CoV3D: A database of high resolution coronavirus protein structures

Ragul Gowthaman[1,2], Johnathan D. Guest[1,2], Rui Yin[1,2], and Brian G. Pierce[1,2,*]

[1]University of Maryland Institute for Bioscience and Biotechnology Research, Rockville, MD 20850, USA

[2]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA

*To whom correspondence should be addressed. Email: pierce@umd.edu

ABSTRACT

Coronaviruses represent a major global threat as emerging pathogens, most recently SARS-CoV-2, the etiologic agent behind COVID-19, which is currently causing a global pandemic. The urgent need for effective vaccines and therapies is leading to a rapid rise in the number of high resolution structures of SARS-CoV-2 proteins that collectively reveal critical targets on the virus and a map of vulnerabilities. To enable structure-based design of vaccines and therapeutics, as well as insights into structural and molecular features of coronaviruses, we have developed the CoV3D database, which provides a view of SARS-CoV-2 and other coronavirus protein structures, updated on a weekly basis. This database includes and annotates structures of coronavirus proteins and their complexes with antibodies, receptors, and small molecules. Additionally, CoV3D provides information on spike glycoprotein sequence variability and polymorphisms, mapping these features onto the spike structure, and includes spike structures with modeled glycosylation. CoV3D is available at: https://cov3d.ibbr.umd.edu.

INTRODUCTION

Coronaviruses have been responsible for several outbreaks over the past two decades, including SARS-CoV in 2002-2003, MERS-CoV in 2012 (de Wit et al., 2016), and the current COVID-19 pandemic, caused by SARS-CoV-2, which began in late 2019 (Tse et al., 2020). The scale of the COVID-19 pandemic has led to unprecedented efforts by the research community to rapidly identify and test therapeutics and vaccines, and to understand the molecular basis of SARS-CoV-2 entry, pathogenesis, and immune targeting.

Between February and April 2020, a large number of structures have been generated and deposited in the Protein Data Bank (PDB) (Rose et al., 2011): 12 spike glycoprotein structures, over 100 main protease structures, and 23 structures of SARS-CoV-2 non-structural proteins (NSPs). These high resolution protein structures are of immense importance for understanding viral assembly and for rational vaccine design. The first structures of SARS-CoV-2 trimeric spike glycoproteins were reported in February and early March 2020 (Walls et al., 2020; Wrapp et al., 2020). These are the major target of SARS-CoV-2 vaccines and antibody therapeutics, and previously determined spike glycoprotein structures have enabled advances including rational design of CoV spike glycoproteins with proline substitutions to stabilize the prefusion conformation, yielding improved protein expression and immunogenicity (Pallesen et al., 2017). Given that the rapid rate of coronavirus protein structural determination and deposition is likely to continue, and their importance to the research community, a simple and updated resource detailing these structures would provide a useful reference.

Here we describe a new database of experimentally determined coronavirus protein structures, CoV3D. CoV3D is updated automatically on a weekly basis, as new structures are released in the PDB. Structures are classified by CoV protein, as well as bound molecule, such as monoclonal antibody, receptor, and small molecule ligand. To enable insights into the spike glycoprotein, we also include information on SARS-CoV-2 residue polymorphisms, overall coronavirus sequence diversity of betacoronaviruses

mapped onto spike glycoprotein structures, and structures of spike glycoproteins with modeled glycosylation. This resource can enable efforts in rational vaccine design, targeting by immunotherapies, biologics, and small molecules, in addition to basic research into coronavirus structure and recognition, and is publicly available at https://cov3d.ibbr.umd.edu.

MATERIALS AND METHODS

CoV3D is implemented using the Flask web framework (https://flask.palletsprojects.com/) and the SQLite database engine (https://www.sqlite.org/). Structures are identified from the PDB on a weekly basis using NCBI BLAST command line tools (Camacho et al., 2009), with coronavirus protein reference sequences from SARS-CoV, MERS-CoV, and SARS-CoV-2 as queries. The spike glycoprotein reference sequences (GenBank identification NP_828851.1, YP_009047204.1 and QHD43416.1 for SARS-CoV, MERS-CoV and SARS-CoV-2 virus respectively) were used as queries to identify all available spike glycoprotein structures. Peptide-MHC structures containing coronavirus peptides are identified in the PDB through semi-manual searches of the PDB site and literature, though future automated updates are planned in conjunction with an expanded version of the TCR3d database (Gowthaman and Pierce, 2019). Structural visualization is performed using NGL viewer (Rose and Hildebrand, 2015). SARS-CoV-2 spike glycoprotein sequences and sequence information were downloaded from NCBI Virus (Hatcher et al., 2017), followed by filtering out sequences with missing residues. Sequence polymorphism information was obtained by BLAST search using a reference SARS-CoV-2 spike glycoprotein sequence (QHD43416.1). To develop spike glycoprotein alignments, betacoronavirus spike glycoprotein sequences were downloaded from NCBI Virus and aligned with Clustal Omega (Sievers et al., 2011) in SeaView (Gouy et al., 2010). Sequences that were redundant (>95% similarity) or contained missing residues were removed, with the remaining 70 sequences forming the Pan-betacoronavirus alignment. A subset of 18 sequences from the pan-betacoronavirus alignment was used to generate the SARS-like sequence alignment, which contains every sequence from the pan-betacoronavirus alignment with >70% sequence similarity to the SARS-CoV-2 spike. Sequence logos are generated dynamically for user-specified residue

ranges using the command-line version of WebLogo (Crooks et al., 2004). N-glycans are modeled onto spike glycoprotein structures using a glycan modeling and refinement protocol in Rosetta (Labonte et al., 2017). An example command line and Rosetta Script for this glycan modeling protocol is provided as Supplemental Information.

RESULTS

Database Contents

The main components of the CoV3D database are interrelated tables, datasets and tools for coronavirus protein structures and spike glycoprotein sequences. The structure portion of the database includes dedicated pages and tables for:

- Spike glycoprotein structures
- Main protease structures
- Other coronavirus protein structures
- Peptide-Major Histocompatibility Complex (MHC) structures
- Spike structures with modeled glycans

Spike glycoprotein structures are annotated by binding partner(s), including bound antibody and receptor, as well as domain present in the structure. Main protease structures include annotation of bound ligand, for convenience of those investigating protease inhibitors computationally or experimentally. The sequence portion of CoV3D is focused on spike glycoprotein variability, and includes:

- Spike glycoprotein residue polymorphisms
- Spike glycoprotein sequence logos based on sets of betacoronavirus sequences
- Visualization of conservation on spike glycoprotein structures

Additional features of the database include a page showing all recently released structures, and a downloads page for users to download structural data, sequence data, and tables. Summary statistics are

also given, providing metrics such as number of spike structures released for SARS-CoV, MERS-CoV, and SARS-CoV-2.

Example use case: viewing and conservation of RBD site

The spike glycoprotein structure table on CoV3D includes structures of the SARS-CoV-2 spike in complex with antibodies and the receptor ACE2. One of these is the structure of SARS-CoV-2 spike receptor binding domain (RBD) in complex with human antibody CR3022 (PDB code 6W41) (Yuan et al., 2020), which can be visualized in the browser using the "View" link in the table (**Figure 1A**). Inspection of this complex shows one contiguous region of the spike RBD that makes multiple contacts with the CR3022 antibody (residues 375-390; circled in **Figure 1A**). CR3022-contactiong residues include K378 (lysine), P384 (proline), and K386 (lysine). The sequence logo generator on the CoV3D site can then be used to generate logos representing that sequence range for SARS-like coronaviruses (**Figure 1B**) as well as a broader set of betacoronaviruses, including SARS-CoV, SARS-CoV-2, and MERS-CoV (**Figure 1C**). The SARS-like logo highlights the high conservation of this region of the spike glycoprotein, providing a mechanistic basis for the cross-reactive binding exhibited by CR3022 for SARS-CoV and SARS-CoV-2 RBDs (Yuan et al., 2020). However, aside from the cysteine residue at position 379, there is much lower conservation of this region across the broader set of betacoronaviruses (**Figure 1C**).

Example use case: location of polymorphisms

One useful feature of CoV3D is readily available information on identified SARS-CoV-2 spike polymorphisms, and their mapping onto the spike glycoprotein structures. Under "Sequences", users can navigate to "Spike Polymorphisms" where a table is shown with observed single and multiple substitutions in the spike glycoprotein, along with the counts of sequences containing them. The D614G variant is currently the most prevalent, with more occurrences than the reference sequence represented in spike glycoprotein structures (923 sequences, versus 637 sequences). The position of this substitution on

the spike glycoprotein can provide some indication of its possible structural and functional impact. By clicking "view" next to this substitution, users can visualize the position of this substitution, showing that it is located on the spike surface, and is not located within or adjacent to the receptor binding domain (RBD; **Figure 2A**). However, this site is closer to the RBD than the site of another less prevalent SARS-CoV-2 spike variant (T791I; **Figure 2B**).

Example use case: visualizing oligomannose glycosylation

N-glycosylation of viral glycoproteins can play a key role by masking the glycoprotein from the immune system, or effecting function. Experimentally reported protein structures often lack full N-glycans due to limitations from resolution or intrinsic glycan dynamics or heterogeneity. To enable visualization and additional analysis or modeling of glycosylated spike glycoproteins, CoV3D includes sets of structures with modeled N-glycans at all predicted glycosylation sites, with N-glycans built onto the glycoprotein structures and refined using a protocol in Rosetta (Labonte et al., 2017). Examples of glycosylated structures that can be visualized in CoV3D are shown in **Figure 3**, and these can be downloaded directly by users for further processing. This permits users to view features such as the amount of N-glycosylation present on the ACE2 surface (**Figure 3A**) and the relatively high glycosylation of the spike glycoprotein base (bottom in **Figure 3B**) versus the RBD. Presently, these structures include oligomannose glycans that were found to be prevalent on the SARS-CoV spike based on previous mass spectroscopy analysis (Ritchie et al., 2010), with five branched mannose sugars. In the future, we plan to include more glycosylated structures, with additional options for glycan sizes and types.

DISCUSSION

We have constructed the CoV3D database as a reference for the research community, providing a simple and updated interface to high resolution coronavirus 3D structures and sequence variability. This will allow researchers to identify and classify new coronavirus protein structures as they are released, particularly for SARS-CoV-2 and COVID-19, and it can enable insights into sequence features,

polymorphisms, as well as glycosylation. A recent study combining coronavirus structural and sequence analysis revealed insights regarding the determinants of ACE2 recognition (Wan et al., 2020), and CoV3D can enable prospective comparative studies, in addition to modeling and structure-based design efforts. This database will likely be of interest to virologists, computational biologists, immunologists, and those interested in learning about and targeting SARS-CoV-2 proteins with small molecules and antibodies, as well as those engaged in vaccine design.

ACKNOWLEDGEMENTS

We are grateful to the structural biologists and researchers whose work resulted in the structures enabling the development of this database. We are additionally thankful to Ghazaleh Taherzadeh and Stefan Ivanov (University of Maryland Institute for Bioscience and Biotechnology Research), as well as the RosettaCommons community, for helpful discussions and suggestions. The Institute for Bioscience and Biotechnology Research computing facility and staff, including Christian Presley, provided resources and assistance with server implementation and web hosting.

FUNDING INFORMATION

This work was supported in part by NIH R01 GM126299 (to BGP).

REFERENCES

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC bioinformatics *10*, 421.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

de Wit, E., van Doremalen, N., Falzarano, D., and Munster, V.J. (2016). SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol *14*, 523-534.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol *27*, 221-224.

Gowthaman, R., and Pierce, B.G. (2019). TCR3d: The T cell receptor structural repertoire database. Bioinformatics *35*, 5323-5325.

Hatcher, E.L., Zhdanov, S.A., Bao, Y., Blinkova, O., Nawrocki, E.P., Ostapchuck, Y., Schaffer, A.A., and Brister, J.R. (2017). Virus Variation Resource - improved response to emergent viral outbreaks. Nucleic Acids Res *45*, D482-D490.

Labonte, J.W., Adolf-Bryfogle, J., Schief, W.R., and Gray, J.J. (2017). Residue-centric modeling and design of saccharide and glycoconjugate structures. J Comput Chem *38*, 276-287.

Pallesen, J., Wang, N., Corbett, K.S., Wrapp, D., Kirchdoerfer, R.N., Turner, H.L., Cottrell, C.A., Becker, M.M., Wang, L., Shi, W., *et al.* (2017). Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. Proc Natl Acad Sci U S A *114*, E7348-E7357.

Ritchie, G., Harvey, D.J., Feldmann, F., Stroeher, U., Feldmann, H., Royle, L., Dwek, R.A., and Rudd, P.M. (2010). Identification of N-linked carbohydrates from severe acute respiratory syndrome (SARS) spike glycoprotein. Virology *399*, 257-269.

Rose, A.S., and Hildebrand, P.W. (2015). NGL Viewer: a web application for molecular visualization. Nucleic Acids Res *43*, W576-579.

Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D., *et al.* (2011). The RCSB Protein Data Bank: redesigned web site and web services. Nucleic acids research *39*, D392-401.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., *et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol *7*, 539.

Tse, L.V., Meganck, R.M., Graham, R.L., and Baric, R.S. (2020). The Current and Future State of Vaccines, Antivirals and Gene Therapies Against Emerging Coronaviruses. Frontiers in microbiology *11*, 658.

Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell *181*, 281-292 e286.

Wan, Y., Shang, J., Graham, R., Baric, R.S., and Li, F. (2020). Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. J Virol *94*.

Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S., and McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science *367*, 1260-1263.

Yuan, M., Wu, N.C., Zhu, X., Lee, C.D., So, R.T.Y., Lv, H., Mok, C.K.P., and Wilson, I.A. (2020). A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science *368*, 630-633.

FIGURE LEGENDS

Figure 1. Viewing the structure and sequence diversity of a key site on SARS-CoV-2 spike receptor binding domain (RBD). (A) Visualization of the spike RBD complex with antibody CR3022 (PDB code 6W41) in CoV3D, with spike colored blue, CR3022 red and yellow, and an interacting region of interest (residues 375-390) circled for reference. (B) Sequence logo for residues 375-390 using 17 SARS-related spike glycoprotein reference sequences. (C) Sequence logo for residues 375-390 using 70 pan-betacoronavirus spike glycoprotein reference sequences.

Figure 2. Visualization of SARS-CoV-2 spike polymorphisms in CoV3D. Shown are (A) residue 614, the site of the D614G substitution, and (B) residue 791, site of the T791I substitution. Polymorphisms are shown in the context of the prefusion spike structure (Wrapp et al., 2020) (PDB code 6VSB) (gray), with sites in spacefill representations of the wild-type residue on each spike trimer subunit, colored by atom type (cyan: carbon, blue: nitrogen, red: oxygen). Black dashed circles denote the location of the receptor binding domains (residues 319-529) and red dashed circles denote mutation sites, for reference.

Figure 3. Spike structures with modeled oligomannose N-glycans. (A) The SARS-CoV-2 spike RBD-ACE2 complex (PDB code 6LZG) with RBD and ACE2 shown as red and blue cartoons, respectively, with modeled N-glycans shown as gray, red, and blue spheres. One N-glycan is modeled on the spike RBD and six are modeled on ACE2. (B) The spike trimer structure (PDB code 6VSB) with spike shown as gray surface and modeled N-glycans shown as gray, red, and blue spheres. 47 modeled N-glycans are present on the spike structure.

Figure 1.

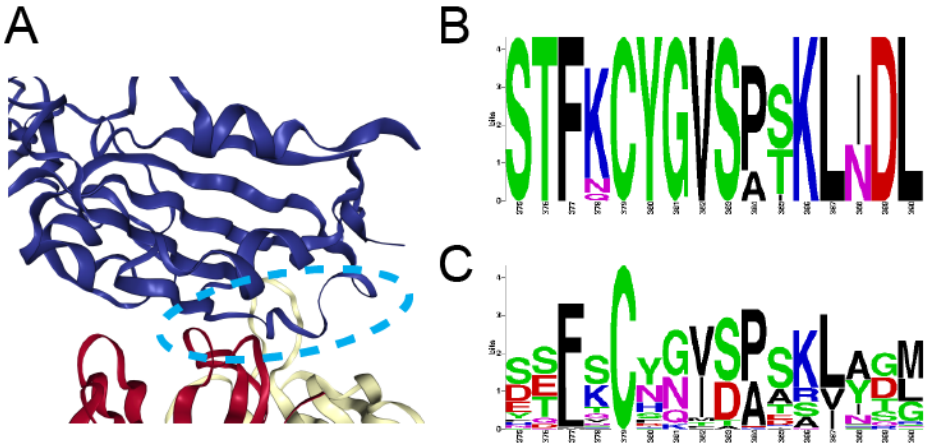

Figure 2.

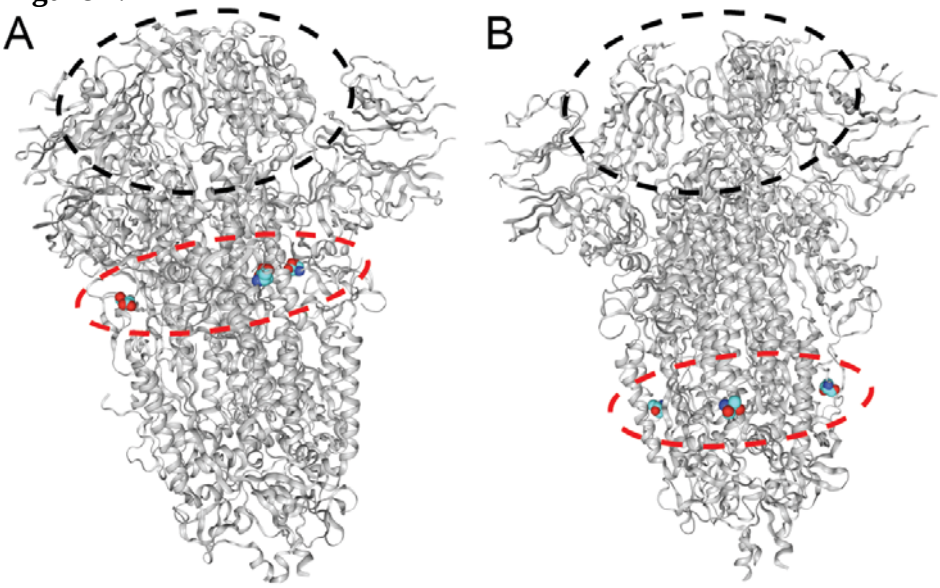
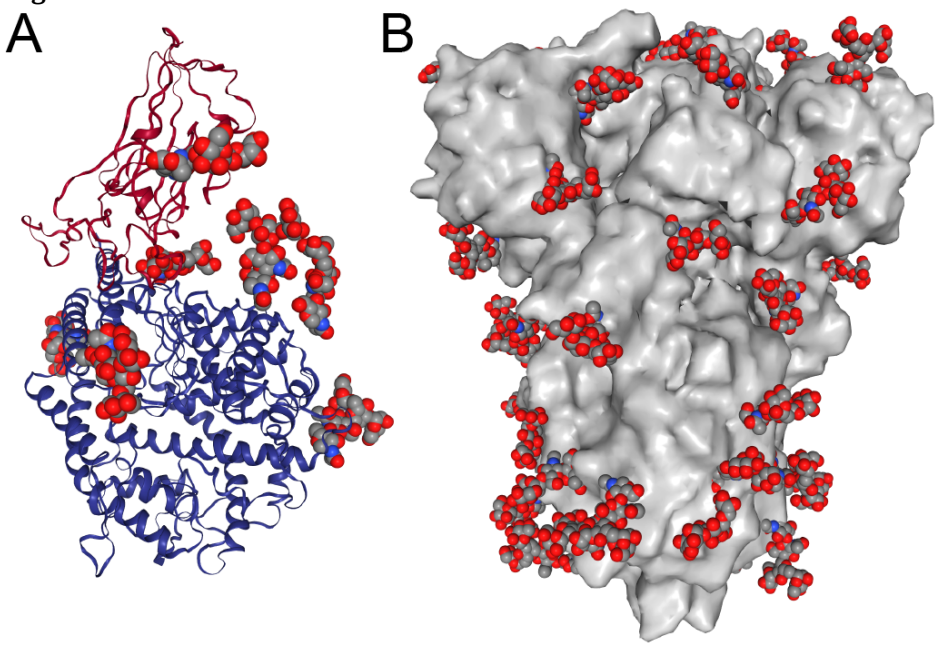
Figure 3.