Applying Protein Structure Comparison Methods for Studying SARS-CoV-2 Spike Protein

Dobrinka Petrova

Technical University of Sofia, Plovdiv Branch,
25, Tsanko Dyustabanov Str., Plovdiv, Bulgaria

dpetrova@tu-plovdiv.bg

Abstract. The importance of SARS-CoV-2 molecule and especially its spike (S) protein defines the need of studying its structure and functions in deep. Main structural studies of viral fusion proteins, largely limited to X-ray crystallographic analysis for a long time, rely on cryo-electron microscopy nowadays, so the amount of high resolution data of spike protein is yet to growing up. The opportunity of applying computational methods for similarity detection between protein molecules for further analysis of SARS-CoV-2 spike protein, based on these precise-enough data is considered in this paper. Biological background and the research problem definition are presented first, followed by a survey of protein structure comparison methods, which are reportedly used for studying SARS-CoV-2 spike protein. Conclusions are made about future investigations in the area.

INTRODUCTION

People have suffered coronavirus's outbreaks more than once [1]. However, the last one is unprecedented with its worldwide spread and consequences in all spheres of life. At the end of 2019 cases of unidentified pneumonia with clinical characteristics, similar to those of viral pneumonia, were reported in Wuhan, China [2]. The analysis of the cases showed, that the pneumonia is caused by a novel coronavirus [3]. The virus was named "severe acute respiratory syndrome coronavirus 2" (SARS-CoV-2) and the disease – COVID-19. Only three months later, the number of cases outside China increased 13-fold and the number of countries with cases increased 3-fold [4]. The World Health Organization (WHO) on March 11, 2020, has declared the novel coronavirus (COVID-19) outbreak a global pandemic. As of 25 April 2022 there have been more than 507 500 000 confirmed cases, including 6 220 390 deaths [5].

As a coronavirus, SARS-CoV-2 molecule contains four structural proteins: spike (S), envelope (E), membrane (M) and nucleocapsid (N) [6]. A key component is the spike – class 1 viral fusion protein that promotes host attachment and fusion of the viral and cellular membranes during entry [7]. As a consequence, S determines host range and cell tropism. S is also the main target of neutralizing antibodies elicited during infection and the focus of vaccine design [6]. These features of spike protein define its importance and determine the need of studying its structure in deep.

Most structural studies of class 1 viral fusion proteins were largely limited to X-ray crystallographic analysis for a long time [6]. In the past few years, however, technical advances in single-particle cryo-electron microscopy (cryoEM) led to the first structures at high-enough resolution to obtain an atomic model and the amount of precise data is yet to growing up. Based on these data, the opportunity of applying computational methods for similarity detection between protein molecules for further analysis of SARS-CoV-2 spike protein is considered in this paper. Biological background and the research problem definition are presented first, followed by a survey of protein structure comparison methods, which are reportedly used for studying SARS-CoV-2 spike protein.

BIOLOGICAL BACKGROUND

SARS-CoV-2 is a single-stranded RNA-enveloped virus, which is approximately 30000 nucleotide bases in length [8, 9]. Part of them encodes a surface glycoprotein – the spike protein, which binds to the host-cell receptor and mediates the viral entry [10]. These functions of the spike protein, which determine its importance, are supposed to be closely related with its structure.

Proteins, and spike protein in particular, possess a complex 3D structure. Four different structure levels can be distinguished: primary, secondary, tertiary and quaternary.

Primary structure

Proteins are composed of amino acids, bonded together in polypeptide chains. Each amino acid consists of Cα atom, amino group (NH2), carboxylate group (COOH) and a specific side chain (R) – Fig. 1. The primary structure of a protein refers to the sequence of the amino acids in the polypeptide chain.

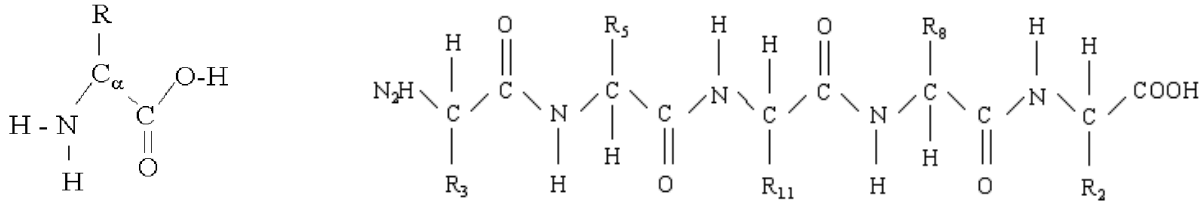

FIGURE 1. (a) Amino acid structure (b) Polypeptide chain

The total length of SARS-CoV-2 S protein is 1273 amino acids and consists of a signal peptide (amino acids 1–13), the S1 subunit (14–685 residues), and the S2 subunit (686–1273 residues); the last two regions are responsible for receptor binding and membrane fusion, respectively [11].

Secondary structure

The chain of the amino acids is not a straight line. It can fold into α-helixes and β-sheets – Fig. 2, due to hydrogen bonds. α-helixes and β-sheets (composed of β-strands) are called secondary structure elements and define the secondary structure of the protein.

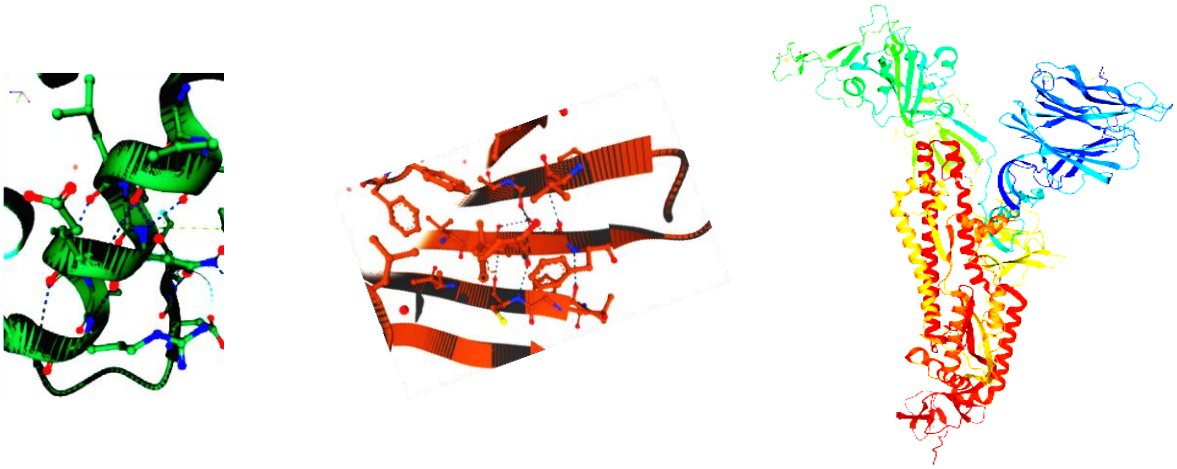

Figure 2. (a) α-helix (b) β-sheet, composed of four strands (c) Structure of SARS-CoV-2 Spike Protein Trimer, chain A – 6ZP2.pdb [12]. All images are components in protein structures, deposited in PDB - rcsb.org [13]

Tertiary and quaternary structure

Secondary structure elements are folded further into a compact structure. When it contains only one protein molecule (polypeptide chain), with possible several domains, it is called tertiary structure – the 3D shape of the protein. When two or more polypeptide chains are bonded together and function as a single unit, the 3D structure of the aggregation is called quaternary structure.

STRUCTURE COMPARISON AND SIMILARITY DETECTION OF SPIKE PROTEINS – PROBLEM SPECIFICS

Three main tasks can be defined in the process of comparison of two protein molecules and evaluation of their structural similarity – Fig. 3:

1. The way of representation of protein structure has to be chosen first – that is the model of the molecule. The geometric properties at preferred structure level are extracted for each protein molecule and they compose the models.
   When the process of study of spike proteins is focused at primary structure level, an enormous number of structure elements (i.e. 1273 amino acids with coordinates of their atoms) has to be considered. In order to decrease the complexity and simplify the presentation of the proteins at this level and its processing, each amino acid can be presented with part of its atoms – for example Cα atoms only.
   When the process of study is focused at secondary structure level, the number of structure elements – helices and sheets or strands, to be considered is 30-40 times less than atoms at primary structure level. The models would be simpler, but not so detailed and accurate.
   Presentations of proteins at tertiary or quaternary structure level would possess higher level of abstraction.
2. Appropriate algorithm has to be applied in order to compare the models.
3. An adequate measure has to be selected at the end of the process to interpret the results of comparison algorithm and to draw a conclusion about the degree of structural similarity between compared protein molecules.

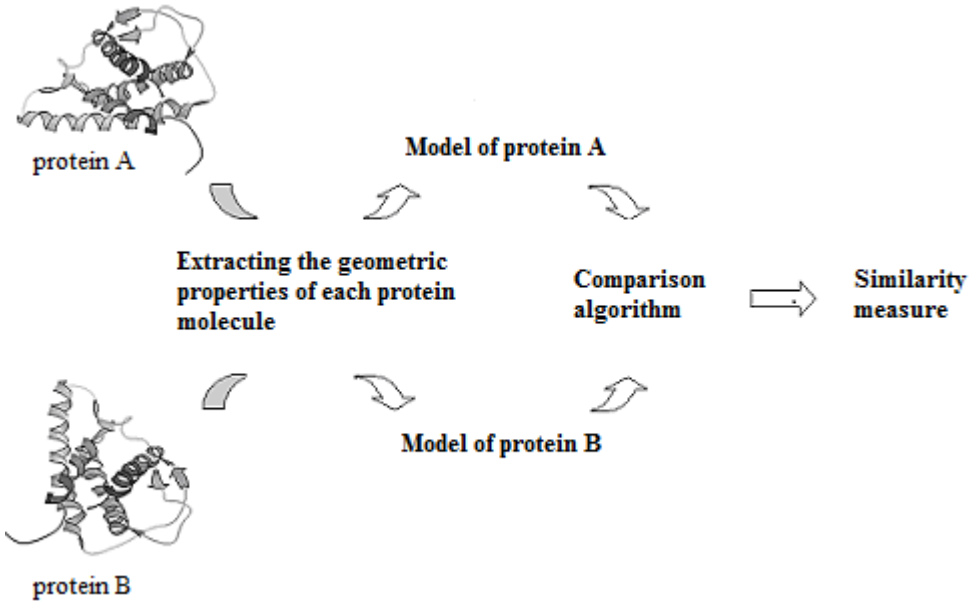

FIGURE 3. The process of structure comparison and similarity detection between protein molecules

PROTEIN STRUCTURE COMPARISON METHODS AND THEIR APPLICATION FOR STUDING SARS-COV-2 SPIKE PROTEIN

Different approaches can be distinguished in protein structure comparison methods, which are reportedly used for studying SARS-CoV-2 spike protein - Fig. 4. Models of the proteins are mainly at primary or secondary

structure level, a rare case is a tertiary structure model. There is also a combined presentation at both – primary and secondary level. Comparison is achieved by superimposing the models or by applying protein structure comparison algorithms. There is also a variety of measures, which are used to evaluate the degree of structural similarity between compared protein molecules.

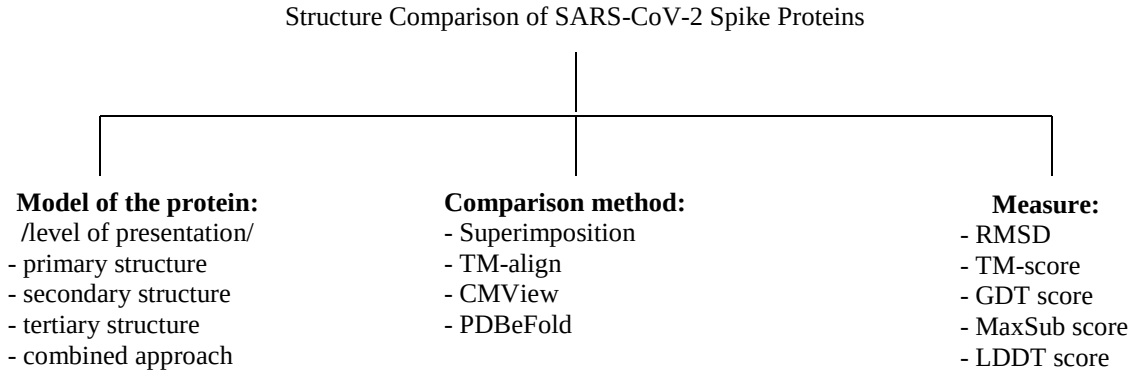

FIGURE 4. Approaches, reportedly used in structure comparison of SARS-CoV-2 spike proteins

Structure Comparison and Similarity Detection of Spike Proteins Using Superimposition

Superimposition is one of the approaches used for structure comparison of spike proteins in research papers considered [14], [15], [16]. Protein molecules are presented at primary (crystal) structure level mainly and models are superimposed. The degree of similarity is evaluated visually or by applying a measure like RMSD [17], TM score [18], MaxSubs score [19] or GDT score [20]. In cases with visual collation, an appropriate viewer is needed [21], [22] for detailed representation of compared proteins.

Analysis of the approach: In the process of similarity detection between protein molecules by superimposition, the comparison method consists of "overlaying" two compared protein structures to find the best fitting between them. The most significant part in the process is the measure, chosen to interpret the results of the fitting and to decree the degree of similarity.

- RMSD – Root-mean-square deviation (RMSD) is one of the preferred quantitative measures for similarity between two superimposed protein structures. It is calculated by eq. 1 and uses the distances between the coordinates of equivalent atoms after superimposition.

$$RMSD(A,B) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}(\| x(i) - y(i) \|^2)} \quad (1)$$

  RMSD is an excellent choice for a measure, when compared protein structures are almost identical [23], but even small differences decrease its effectivity. The average value of RMSD depends on the length [24] and the resolution [25] of compared structures. It is also strongly affected by the most deviated fragments in compared structures.
- TM score and GDT score overcome some of the dependencies of RMSD. TM score is a scoring function, which overcomes the length dependency of RMSD and it is suitable also for comparing proteins with different lengths. GDT score can be applied successfully for structures with deviated fragments.
- Visual comparison of protein structures can be used for similarity detection between proteins with high degree of similarity, as well as can be applied for studying visually in details the small differences between them.

Superimposition is an approach, which is reportedly applied for comparison of protein molecules at primary structure level mainly, rare at secondary structure level. Regardless of the level of the model or the similarity measure used, superimposition can be effectively applied for comparison of protein molecules only if there are some known pivots at the beginning of the process, which are supposed to be corresponding elements (atoms, helices or sheets) at the end of the comparison.

Studying Spike Proteins with Protein Structure Comparison Algorithms

Generally, in the process of studying SARS-Cov-2 spike protein by similarity detection, tools for protein structure analysis are preferred [14], [26], [27], [28], [29], based on algorithms for protein structure comparison. Three of them stand out for comparison of spike proteins in considered research papers – TM-align [30], CMView [31] and PDBeFold [32].

Comparative analysis: A summary of comparative analysis of these tools, especially their protein structure comparison features, which are used in studying spike proteins, is presented in Table 1.

TABLE 1. Comparative analysis of the algorithms

Tool Name	Level of Presentation	Basic Data Structure for Presentation	Techniques and Approaches, Used in Comparison	Complexity of the Problem
TM-align	Both – secondary and primary structure	2D array	Dynamic programming and heuristic iterative alignment	NP
CMView	Primary structure/ Combined approach	2D array	Branch and bound approach / Softassign and dynamic programming	NP
PDBeFold	Both – secondary and primary structure	Graph	Graph matching algorithm and fast optimal superposition	NP

All three tools propose features, which present compared protein structures at both – secondary and primary structure level. TM-align employs the coordinates of backbone C_α atoms of the given protein structures in conjunction with secondary structure elements. The distances between structural elements are modelled with two-dimensional arrays – distance score matrices, secondary structures score matrices or combination of both. CMView also uses two-dimensional arrays for models – distance matrices and contact maps. A contact map [33] is defined as a matrix, filled with all distances between pairs of preferred structural elements to be compared - C_α atoms, amino acids or secondary structure elements. PDBeFold differs in basic data structure type – it presents proteins with graphs, built on the protein's secondary-structure elements. α-helixes and β-strands are presented as vertices in the graph model. An edge between two vertices is labelled with a property vector, composed of edge length and angles, which define mutual positions and orientations of all vertices in the graph.

After constructing, the models of the proteins have to be compared and the tasks are known as “distance matrices alignment problem”, “contact map overlap problem” and “maximum common subgraph problem”, respectively. All these problems are proven to be NP-complete [34]. It is a “challenge” for comparison algorithm to find an optimal solution in an acceptable time. Different techniques and approaches are applied to solve the problem. TM-align uses three kinds of quickly identified initial alignments at the beginning. The first initial alignment is between the secondary structures (SS) of two proteins using dynamic programming [35]. The second type of initial alignment is based on the gapless matching of two structures. The third initial alignment is also obtained by dynamic programming, but the score matrix is a combination of the SS score matrix and the distance score matrix selected in the second initial alignment. The above-obtained initial alignments are submitted to a heuristic iterative algorithm. CMView aligns distance matrices by applying Branch and bound approach [36] or contact maps – by softassign and dynamic programming [37]. PDBeFold uses graph matching algorithm, the results from which are set as a starting point of an iterative three dimensional alignment of protein backbone C_α atoms.

Comparison of protein molecules at atomic level (primary structure level), which can be found as a feature in all three comparison algorithms, discussed above, has a significant advantage: the models of the proteins can be classified as detailed and accurate, with plenty of structural information. The huge amount of information (1273 C_α atoms or distances between them), however, is hard to be processed in an acceptable time, because the problem is NP-complete. In order to speed up the comparison, different approaches are used: heuristic algorithms are used or some initial alignments at higher level of abstraction are applied first.

Models at secondary structure level, on the other hand, are more compact and easy for construction and processing. However, the structural information on this level only, is not sufficient to achieve accurate comparison. This is also a reason, algorithms to use comparison at secondary structure level as initial fast steps followed by refinement of comparison at primary structure level. To increase the accuracy of the model, the geometric properties

of the molecules – mutual positions and orientations of the modelled structural elements, distances, angles, overlapping, have to be considered.

Application of Structure Comparison and Similarity Detection in Studying SARS-CoV-2 Spike Protein

Protein structure comparison methods, discussed above, give a solution to one basic task when studying protein molecules. Generally, the degree of similarity between two compared proteins can be determining in structural analysis and classification of proteins, can be applied for model assessment in protein structure prediction, can help in structure-based functional analysis or drug design. In particular, quality assessment of predicted protein structure models and structural analysis of spike proteins can be distinguished as main applications of the discussed protein structure comparison methods in research papers considered [14], [15], [16], [26], [27], [28], [29].

- **Quality assessment of predicted protein structure**: due to technical and time limitations, the amount of high resolution data of spike protein is yet to growing up. So, an essential task at the beginning of the process of structural investigations of spike protein often is prediction of its unknown secondary and tertiary structure, based on the sequence of the amino acids, which compose it. The predicted model is compared with an experimental structure and the quality of the prediction is evaluated by the degree of similarity between them. In most cases, quality assessment is achieved by comparing predicted and experimental structure models (which are supposed to be highly similar) with superimposition and evaluating the degree of similarity with a measure like like RMSD, TM-score, MaxSubs score, GDT score or superimposition-free LDDT score [38]. TM-align and CMView are also used for quality assessment of predicted models [27].
- **Structural analysis of spike proteins: a**ccording to [16], clustering analysis based on structural similarity between SARS-CoV-2 strains could reflect the current characteristics of the pandemic more accurately than those based on the protein sequence. The study suggests that structural similarity can be a new way to classify SARS-CoV-2 strains and applies comparison methods in conjunction with similarity measures like RMSD and TM-score for that purpose. In [14] comparison of the SARS CoV 2 spike protein with other human-infecting coronaviruses is based on superimposition, TM-align and CMView. In [28] receptor binding motifs of SARS CoV 2 spike protein are compared by TM-align and PDBeFold.
- Protein structure comparison algorithms are also included in researches, dedicated to drug screening and design [26].

CONCLUSION

Despite the small amount of high resolution data of SARS-CoV-2 spike proteins at the atomic level, there are researches in the area, in which protein structure comparison methods are applied in order to study the S protein. Proposed approaches present spike proteins at primary, secondary, rare tertiary or combined (primary and secondary) structure level and use superimposition or protein structure comparison algorithms to detect similarity between compared proteins. All protein structure comparison methods, chosen for studying SARS-CoV-2 spike protein in research papers considered, can be described as well known, online available tools with set of options the user to select from. So, two directions for future work in the area can be defined: 1) protein structure comparison method to be proposed, in which balance is found between precise enough model of protein molecules, fast and accurate comparison algorithm and adequate measure for similarity detection; and 2) a complex tool for protein structure analysis to be developed, based on proposed method, with appropriate features and user-friendly interface for scientists to work with.

ACKNOWLEDGMENTS

The author would like to thank the Research and Development Sector at the Technical University of Sofia for the financial support.

REFERENCES

1. Standl, F.; Jöckel, K.-H.; Brune, B.; Schmidt, B.; Stang, A. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *Lancet Infect. Dis.* **2020.**
2. Huang C , Wang Y , Li X , Ren L , Zhao J , Hu Y , et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395:497–506.
3. Lisheng Wang, Yiru Wang, Dawei Ye, Qingquan Liu, Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence, International Journal of Antimicrobial Agents, Volume 55, Issue 6, 2020, 105948, ISSN 0924-8579, https://doi.org/10.1016/j.ijantimicag.2020.105948.
4. Cucinotta D, Vanelli M. WHO Declares COVID-19 a Pandemic. Acta Biomed. 2020 Mar 19;91(1):157-160. doi: 10.23750/abm.v91i1.9397. PMID: 32191675; PMCID: PMC7569573.
5. WHO Coronavirus (COVID-19) Dashboard: https://covid19.who.int/
6. M. Alejandra Tortorici, David Veesler, Chapter Four - Structural insights into coronavirus entry, Editor(s): Félix A. Rey, Advances in Virus Research, Academic Press, Volume 105, 2019, Pages 93-116, ISSN 0065-3527, ISBN 9780128184561, https://doi.org/10.1016/bs.aivir.2019.08.002.
7. Bosch BJ, van der Zee R, de Haan CA, Rottier PJ. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol. 2003 Aug;77(16):8801-11. doi: 10.1128/jvi.77.16.8801-8811.2003. PMID: 12885899; PMCID: PMC167208.
8. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020;395:565–74.
9. Huang, Y., Yang, C., Xu, Xf. et al. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin 41, 1141–1149 (2020). https://doi.org/10.1038/s41401-020-0485-4.
10. Li, F. Structure, function, and evolution of coronavirus spike proteins. Annu. Rev. Virol. 3, 237–261 (2016).
11. Huang, Y., Yang, C., Xu, Xf. et al. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin 41, 1141–1149 (2020). https://doi.org/10.1038/s41401-020-0485-4.
12. Xiong, X., Qu, K., Ciazynska, K.A. et al. A thermostable, closed SARS-CoV-2 spike protein trimer. Nat Struct Mol Biol 27, 934–941 (2020). https://doi.org/10.1038/s41594-020-0478-5
13. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.
14. Cueno ME, Imai K. Structural Comparison of the SARS CoV 2 Spike Protein Relative to Other Human-Infecting Coronaviruses. Front Med (Lausanne). 2021 Jan 14;7:594439. doi: 10.3389/fmed.2020.594439. PMID: 33585502; PMCID: PMC7874069.
15. Chengxin Zhang, Wei Zheng, Xiaoqiang Huang, Eric W. Bell, Xiaogen Zhou, and Yang Zhang, Protein Structure and Sequence Reanalysis of 2019-nCoV Genome Refutes Snakes as Its Intermediate Host and the Unique Similarity between Its Spike Protein Insertions and HIV-1, Journal of Proteome Research 2020 19 (4), 1351-1360, DOI: 10.1021/acs.jproteome.0c00129
16. Qiangzhen Yang, Xuemin Jian, Ali Alamdar Shah Syed, Aamir Fahira, Chenxiang Zheng, Zijia Zhu, Ke Wang, Jinmai Zhang, Yanqin Wen, Zhiqiang Li, Dun Pan, Tingting Lu, Zhuo Wang, Yongyong Shi, "Structural Comparison and Drug Screening of Spike Proteins of Ten SARS-CoV-2 Variants", Research, vol. 2022, Article ID 9781758, 20 pages, 2022. https://doi.org/10.34133/2022/9781758
17. Eidhammer I, Jonassen I, Taylor WR. Structure comparison and structure patterns. J Comput Biol. 2000;7(5):685-716. doi: 10.1089/106652701446152. PMID: 11153094.
18. Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," Proteins, vol. 57, no. 4, pp. 702–710, 2004.
19. N. Siew, A. Elofsson, L. Rychlewski, and D. Fischer, "MaxSub: an automated measure for the assessment of protein structure prediction quality," Bioinformatics, vol. 16, no. 9, pp. 776–785, 2000.
20. A. Zemla, "LGA: a method for finding 3D similarities in protein structures," Nucleic Acids Research, vol. 31, no. 13, pp. 3370–3374, 2003.
21. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/
22. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
23. Cohen, FE, Sternberg, MJE, On the prediction of protein structure – the significance of the Root-Mean-Square Deviation, J Mol Biol, 1980, Vol: 138, p. 321 - 333, ISSN: 0022-2836.

24. Betancourt, M.R, J. Skolnick. Universal similarity measure for comparing protein structures. Biopolymers, 2001, p. 305–309.
25. Carugo O. How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. J. Appl. Crystallogr., 2003, p. 125–128.
26. Siqin Feng, Xiaodong Luan, Yifei Wang, Hui Wang, Zhiyu Zhang, Yiyang Wang, Zhuang Tian, Meixi Liu, Ying Xiao, Yong Zhao, Ruilin Zhou, Shuyang Zhang, Eltrombopag is a potential target for drug intervention in SARS-CoV-2 spike protein, Infection, Genetics and Evolution, Volume 85, 2020, 104419, ISSN 1567-1348,
27. Shishir TA, Jannat T, Naser IB (2022) An in-silico study of the mutation-associated effects on the spike protein of SARS-CoV-2, Omicron variant. PLOS ONE 17(4): e0266844. https://doi.org/10.1371/journal.pone.0266844.
28. Beaudoin CA, Jamasb AR, Alsulami AF, Copoiu L, van Tonder AJ, Hala S, Bannerman BP, Thomas SE, Vedithi SC, Torres PHM, Blundell TL. Predicted structural mimicry of spike receptor-binding motifs from highly pathogenic human coronaviruses. Comput Struct Biotechnol J. 2021 Jul 2;19:3938-3953. doi: 10.1016/j.csbj.2021.06.041. PMID: 34234921; PMCID: PMC8249111.
29. Wu S, Tian C, Liu P, Guo D, Zheng W, Huang X, Zhang Y, Liu L. Effects of SARS-CoV-2 mutations on protein structures and intraviral protein-protein interactions. J Med Virol. 2021 Apr;93(4):2132-2140. doi: 10.1002/jmv.26597. Epub 2020 Nov 1. PMID: 33090512; PMCID: PMC7675365.
30. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. (2005) 33:2302–9. doi: 10.1093/nar/gki524
31. Corinna Vehlow, Henning Stehr, Matthias Winkelmann, José M. Duarte, Lars Petzold, Juliane Dinse, Michael Lappe, CMView: Interactive contact map visualization and analysis, Bioinformatics, Volume 27, Issue 11, 1 June 2011, Pages 1573–1574, https://doi.org/10.1093/bioinformatics/btr163
32. Krissinel E., Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr Sect D Biol Crystallogr. 2004;60(12):2256–2268. doi: 10.1107/S0907444904026460
33. Sonnhammer, E.L., Wooton, J. C., Dynamic contact maps of protein structures. 1998, Journal of Molecular Graphics and Modelling, Vol.16, p. 1-5.
34. Kolodny R., Koehl P., M. Levitt. Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures. J. Mol. Biol., 2005, p. 1173-1188.
35. Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 1970;48:443–453.
36. Holm L., C. Sander, Protein structure comparison by alignment of distance matrices, J Mol Biol., 1993 Sep 5;233 (1):123-38 8377180.
37. Jain, Brijnesh J.; Lappe, Michael. Joining softassign and dynamic programming for the contact map overlap problem. In: International Conference on Bioinformatics Research and Development. Springer, Berlin, Heidelberg, 2007. p. 410-423.
38. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;vol. 29(21), p: 2722-2728. doi:10.1093/bioinformatics/btt473.