BITS :: Call for Abstracts 2023 - Oral communication

Type	Oral communication
Session	Structural Bioinformatics
Title	The importance of data standardization for the analysis of structural files: a case report
All Authors	Bernardina Scafuri (1 *), Nancy D’Arminio (1 *), Deborah Giordano (2 *), Angelo Facchiano(2) , Anna Marabotti (1)

Affiliation

1 Department of Chemistry and Biology “A. Zambelli”, University of Salerno,
Fisciano (SA), Italy
2 National Research Council, Institute of Food Science, Avellino, Italy
* equally contributed to the work

Motivation

Recently, huge advances in structural biology have made it possible to rapidly increase the number of structures in the Protein Data Bank (PDB) archive. Since PDB inception, the files containing all the information related to deposited protein structures have undergone several changes, moving from the original PDB file format (legacy) to the current PDBx/mmCIF file format [1]. Over the years, numerous efforts have been made to overcome the problems related to the original PDB file format, and new standards have been introduced to consider more structural and technical characteristics, arriving at the adoption of the PDBx/mmCIF format, which is more suitable for automated analysis by software tools [2].

Following the COVID-19 pandemic event, the structural biology community has spent a great effort in determining the structures of SARS-CoV-2 proteins. The attention of scientists has been focused on spike protein, whose protein structure is the key to understanding the interaction of SARS-CoV-2 with the human receptor angiotensin-converting enzyme 2 (ACE2) and subsequent entry of coronaviruses into cells [3], and more than 3000 structures of this protein have been determined in a limited amount of time.

Recently, we started a study to predict the influence of mutations of different SARS-CoV-2 variants on antibody binding to the spike protein by in silico analysis [4]. In order to obtain the desired information, we decided to perform an automated analysis of hundreds of structures of the spike protein complexed with the light and heavy chains of different types of antibodies.

Unfortunately, the automated analyses, which were expected to get results quickly, have proved difficult due to several problems encountered in the PDB and PDBx/mmCIF file, dealing in particular with the lack of standardization of the structural information.

We present some of the issues observed in our specific case but these can probably be generalized to many other structures collected in the PDB database

Methods

From PDB Database, we collected 172 spike-Ab complexes. For each complex, we automatically model the omicron variants of Sars-CoV-2, developing in-house a Perl script.

To study the change in the interactions we calculated the H-bonds and hydrophobic interactions between each different spike chain and the antibody in the selected complexes using LigPlot [5] and detected the ionic interactions with an in-house developed Perl script.

We compared the non-mutated interactions with the predicted ones using R programming (https://www.r-project.org).

The difficulties that emerged during the analysis, due to the non-heterogeneity of the information in the files of the complexes, forced us to manually check the individual files before submitting them to the automated analyses.

Results

To predict the effect of mutations on the interaction between spike and antibodies, an essential information is the association of the molecular entity with the chain that identifies it. In the files we analyzed this information is not homogeneous. For example, in some cases the spike protein is described as spike protein S1 but in others as spike glycoprotein or still other ways. The same thing happens with the antibodies’ names [6].

The second issue we dealt with concerned the chain identifiers. Often chain identifiers used for spike protein in some files are used for antibodies in other files. The lack of uniformity of chain identifiers makes it more difficult to automatically extract the information from the output of software analysis performed on hundreds of PDB files, especially if the study is focused on identifying the interaction between two molecules, as in our case [6].

Another aspect that should be taken into consideration is the order in which the chains of the complex are reported in the ATOM records 6]. This is useful because having the same molecule always in the same position in a series of files could help the researchers who use the structures for alignment or modeling.

We acknowledge the important work of harmonization and data FAIRification made by the structural biology community, especially with the introduction of the new PDBx/mmCIF format; however, these problems underline the need to define further rules to standardize the way to report the information contained in these files.

We hope that this case report will suggest further improvements in the way structure-related information is provided to those who will use these files for subsequent analysis.

Info

References

[1] Adams PD, Afonine PV, Baskaran K, Berman HM, Berrisford J, Bricogne G, Brown DG, Burley SK, Chen M, Feng Z, Flensburg C, Gutmanas A, Hoch JC, Ikegawa Y, Kengaku Y, Krissinel E, Kurisu G, Liang Y, Liebschner D, Mak L, Markley JL, Moriarty NW, Murshudov GN, Noble M, Peisach E, Persikova I, Poon BK, Sobolev OV, Ulrich EL, Velankar S, Vonrhein C, Westbrook J, Wojdyr M, Yokochi M, Young JY. Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). Acta Crystallogr D Struct Biol. 2019 Apr 1;75(Pt 4):451-454;

[2] Westbrook JD, Young JY, Shao C, Feng Z, Guranovic V, Lawson CL, Vallat B, Adams PD, Berrisford JM, Bricogne G, Diederichs K, Joosten RP, Keller P, Moriarty NW, Sobolev OV, Velankar S, Vonrhein C, Waterman DG, Kurisu G, Berman HM, Burley SK, Peisach E. PDBx/mmCIF Ecosystem: Foundational Semantic Tools for Structural Biology. J Mol Biol. 2022 Jun 15;434(11):167599;

[3] Papageorgiou AC, Mohsin I. The SARS-CoV-2 Spike Glycoprotein as a Drug and Vaccine Target: Structural Insights into Its Complexes with ACE2 and Antibodies. Cells. 2020 Oct 22;9(11):2343;

[4] D'Arminio N, Giordano D, Scafuri B, Biancaniello C, Petrillo M, Facchiano A, Marabotti A. In Silico Analysis of the Effects of Omicron Spike Amino Acid Changes on the Interactions with Human Proteins. Molecules. 2022 Jul 28;27(15): 4827;

[5] Barnes CO, Jette CA, Abernathy ME, Dam KA, Esswein SR, Gristick HB, Malyutin AG, Sharaf NG, Huey-Tubman KE, Lee YE, Robbiani DF, Nussenzweig MC, West AP Jr, Bjorkman PJ. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. Nature. 2020 Dec;588(7839):682-687. doi: 10.1038/s41586-020-2852-1. Epub 2020 Oct 12. PMID: 33045718; PMCID: PMC8092461;

[6] D'Arminio N, Giordano D, Scafuri B, Facchiano A, Marabotti A. Standardizing macromolecular structure files: further efforts are needed. Trends Biochem Sci. 2023 Apr 6:S0968-0004(23)00078-6.

D.G. was supported within the framework of “CIR01_00017 “CNRBiOmics Centro Nazionale di Ricerca in Bioinformatica per le Scienze Omiche - Rafforzamento del capitale umano” CUP B56J200009600. N.D.A., B.S. and A.M. were supported by UNIVERSITY OF SALERNO, grant numbers ORSA199808, ORSA208455, and ORSA219407; by MIUR, grant FFABR2017 and PRIN 2017 program, grant number: 2017483NH8; and by BANCA D’ITALIA. The corresponding authors are members of the 3D-Bioinfo Community of ELIXIR infrastructure (https://elixir-europe.org/).

filename	-
Figure	
-	
Availability	-

Corresponding Author

Name, Surname	Anna , Marabotti
Email	amarabotti@unisa.it
Submitted on	02.05.2023