomedical community with an easy way of finding trusted structural information about concrete parts of the virus.

It is an evolving project. We are working on visualizations connecting genetic information with the virus structure. We will also use our machine learning system [5] to validate small-molecule ligands in the models.

[1] Berman HM, Westbrook J, Feng Z, Glliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28, 235–242

[2] Grabowski M, Langner KM, Cymborowski M, Porebski PJ, Sroka P, Zheng H, Cooper DR, Zimmerman MD, Elsliger MA, Burley SK, Minor W (2016) A public database of macromolecular diffraction experiments. Acta Cryst D72,1181-1193

[3] Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, Verma V, Keedy DA, Hintze BJ, Chen VB et al. (2018) MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci 27, 293-315

[4] Weichenberger CX, Pozharski E & Rupp B (2013). Visualizing ligand molecules in twilight electron density. Acta Cryst F69, 195-200

[5] Kowiel M, Brzezinski D, Porebski PJ, Shabalin IG, Jaskolski M & Minor W (2019) Automatic recognition of ligands in electron density by machine learning. Bioinformatics 35, 452-461

Keywords: proteins; structural biology; structure quality; X-ray crystallography; data mining

This work was supported in part by NIGMS grant R01-GM132595, federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health and the United States Department of Health and Human Services under contract HHSN272201700060C, the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. We acknowledge support of the National Science Center (NCN, Poland), grant No. 2020/01/0/NZ1/00134. The project was also partly funded by the Polish National Agency for Academic Exchange under Grant No. PPN/BEK/2018/1/00058/U/00001 and by the FWF (Austrian Science Foundation) grant P 32821.