**Functional conservation analysis of the novel coronavirus SARS-Cov-2 main protease Mpro**

Manjot Hunjan

20622850

**Introduction**

Currently, the novel coronavirus severe acute respiratory syndrome coronavirus (SARS-Cov) presents a viral epidemic and as such, it is of importance to understand its molecular inner workings. The main protease (Mpro) is a key target for drug therapy given its critical importance in coronavirus replication. The present study aims to analyze the key domains of Mpro and their conservation in order to understand if it is functionally like other coronaviruses. This would lead to a better understanding of its mechanisms, and lead drug design when developing therapies or vaccines.

Recently, whole genome sequencing of several isolates in various countries have been conducted, leading to further understanding of the genomic organization, and viral structure. Classically, coronaviruses are characterized as enveloped positive single-stranded RNA viruses wrapped in a nucleocapsid protein inside of an envelope (Medhi et al., 2020). Since this is a novel species, in order to understand its functional conservation, it is relevant to understand the typical coronavirus. Firstly, there are 4 genera in this virus family being α, β, γ and δ-coronavirus, and of interest to the current epidemic are the α and β genera, which are typically found in bats (Cascella et al., 2020). One study found through a phylogenetic analysis of the whole genome that it was most closely related to the β genera, specifically with 89.1 % nucleotide similarity (Wu et al., 2020). The ssRNA is translated into two polyproteins following host cell entry (Cascella et al., 2020). The β genera genus typically encodes Mpro, which coupled with a papain-like protease, act to cleave the polyproteins to form the non-structural proteins, in which some form the replicase-transcriptase complex (RTC) (Fehr and Perlman, 2020). This function is two-fold, in that it produces the means by which the virus can replicate, but also in how it impacts host function. The RTC is used for RNA replication and transcription (Fehr and Perlman, 2020). As for the non-structural proteins. these include immune perturbing proteins, transmembrane scaffolds, helicases and more (Fehr and Perlman, 2020). Thus, Mpro presents a critical protein in the life cycle of a typical β-coronavirus. Based on the high-degree of similarity between the novel coronavirus and the β-genus, this presents as a prime target for functional conservation.

Mpro is also a well-characterized anti-viral drug therapy target for human coronaviruses, being a three-domain cysteine protease (Xue et al., 2008). It is known to contain a two-domain spanning chymotrypsin-like domain at the N-terminal region of the protein, with the C-terminal domain thought to be used for dimerization activity (Xue et al., 2008). Cysteine proteases represent a classification of protease characterized by their catalytic sites. Found in most organisms, there are several proteases of note, papain and cathepsins (Verma et al., 2016). Papain-like and cathepsin-like proteases have bee noted throughout literature to exist in coronaviruses. Cysteine proteases are characterized by a cysteine-histidine-asparginine active site triad (Verma et al., 2016). As for their synthesis, they are typically made as zymogens with a regulatory prodomain, and a catalytic mature domain (Verma et al., 2016). There are several activation mechanisms for a cysteine protease. To release itself from a zymogen, typically auto-catalysis based on a decrease in pH, trans-activation or unusual processing mechanisms (Verma et al., 2016). These activation mechanisms may highlight domains of functional conservation between the sars-cov-2 and other coronaviruses. In auto-processing, the pH change is thought to disrupt the interacting pro- and mature domain (Verma et al., 2016). This presents another key region for functional conservation analysis in Mpro.

**Methods**

**Databases.** Several databases were utilized for this study, focusing on protein databases in majority. RCSB PDB was used to browse various SARS-CoV-2 proteins (Berman et al., 2000). For tool outsourcing of various analyses, the ExPASy SIB Bioinformatics resource portal was utilized (Artimo et al., 2012). For homology-based modelling, the UNIREF-90 database was chosen based on being clustered at a 90% sequence identity threshold. Given the novel nature of SARS-Cov-2, this seemed an appropriate database for generating diversity, but homologous sequences.

**Mpro crystallized structure.** The PDB structure 6LU7, used in this study represents the first published crystal structure of the SARS-CoV-2 Mpro. The structure comes courtesy of a pre-print paper mapping the crystal structure of COVID-19 main protease in complex with an inhibitor N3 (Jin et al., 2020). Mpro is shown to follow dimerization of two molecules, each individually 3 domain proteins (Jin et al., 2020). For a comparative model, the PDB crystal structure 2A5I was used which represents a cysteine peptidase Mpro, from the SARS-CoV (Lee et al., 2005). Note that all Mpro characterizations were based on chain identification A. Sequences for both structures are provided in the appendix.

**Protein structure homology-modelling.** For model confirmation, SWISS-MODEL was used to generate a model for the chain A 6LU7 protein sequence, based on the template structure 2A5I. SWISS-MODEL uses BLAST and HHblits for homology-based template searching, followed by section via Global Model Quality Estimate (GMQE) and Quaternary Structure Quality Estimate (QSQE) for quality selection of templates (Waterhouse et al., 2018). Models are built based on alignment of conserved atom coordinates via OpenStructure and ProMod3 modelling engine, followed by model quality estimation via QMEAN scoring (Waterhouse et al., 2018).

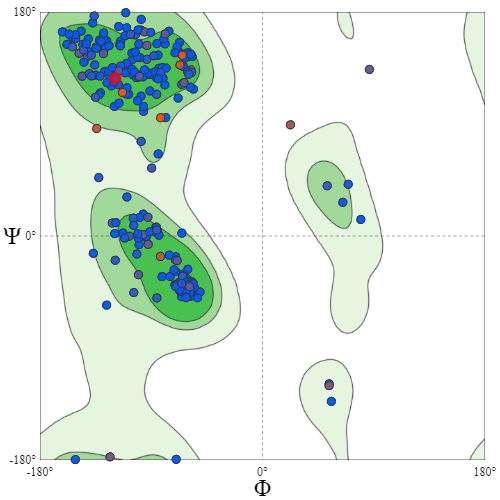
**Conservation analysis.** For conservation analysis of each model, the ConSurf server was used with both PDB protein structures (Landau et al., 2005). The ConSurf webserver was run using the following parameters for both models via respective PDB ID’s, for chain identifier A. Sequence homologs were searched against UNIREF90 using HMMER at an E-value of 0.0001 for 1 iteration, with a minimal %ID of 35 for homologs, and a maximal %ID of 95 between sequences. Multiple sequence alignment (MSA) was generated by the MUSCLE algorithm, following by which the best fit evolutionary model was chosen as LG, based on best fit. A phylogenetic tree was generated using neighbor joining with ML distance for conservation mapping, and is provided in the appendix, visualized using Interactive Tree of Life (iTOL) (Letunic and Bork, 2006). Conservation scores were calculated via Bayesian scoring, with a Best fit model of substitution. Note that conservation scores by ConSurf grading is highly impacted by the size of the MSA set, thus an MSA of unique hits, at least >30 sequences were used. ConSurf server results webpage links for both models are provided in the appendix. Conservation scores were mapped onto both models using UCSF Chimera (Pettersen et al., 2004). Code for generating the structures is provided in the appendix section as well. Finally, conservation of specific residues was visualized using Wasabi (Veidenberg et al., 2015).

**Domain analysis.** An initial scan of Mpro domains was conducted using the MyHits tool Motif Scan (Pagni et al., 2007). 6LU7 was scanned against HAMAP, Pfam and PROSITE profiles for a broad analysis approach to Mpro domain structure. ScanProsite and HMMSCAN were used as confirmatory analysis for domain mapping.

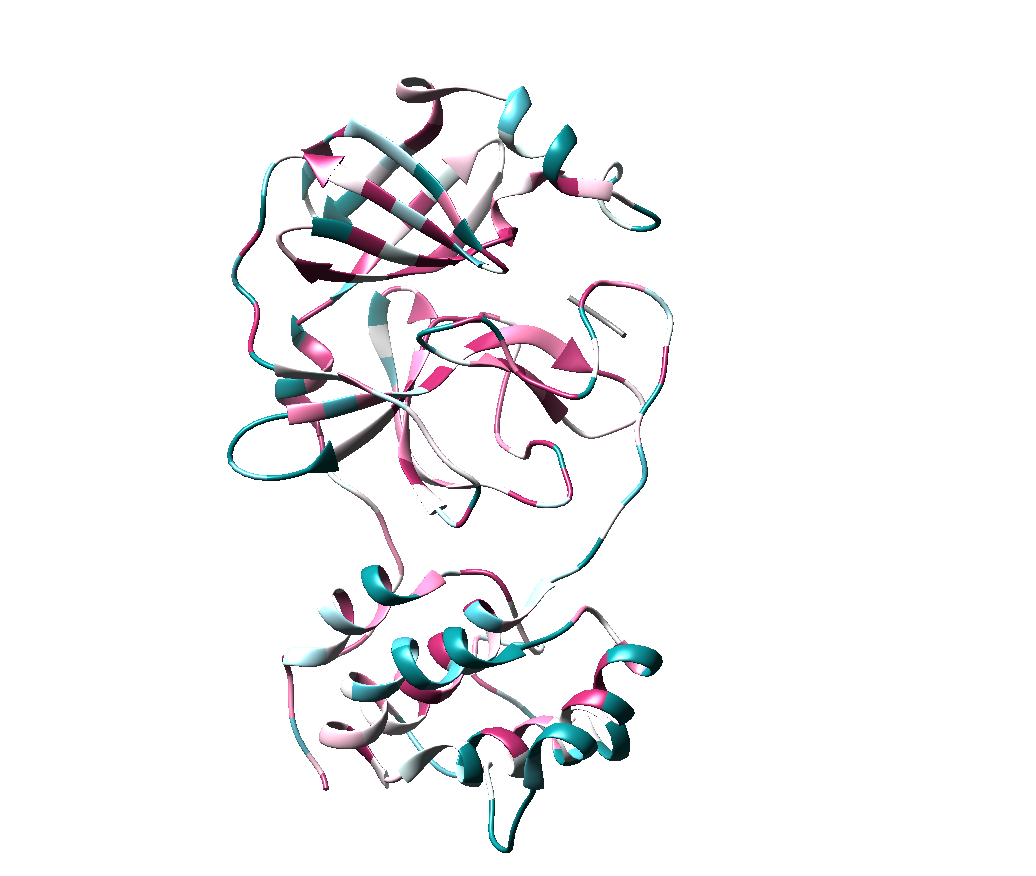
**Results**



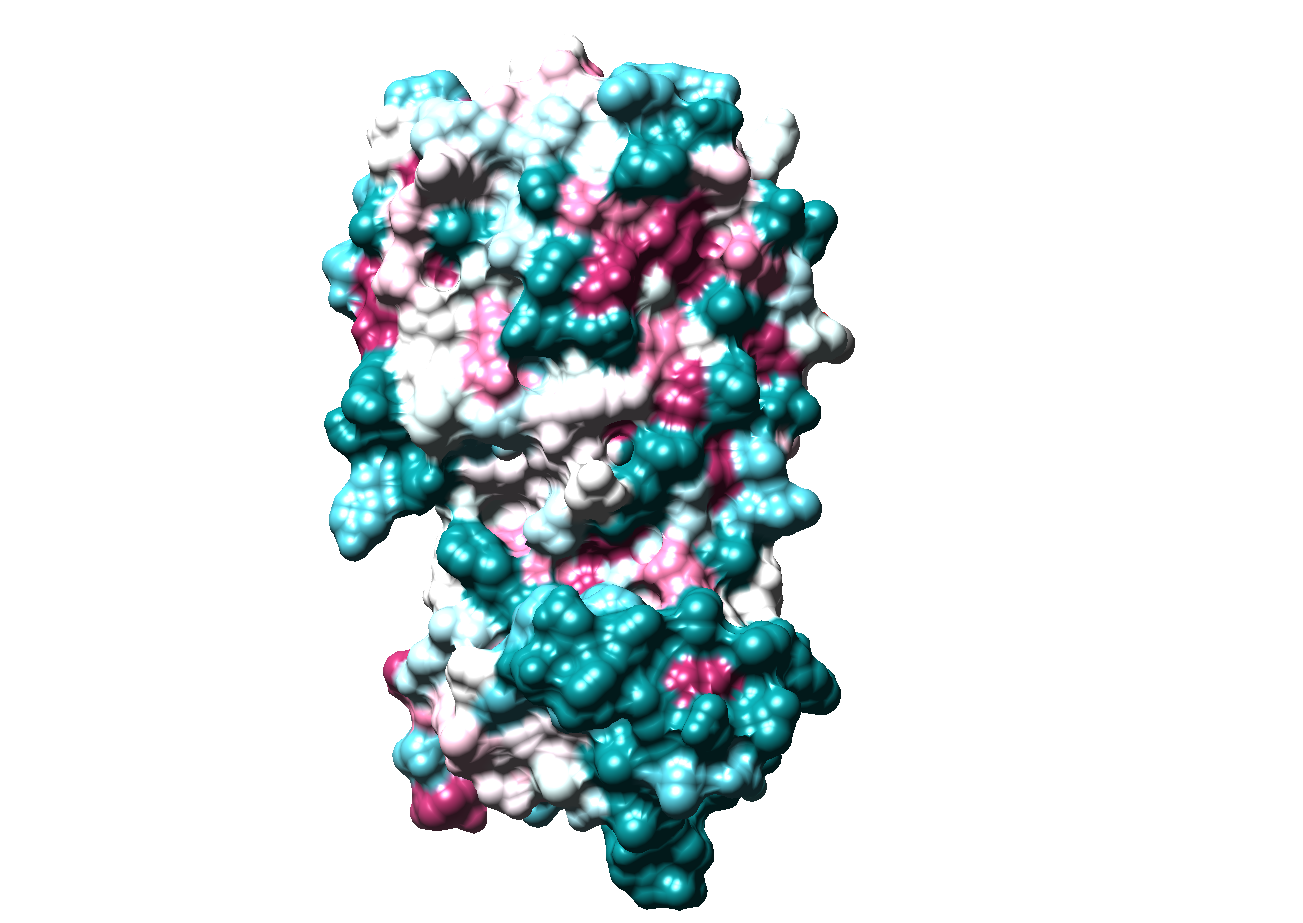
**Figure 1.** Mpro homodimer in complex with 2 AZP ligands. 6LU7**-**2A5I target-template homology-based model designed in SWISS-MODEL. Sequence identity was 96.08%, QMEAN of 0.03 and GMQE of 0.99.



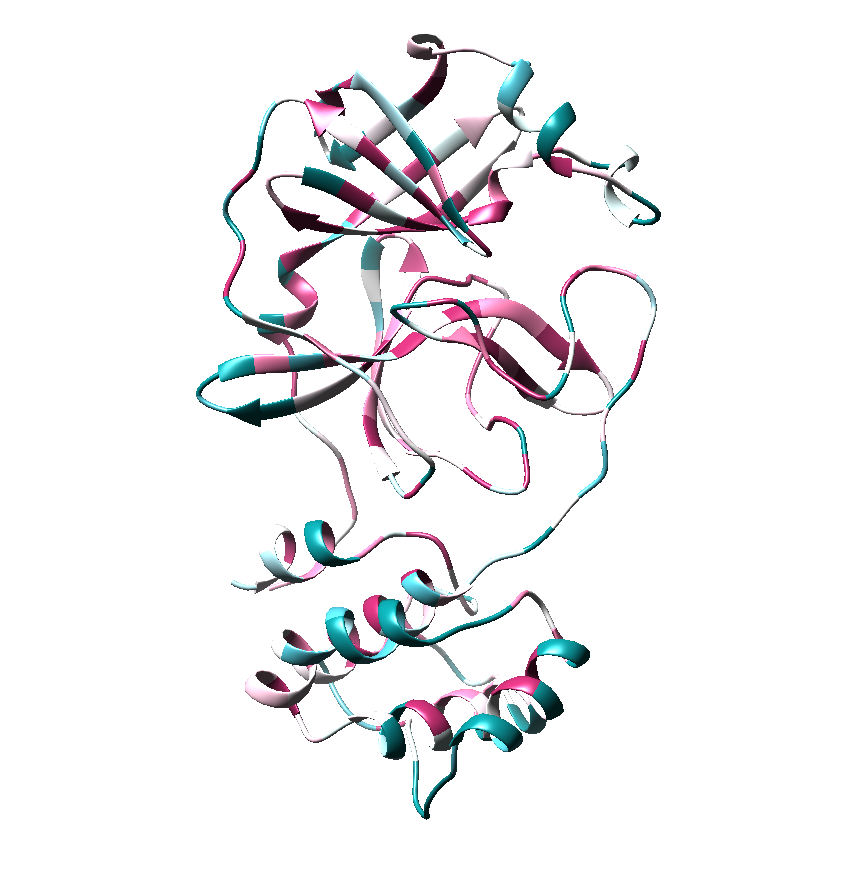
**Figure 2.** Ramachandran plots for Mpro Homodimer superimposed over background distribution. Although chains A and B are both shown, plot points stack, as they are dimerized. Generated through SWISS-MODEL structure assessment tool. Blue indicates higher quality residues, red indicates lower quality resiudes.



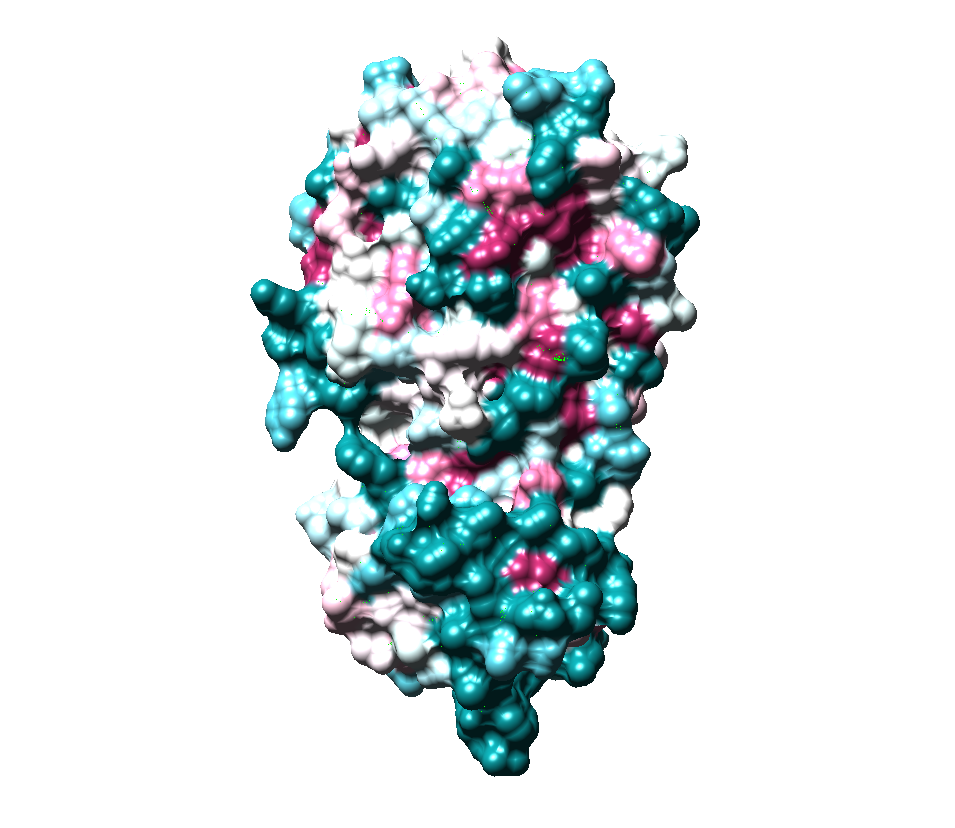
**Figure 3.** Backbone conservation mapping of 6LU7 chain A in Chimera. Conservation scores are color-coded in a gradient scale, dark-blue to dark-pink, from variable to conserved respectively.



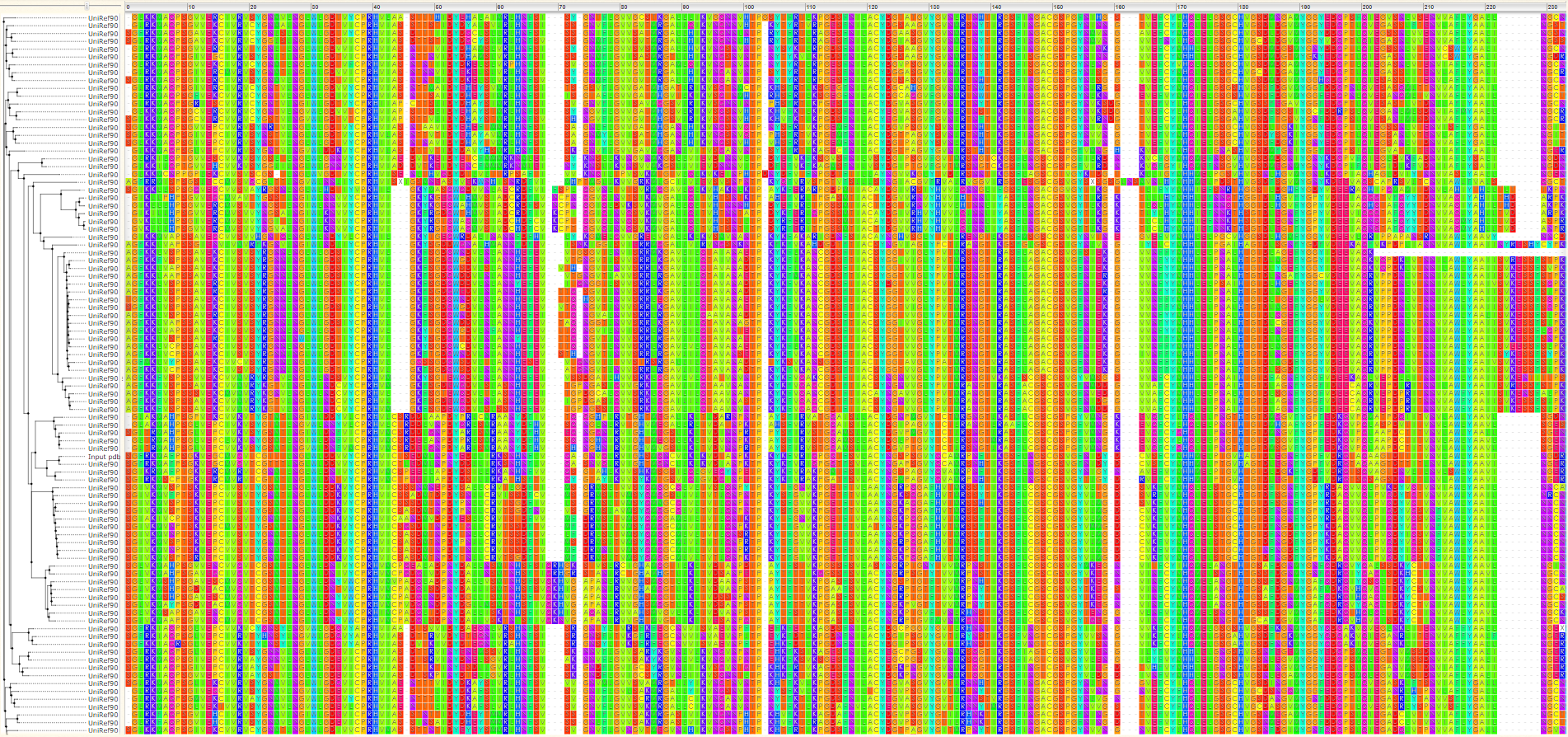
**Figure 4.** Surface conservation mapping of 6LU7 chain A in Chimera. Conservation scores are color-coded in a gradient scale, dark-blue to dark-pink, from variable to conserved respectively.



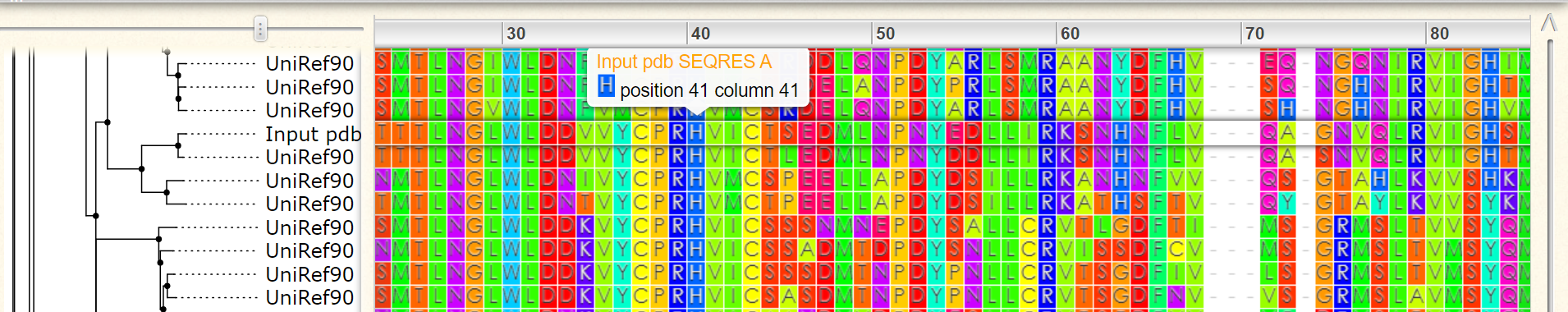
**Figure 5.** Backbone conservation mapping of model chain A in Chimera. Conservation scores are color-coded in a gradient scale, dark-blue to dark-pink, from variable to conserved respectively.

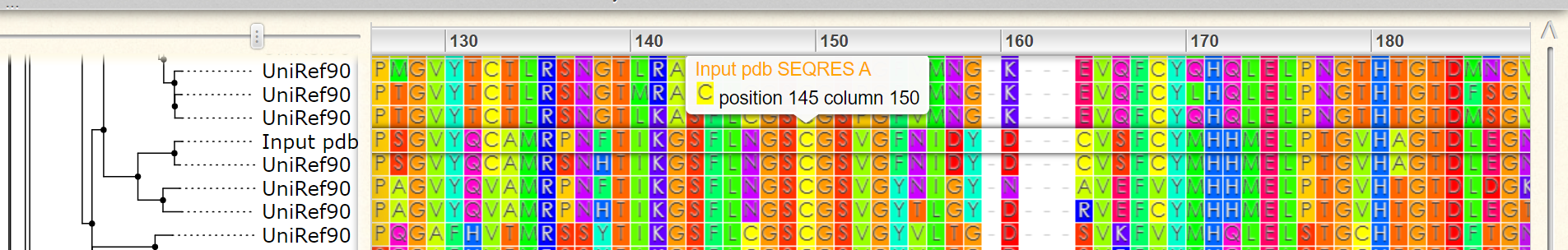
****

**Figure 6.** Surface conservation mapping of model chain A in Chimera. Conservation scores are color-coded in a gradient scale, dark-blue to dark-pink, from variable to conserved respectively.

****

**Figure 7.** Full MSA residue conservation map with phylogenetic tree. The image above was generated via ConSurf, visualized in WASABI. Residues are color-coded by residue, not conservation. The Mpro amino acid sequence is indicated by the descriptor “input\_pdb”.



****

**Figure 8.** Conservation of active dyad. Residues were focused based on domain analysis conducted in the following Tables. The images above were generated via ConSurf, visualized in WASABI. Residues are color-coded by residue, not conservation. The Mpro amino acid sequence is indicated by the descriptor “input\_pdb”.

**Table 1.** Motif-Scan list of matches for query 6LU7

|  |  |  |
| --- | --- | --- |
| Motif Database | Residues | Motif Information |
| PROSITE patterns | 133-136 | Asn glycosylation [?] |
| 142-145 |
| 45-48 | CK2 phosphorylation site [?] |
| 175-178 |
| 226-229 |
| 267-270 |
| 292-295 |
| 109-114 |
| 120-125 | N-myristoylation site [?] |
| 138-143 |
| 170-175 |
| 195-200 |
| 10-12 | PKC phosphorylation site [?] |
| 98-100 |
| 135-137 |
| 267-269 |
| PROSITE profiles | 1-306 | Main protease [!] |
| Pfam local | 29-306 | Peptidase C30 [!] |
| Pfam global | 57-76 | DUF321 [?] |
| 29-305 | Peptidase C30 [!] |

The above generated match list, within the motif information section, includes match status codes for each predicted motif, as defined by the MyHits SIB resource. Status codes represent classifications of true positives or false positives in sequence matching. [?] indicates a questionable/weak match, requiring additional data to distinguish between true/ false positives. [!] indicates a strong match. Note that DUF321 is of unknown function.

**Table 2.** PROSITE predicted features for Mpro

|  |  |  |
| --- | --- | --- |
| Predicted Features | Residues | Amino Acid |
| Peptidase C30 | 1-306 | N/A |
| Active site | 41 | H |
| Active site | 144 | C |

Features were predicted using ScanProsite quick Scan mode, excluding motifs with a high probability of occurrence from the scan itself.

**Table 3.** HMMSCAN results for Mpro feature prediction

|  |  |  |
| --- | --- | --- |
| Predicted Features | Residues | Amino Acid |
| Endopeptidase C30 | 29-306 | N/A |
| Active site | 41 | H |
| Active site | 145 | C |

Predicted features were found against the Pfam database. Results also indicated the presence of disorder, coiled-coil , a transmembrane domain and a signal peptide.

**Discussion**

Mpro presents with highly similar patterns to other homologs within the SARS-CoV classification of coronaviruses, as indicated by the results of this study. Firstly, in terms of SARS-CoV-2 protease structures, resolved structures were highly identical, as indicated in the SWISS-MODEL Homology Modelling Report. Based on this model, it is suggested that 6LU7 is an accurate model for Mpro. Ramachandran plot visualizes phi/psi combinations and essentially dictates which torsional angles can occur in a model’s structure. As seen in Figure 2, clustering seems to be occurring as expected, with the upper left quadrant showing beta sheets, lower left showing right-handed alpha-helix and upper right showing a left-handed helix. The superimposed plot indicates clustering near expected regions based on the generated background plot, with some residues as outliers. Note that these residues, indicated in red, had low quality thus this may have skewed the data. This analysis of the model supports model accuracy, based on correct phi/psi combinations. Based on homology, and model quality, the protein structures were deemed adequate for conservation analysis, specifically for functional mapping.

For initial conservation analysis, both backbone and surface mapping of ConSurf generated conservation scores were mapped onto both 6LU7 and the generated model. Firstly, in terms of Figures 3 and 4, the alpha-helix portion of the structure towards the tail end is highly variable, especially in the surface map, representing residues 200-306. The beta sheet region in the core of the backbone structure follows a much more conserved outlook, mapping locally around residues 160-180. For confirmation of mapping, this was done using the generate model, and as expected results were highly similar with little variation, as seen in Figure 5 and 6. This was mostly confirmatory, which is illustrated through the almost identical surface and backbone mapping. It was expected because when initial SWISS-model homology modelling was done, the sequence identity was highly similar. Note that this does not indicate conservation between the templet and target used for modelling, as noted in Figure 1. Rather, this provides confidence in the original conservation mapping in Figures 3 and 4 using the MSA generated via HMMR using Uniref90.

Having established confidence in the conservation map, it was of interest to see exactly where residues mapped, in terms of amino acid substitutions at a residue. In Figure 7, the entire alignment is provided, and there are blocks of conserved residues throughout the alignment, specifically towards the second half of the protein sequence. This can be better seen in Figure 8, which provides a closer look at some of the conserved residues of interest, based on domain analysis results. Domain analysis using Motif-Scan, as recorded in Table 1, identified mostly weak matches with some strong matches. Based on literature, it is unlikely that these weak matches represent motifs present in the protease, simply because they are not typical of cysteine proteases. In viruses, glycosylation is found, and typically highly variable, and in terms of cysteine proteases, there is evidence of Asn glycosylation in viruses (Rider et al., 2017). In terms of Coronavirus, for SARS-CoV, studies have also found the existence of aspargine-linked glycosylation sites on spike proteins (Zheng et al., 2018). This is however only suggestive and does not imply this is true for Mpro. Note that upon examination using Wasabi, these residues were also highly conserved, as can be seen in Figure 8 at the corresponding residues. There is no evidence in literature for CK2, PKC phosphorylation, or N-myristoylation sites in cysteine proteases. Most of the protein sequence matches the peptidase C30, which was confirmed through PROSITE (Table 2) and HMMSCAN (Table 3) with some discrepancies to where this domain is mapped. This domain is classified as expected, mapping to the SARS-CoV 3CL-like protease, also known as Mpro. Active sites were found through PROSITE and HMMSCAN as a catalytic histidine-cysteine dyad. Note that these are based on sequence alignments so it could be argued that function may differ between protease homologues. However, as these regions are conserved and map to similar protein domains, it is likely that Mpro is functionally conserved at its catalytic dyad. Figure 8 highlights these sites and shows no amino acid substitution at this residue and around this residue within a block of 7 (histidine) and 5 (cysteine) amino acids.

In summary, the homology-based model mapped confirmed that the 6LU7 was enough for conservation analysis. The SARS-CoV-2 Mpro chain A seems to be highly conserved in beta sheet regions, specifically at N-myristoylation and active sites. Based on the results provided in this study, SARS-CoV-2 Mpro chain A is functionally like other coronavirus proteases. It is suggested that a follow-up analysis be done based on chain B, as well as alignments be done using structures as well, rather than sequences.

**References**

Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, and Stockinger H. ExPASy: SIB bioinformatics resource portal, *Nucleic Acids Res*, 40(W1):W597-W603, 2012.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). *The Protein Data Bank*. *Nucleic Acids Research* (Vol. 28). Retrieved from http://www.rcsb.org/pdb/status.html

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2020). *Features, Evaluation and Treatment Coronavirus (COVID-19)*. *StatPearls*. StatPearls Publishing. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/32150360

Fehr, A. R., & Perlman, S. (2015). Coronaviruses: An overview of their replication and pathogenesis. In *Coronaviruses: Methods and Protocols* (Vol. 1282, pp. 1–23). Springer New York. https://doi.org/10.1007/978-1-4939-2438-7\_1

Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., … Yang, H. (2020). Structure of Mpro from COVID-19 virus and discovery of its inhibitors. *BioRxiv*, 2020.02.26.964882. https://doi.org/10.1101/2020.02.26.964882

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., & Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. https:/

Lee, T. W., Cherney, M. M., Huitema, C., Liu, J., James, K. E., Powers, J. C., … James, M. N. G. (2005). Crystal structures of the main peptidase from the SARS coronavirus inhibited by a substrate-like aza-peptide epoxide. *Journal of Molecular Biology*, *353*(5), 1137– 1151. https://doi.org/10.1016/j.jmb.2005.09.004

Letunic, I., & Bork, P. (2006). Phylogenetics Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation, *23*(1), 127–128. https://doi.org/10.1093/bioinformatics/btl529

Medhi, B., Prajapat, M., Sarma, P., Shekhar, N., Avti, P., Sinha, S., … Bansal, S. (2020). Drug targets for corona virus: A systematic review. *Indian Journal of Pharmacology*, *52*(1), 56–65. https://doi.org/10.4103/ijp.IJP\_115\_20

Pagni M, Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Hau J, Martin O, Kuznetsov D, Falquet L.MyHits: improvements to an interactive resource for analyzing protein sequences.*Nucleic Acids Res.* 2007 Jul; **35**(Web Server issue):W433-7.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Rider, P. J. F., Naderi, M., Bergeron, S., Chouljenko, V. N., Brylinski, M., & Kousoulas, K. G. (2017). Cysteines and N-Glycosylation Sites Conserved among All Alphaherpesviruses Regulate Membrane Fusion in Herpes Simplex Virus 1 Infection. *Journal of Virology*, *91*(21). https://doi.org/10.1128/jvi.00873-17

Veidenberg, A., Medlar, A., & L€, A. (2015). Wasabi: An Integrated Platform for Evolutionary Sequence Analysis and Data Visualization. https://doi.org/10.1093/molbev/msv333

Verma, S., Dixit, R., & Pandey, K. C. (2016, April 25). Cysteine proteases: Modes of activation and future prospects as pharmacological targets. *Frontiers in Pharmacology*. Frontiers Media S.A. https://doi.org/10.3389/fphar.2016.00107

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., … Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, *46*. https://doi.org/10.1093/nar/gky427

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., … Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), 265– 269. https://doi.org/10.1038/s41586-020-2008-3

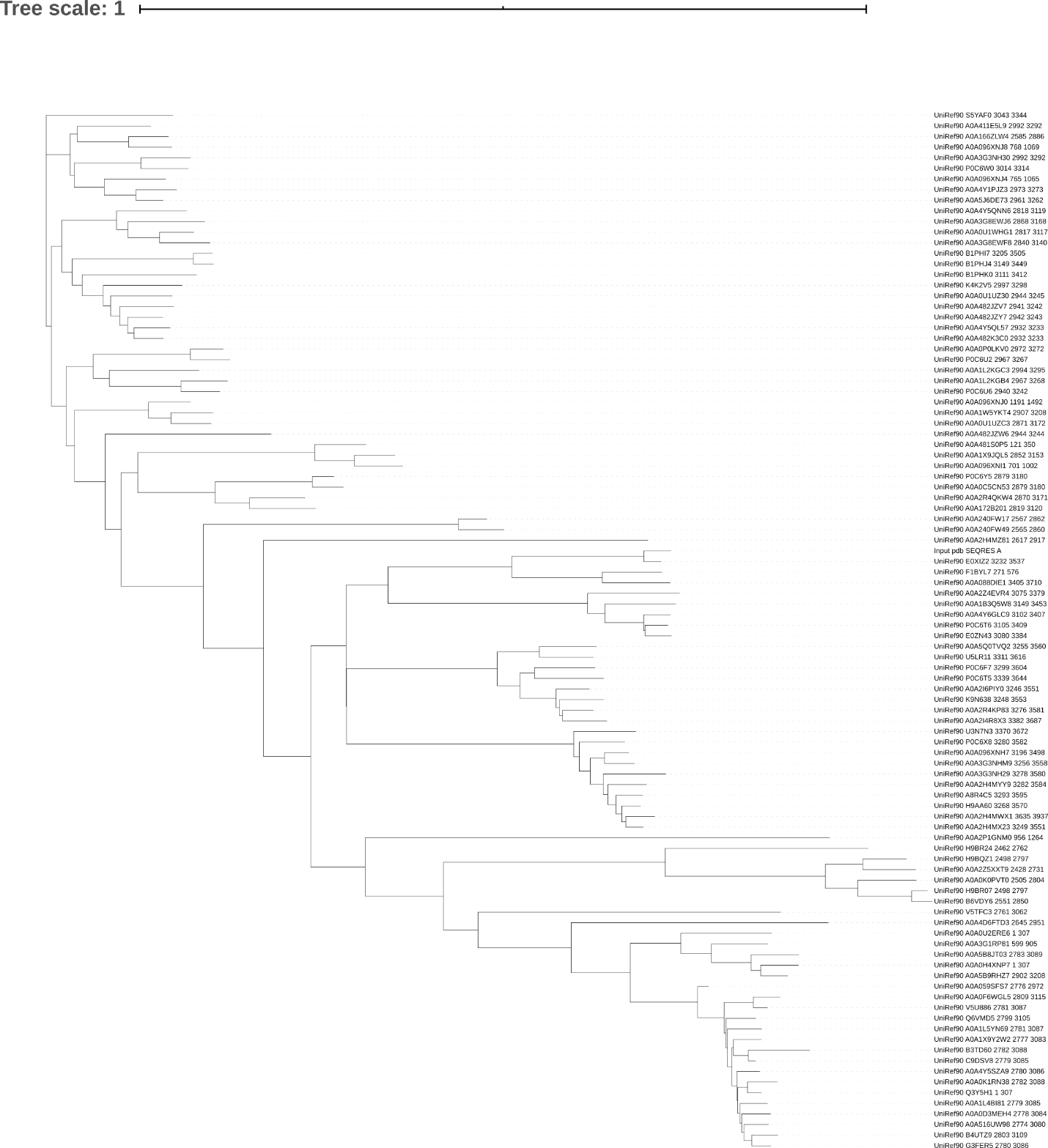
Xue, X., Yu, H., Yang, H., Xue, F., Wu, Z., Shen, W., … Rao, Z. (2008). Structures of Two Coronavirus Main Proteases: Implications for Substrate Binding and Antiviral Drug Design. *Journal of Virology*, *82*(5), 2515–2527. https://doi.org/10.1128/jvi.02114-07

Zheng, J., Yamada, Y., Fung, T. S., Huang, M., Chia, R., & Liu, D. X. (2018). Identification of N-linked glycosylation sites in the spike protein and their functional impact on the replication and infectivity of coronavirus infectious bronchitis virus in cell culture. *Virology*, *513*, 65–74. https://doi.org/10.1016/j.virol.2017.10.003

**Appendix**

**Table 4.** Links to reports and data generated in this study

|  |  |
| --- | --- |
| Data/Reports | Link |
| Plots, Models, MSA’s, Fasta Files and Reports | <https://github.com/mshunjan/SARS-CoV-2-Mpro> |
| SWISS-MODEL Homology Modelling Report | <http://htmlpreview.github.io/?https://github.com/mshunjan/SARS-CoV-2-Mpro/blob/master/report.html> |
| ConSurf Job Status for 6LU7 | <https://consurf.tau.ac.il/results/1585706801/output.php> |
| ConSurf Job Status for model | <https://consurf.tau.ac.il/results/1585884007/output.php> |



**Figure 7.** Mpro phylogenetic tree. Generated in Newick format by ConSurf neighbor joining with ML, then visualized using iTOL.

**Table 5.** Conservation score mapping in Chimera via command line

|  |  |
| --- | --- |
| Figure | Code |
| 1 | Open 6LU7  set bgColor white  lighting mode single |
| 2 | Open 6LU7  set bgColor white  lighting mode single  represent surface |
| 3 | Open model  set bgColor white  lighting mode single  ~display ~ :.\_  ~display ~ :.B |
| 4 | Open model  set bgColor white  lighting mode single  represent surface  ~display ~ :.\_  ~display ~ :.B |
| Coloring Script | colordef CONSGRAY 0.50 0.50 0.50  color CONSGRAY,a C  color CONSGRAY,r  color byhet  colordef CONS10 1.00 1.00 0.59  colordef CONS9 0.63 0.16 0.37  colordef CONS8 0.94 0.49 0.67  colordef CONS7 0.98 0.78 0.86  colordef CONS6 0.98 0.92 0.96  colordef CONS5 1.00 1.00 1.00  colordef CONS4 0.84 0.94 0.94  colordef CONS3 0.65 0.86 0.90  colordef CONS2 0.29 0.69 0.75  colordef CONS1 0.04 0.49 0.51  color CONS10 @/bfactor=10  color CONS9 @/bfactor=9  color CONS8 @/bfactor=8  color CONS7 @/bfactor=7  color CONS6 @/bfactor=6  color CONS5 @/bfactor=5  color CONS4 @/bfactor=4  color CONS3 @/bfactor=3  color CONS2 @/bfactor=2  color CONS1 @/bfactor=1 |

Note that PDB files provided with ConSurf updated colors were used for conservation mapping in conjugation with the coloring script