Systematic modeling of SARS-CoV-2 protein structures

Seán I. O'Donoghue[1,2,3], Andrea Schafferhans[4,5], Neblina Sikta[1], Christian Stolte[1], Sandeep Kaur[1,3], Bosco Ho[1], Stuart Anderson[2], James Procter[6], Christian Dallago[5], Nicola Bordin[7], Burkhard Rost[5], Matt Adcock[2]

1: Garvan Institute of Medical Research, Sydney, Australia. 2: CSIRO Data61, Sydney, Australia. 3: School of Biotechnology and Biomolecular Sciences (UNSW), Sydney, Australia. 4: Department of Bioengineering Sciences, Weihenstephan-Tr. University of Applied Sciences, Freising, Germany. 5: Department of Informatics, Bioinformatics & Computational Biology - i12, TUM (Technical University of Munich), Munich, Germany. 6: The University of Dundee, UK. 7: Institute of Structural and Molecular Biology, University College London, London, WC1E 6BT, UK

Abstract

In response to the COVID-19 pandemic caused by the SARS-CoV-2 virus, structural biologists are using experimental structural determination methods to better understand the viral proteome. Our goal in this work was to help researchers use these rapidly emerging structural data to gain detailed insights into the molecular mechanisms underlying COVID-19 infection. Our analysis was based on the protein sequences defined by UniProt as comprising the viral proteome. We systematically compared each SARS-CoV-2 protein sequence against all available protein 3D structures derived from any organism (164,250 PDB entries), using pairs of hidden Markov models built with the HHblits tool. We found 872 sequence-to-structure alignments assessed to have significant similarity (E < 10e-10) to infer structural similarity. The resulting 872 3D template models now provide a wealth of new detail, currently not available from related resources. To help make this large, complex dataset accessible and usable for other researchers, we also developed a tailored layout strategy to visually organise the 3D models by mapping them to the viral genome. The resulting graph provides an immediate and comprehensive visual overview of what is known - and not known - about the 3D structure of the viral proteome, thereby helping direct future research. The graph also clearly reveals all available structural evidence of viral mimicry or hijacking of human proteins, as well as all evidence of interactions between viral proteins. We have created PDF and online versions of the graph, in which users can click on any node in the graph to open the corresponding 3D model in the Aquaria molecular graphics system. In Aquaria, these models can then be colored to show sequence features, such as single nucleotide polymorphisms and posttranslational modifications. Previous versions of Aquaria showed only features from UniProt; however, as part of this study, we have now added features from PredictProtein and CATH, thus providing a total of 32,717 features for SARS-CoV-2 protein sequences. In this work, we present novel insights found, using the above approach, into how SARS-CoV-2 mimics and hijacks host proteins, and how viral proteins self-assemble during infection. The resulting Aquaria-COVID resource is freely available online at https://aquaria.ws/covid19, and an accompanying video (https://youtu.be/J2nWQTlJNaY) explains how researchers can use the resource.

INTRODUCTION

In response to the COVID-19 pandemic, structural biologists are applying experimental structural determination methods to SARS-CoV-2 (a.k.a. Human Severe Acute Respiratory Syndrome Coronavirus 2), so far resulting in 284 Protein Data Bank entries (PDB; Berman et al. 2000). These structures, in turn, are being augmented by homology modeling studies, of which the majority explore a specific research question in depth, such as interaction of the viral spike glycoprotein with ACE2 (e.g., Jaimes et al., 2020).

A less common approach focuses on breadth of coverage, applying homology modelling to the entire SARS-CoV-2 proteome; this has been done using AlphaFold (Senior et al., 2020), C-I-TASSER (Zheng et al., 2019), Rossetta (Rohl et al., 2004), and SwissModel (Waterhouse et al., 2018). Unfortunately, the resulting predicted models vary greatly (Heo and Feig, 2020), raising quality issues that are explored in the current CASP activity (Kryshtafovych et al., 2019). In addition, breadth-of-coverage approaches focus on determining one structural model for each viral protein or domain, largely ignoring viral and host protein interactions, and resulting in relatively few total models (e.g., 24 for AlphaFold and 116 for SwissModel).

Our resource addresses these limitations by systematically modelling each viral protein sequence against all available 3D structures from any organism (164,250 PDB entries). This generates many more models and finds all viral and host protein interactions with supporting structural evidence. This approach combines both breadth and depth of coverage, but incurs a computational cost and so requires a simpler strategy ('minimal modelling') in which we use profile-profile comparisons (Steinegger et al., 2019) to align viral protein sequences onto experimentally derived 3D structures (O'Donoghue et al., 2015). Thus, in our resource, the 3D coordinates are not modified in any way from the experimental structure, but are shown mapped onto SARS-CoV-2 sequences using a coloring scheme that makes model quality immediately apparent (Heinrich et al., 2015).

A key benefit of using minimal models is that it can be straightforward for a researcher to understand how they were derived and to assess the validity of any insights gained. This makes minimal models broadly useful for scientists who are not modelling experts. By contrast, homology models generated by more sophisticated methods (e.g., Senior et al., 2020) can be more accurate (Kryshtafovych et al., 2019), but a researcher using them needs to invest considerable time and effort to understand their strengths and weaknesses, thus limiting their usefulness.

The large number of models that can be generated by such minimal strategies (e.g., HOMCOS; Kawabata, 2016) raises a new problem: how to visually organise these complex datasets so they are accessible and usable for other researchers. We believe this is a critical issue, and that the current paucity of effective visual organisation of key datasets is a rate limiting step impeding not just COVID-19 research, but life science research generally (O'Donoghue et al.,

2018). For our resource, we have developed a tailored layout strategy to visually organise SARS-CoV-2 3D models by mapping them to the viral genome.

An effective way to use each of the resulting 3D models is to map sequence features, such as domains or Post-Translational Modifications (PTMs). In our resource, we integrated the SARS-CoV-2 3D models into the Aquaria resource (O'Donoghue et al., 2015), originally designed to facilitate mapping of UniProt sequence features (The UniProt Consortium 2019). For this work, we have added features from PredictProtein (Yachdav et al., 2014) and CATH (Dawson et al., 2017).

The resulting Aquaria-COVID resource provides researchers with a rich set of SARS-CoV-2 protein models and sequence features that are easy to use, and currently not available from related resources. Our resource also identifies structurally dark regions of the viral proteome, i.e., regions with no detectable sequence similarity to any protein that has been observed by experimental structure determination (Perdigão et al., 2015; Schafferhans et al., 2018). In this work, we describe how we created the resource and present novel insights we found using it. These insights shed light on molecular mechanisms underlying COVID-19 infection, especially on the self-assembly of SARS-CoV-2 proteins, and how they mimic and hijack host proteins.

RESULTS

The core of our resource is a set of 872 sequence-to-structure alignments derived by matching all sequences in the SARS-CoV-2 proteome against sequences of all available 3D structures in PDB. Of these alignments, almost all were matched to structures of viral proteins - in some cases, these viral proteins were in complex with host proteins, indicating potential hijacking. In a small number of cases, regions of the SARS-CoV-2 proteome aligned directly onto human proteins, indicating potential mimicry. In the text below, and in Fig. 1, we present all such cases of potential hijacking or mimicry of human proteins found by our systematic analysis. For brevity, we have largely omitted mention of cases involving non-human host proteins. These matching structures were incorporated into the Aquaria resource, and augmented with 32,717 sequence features from PredictProtein, SNAP2, and CATH. Below, we highlight key findings revealed by using the Aquaria interface to systematically explore these matching structures in combination with these features.

Polyprotein 1a (pp1a)

Nsp1 and Nsp2. The first 180 residues of pp1a, which comprise nsp1, had no CATH assignments and only three matching structures. The top-ranked match (P0DTC1/2gdt from SARS-CoV, 85% identity, E = 10^{-20}) spanned residues 12–126 (Almeida et al., 2007). The second match was to an earlier dataset from the same research team. The third match (P0DTC1/3ld1 from IBV, 11% identity, E = 10^{-32}) spanned pp1ab 104–562, which also covered the N-terminal half of nsp2 and is the only structural match for nsp2. Of the multiple dark regions found in nsp1 and nsp2, only the first region (pp1ab 1–11) is explainable by features considered in this study, as this region had residues with extremely high relative B-value and disorder. The

three matching structures for nsp1 and nsp2 showed no matches to, or interaction with, human proteins.

Nsp3. Residues 819–2763 of pp1a comprise nsp3, which was predicted to occur predominantly in the cytoplasm and had 188 matching structures that clustered to define nine distinct sequence regions, outlined below.

(1) The first region (pp1a 819–929) was highly conserved and had three matching structures (P0DTC1/2gri from SARS-CoV, 77% identity, E = 10^{-23}) with ubiquitin-like fold (CATH 3.10.20.350).

(2) This was followed by a dark region (pp1a 930–1026) with low conservation, low mutational sensitivity, high relative B-factor, and no CATH assignments.

(3) Next was a macro domain region (pp1a 1027–1193; CATH 3.40.220.10) with high conservation and with 144 matching structures (P0DTC1/6woj, 100% identity, E = 10^{-24}). Of these, 47 structures showed alignment between the viral proteome and human proteins, indicating that the viral macro domain may potentially mimic human proteins (Fig.1). The matched proteins were: GDAP2 (P0DTC1/4uml, 19% identity, E = 10^{-22}), MACROD1 (P0DTC1/2x47, 25% identity, E = 10^{-22}), MACROD2 (P0DTC1/4iqy, 24% identity, E = 10^{-22}), MACROH2A1 (P0DTC1/1zr5, 19% identity, E = 10^{-18}), MACROH2A2 (P0DTC1/2xd7, 19% identity, E = 10^{-21}), OARD1 (P0DTC1/2eee, 13% identity, E = 10^{-12}), PARP9 (P0DTC1/5ail, 22% identity, E = 10^{-17}), PARP14 (P0DTC1/3q6z, 25% identity, E = 10^{-19}), and PARP15 (P0DTC1/3v2b, 18% identity, E = 10^{-16}). An additional 73 structures matched to viral proteins, two in complex with RNA (P0DTD1/4tu0, 24% identity, E = 10^{-17}; P0DTD1/3gpq, 24% identity, E = 10^{-17}). For brevity, here and in Fig. 1, we have not provided details about the remaining 41 matching structures, which show viral proteins matched to host proteins in other organisms (including archaea, bacteria, yeast, and mouse).

(4) Next was another dark region (pp1a 1027–1368) with low conservation and comprising a short disordered region followed by another macro-like domain called SUD-N (CATH 3.40.220.30).

(5) The SUD-M region (pp1a 1389–1493; CATH 3.40.220.20) comprised a third macro-like domain called SUD-M (CATH 3.40.220.20) with seven matching structures (P0DTC1/2jzd, 81% identity, E = 10^{-34}), all determined using SARS-CoV nsp3. This was the first of four slightly overlapping regions with matching structures (Fig. 1).

(6) The SUD-C region (pp1a 1494–1562; CATH 2.30.30.590) had high conservation and comprised a domain called SUD-C (CATH 2.30.30.590) that had two matching structures: one structure spanned the region (P0DTC1/2kqw, 76% identity, E = 10^{-41}), while the second, more distant matching structure (P0DTC1/4ypt from mouse hepatitis virus strain A59, 28% identity, E = 10^{-62}) also spanned the following PL-PRO region.

(7) The PL-Pro region (pp1a 1563–1878) comprises three sub-domains: a highly conserved ubiquitin-like domain (CATH 3.10.20.540), a catalytic 'thumb' domain (CATH 1.10.8.1190) that is also highly conserved, and a final domain with lower conservation containing a

'fingers' and 'palm' motif (CATH 3.90.70.90; (Wang et al., 2020)). The palm motif includes a C4-type zinc finger motif, annotated in the UniProt features (pp1a 17520–1789). PL-Pro had 38 matching structures (P0DTC1/6wrh from SARS-CoV-2, 100% identity), of which 11 feature a viral protein in complex with a human protein related to ubiquitin (Fig.1). Four of these structures showed the PL-Pro region in complex with ISG15 (P0DTC1-5tl6 from SARS-CoV, 82% identity, E = 10^{-70}); two structures showed a complex with UBA52 (P0DTC1/4rf0 from MERS-CoV, 29% identity, E = 10^{-65}); two structures showed a complex with UBB (P0DTC1/5e6j from SARS-CoV, 83% identity, E = 10^{-62}); one structure showed a complex with UBC (P0DTC1/4mm3 from SARS-CoV, 83% identity, E = 10^{-62}), and one structure showed a complex with both UBB and UBC (P0DTC1/5e6j from SARS-CoV, 83% identity, E = 10^{-62}). Note these identity scores and E-values do not indicate similarity to the human, ubiquitin-like proteins, but rather the similarity between SARS-CoV-2 nsp3 and the viral proteins in the matching structures (e.g., P0DTC1/4rf0 has 29% identity as it is based on a structure determined with MERS-CoV nsp3). Two additional matching structures show SARS-CoV-2 nsp3 in complex with inhibitory peptides (P0DTC1/6wuu, 100% identity; P0DTC1/6wx4, 100% identity).

(8) The NAB region (pp1a 1879–2020) had low conservation and only one matching structure (P0DTC1/2k87 from SARS-CoV; 82% identity, E = 10^{-24}) that adopts a unique fold, not seen in any other structure reported to date (CATH 3.40.50.11020).

(9) The final dark region (pp1a 2021–2763) had no CATH matches, and comprised two distinct segments. The first segment (pp1a 2021–2396) had overall low conversation and began with several disordered residues, followed by four transmembrane helices and one or two short sub-regions that occur in the lumen. The final segment (pp1a 2397–2763) had higher conservation, no disorder, no transmembrane domains, and was predicted to be fully located in the cytoplasm, along with regions 1–8 of nsp3. This segment has been assigned two domains, called Y1 and CoV-Y (Lei et al., 2018).

Nsp4. Most of this protein comprised a dark region (pp1a 2764–3172) with no matching structures, no CATH matches, low conservation, and was predicted to have no disorder, but multiple transmembrane helices. The C-terminal 91 residues (pp1a 3173–3263) comprised a conserved domain called nsp4C (CATH 1.10.150.420) with three matching structures; none of which had matches to, or interactions with, human proteins. All matching structures were dimers, however, as reported for the best matching structure (P0DTC1/3vcb, 59% identity, E = 10^{-35}) derived from murine hepatitis virus nsp4, nsp4C is believed to operate primarily as a monomer (Xu et al., 2009).

Nsp5. Also known as 3CL-Pro, this protein (pp1a 3264–3569) was highly conserved, matched two CATH families (2.40.10.10 and 1.10.1840.10), and had 256 matching structures (P0DTC1/5rfa; 100% identity), none with matches to, or interactions with, human proteins. However, 15 structures showed interaction with inhibitory peptides (Fig. 1; P0DTC1/7bqy, 100% identity), including several determined with distant homologs, such as 3CL-Pro from IBV, or infectious bronchitis virus (P0DTC1/2q6f, 42%,E = 10^{-122}).

Nsp6. Amongst the non-structural proteins, nsp6 (pp1a 3570–3859) was the only one that was fully dark, i.e., with no structural matches in Aquaria. In addition, nsp6 had low conservation, no CATH matches, and was predicted to have no disorder but multiple transmembrane helices.

Nsp7. The nsp7 protein (pp1a 3860–3942) had 15 matching structures (P0DTC1/2kys; 98% identity; E = 10^{-32}), all comprising an antiparallel bundle of four helices (CATH 1.10.8.370). Two of the matching structures were monomeric (P0DTC1/2kys and P0DTC1/1ysy; 98% identity; E = 10^{-32}), while in the remaining 13, nsp7 was bound to nps8 (P0DTC1/3ub0; 43% identity; E = 10^{-79}). Of these 13 structures, six also included nsp12 as a binding partner (P0DTC1/7bv1; 99% identity), and in two cases RNA as well (P0DTC1/7bv2; 99% identity and P0DTC1/6yyt; 97% identity; E = 10^{-32}). In these matching structures, nsp7 adopts a range of conformationally distinct structures, depending on the environment and on its interaction partners (Johnson et al., 2010).

Nsp8. The nsp8 protein (pp1a 3943–4140) begins with a highly conserved, predominantly helical segment (pp1a 3943–4041) with some disordered regions (pp1a 4018–4019), and no CATH matches. This is followed by a 'head' domain (pp1a 4042–4140) with a mostly beta barrel fold (CATH 2.40.10.290). Of the 14 matching structures, one featured nsp8 bound to nsp12 only (P0DTC1/6nus; 97% identity; E = 10^{-78}). As noted above, the remaining 13 matching structures all featured nsp8 bound to nsp7 (P0DTC1/3ub0) - as shown in Fig. 1, these two copies of each of these proteins can form a heterotetramer, which can then form a hexadecamer (P0DTC1/2ahm). Six of the 13 matching structures also included nsp12, and two also included RNA.

Nsp9. The nsp9 protein (pp1a 4141–4253) had intermediate to high conservation with a beta barrel architecture similar to thrombin subunit H (CATH 2.40.10.250). It had 14 matching structures (P0DTC1/6wxd; 98% identity; E = 10^{-44}), mostly homodimers, but with no matches to, or interactions with, human proteins.

Nsp10. The final pp1a protein (residues 4254–4392) had no CATH matches, yet was conserved and had 33 matching structures. In one of the matching structures, nsp10 was monomeric (P0DTC1/2fyg; 97% identity; E = 10^{-57}), while in two of the matching structures, nsp10 was assembled into a dodecamer (Fig. 1; P0DTC1/2ga6; 96% identity; E = 10^{-72} and P0DTC1/2g9t; 96% identity; E = 10^{-72}), forming a hollow sphere with twelve C-terminal zinc finger motifs sticking out from the outer surface, and another 12 zinc finger motifs on the inner surface (Su et al., 2006). In four matching structures, nsp10 was in complex with nsp14 (P0DTC1/5c8u; 97% identity; E = 10^{-68}), while the remaining 26 matching structures all showed nsp10 in complex with nsp16 (P0DTC1/6w61; 100% identity).

Polyprotein 1b (pp1b)

The five proteins from polyprotein 1b were all predicted to have no regions of disorder or transmembrane helices.

Nsp12. This protein spans pp1ab residues 4393–5324, and comprises nsp11, i.e., the last 13 residues from pp1a (4393–4405), concatenated with the first 919 residues of pp1b. Aquaria found 47 matching structures (all of viral proteins), of which seven showed interactions with other viral proteins, including: one structure of nsp12 in complex with nsp8 (P0DTC1/6nus from SARS-CoV, 97% identity; E = 10^{-77}), or equivalent proteins in other viruses; four structures of the same two proteins in complex with nsp7 (P0DTD1/7bv1 from SARS-CoV-2, 99% identity, E = 10^{-230}); and two structures of these three proteins in complex with RNA (P0DTD1/6yyt from SARS-CoV-2, 100% identity). An additional 15 structures showed nsp12 in complex with RNA only (P0DTD1/3kmq from FMDV, 18% identity, E = 10^{-11}), while one structure showed nsp12 in complex with both RNA and DNA (P0DTD1/4k4v, 17% identity, E = 10^{-11}).

Nsp13. For nsp13 (pp1ab 5325–5925), Aquaria found 64 matching structures, of which 23 showed potential mimicry of four human proteins. This included: 11 structures showing mimicry of one of the three Rossman fold motifs (CATH 3.40.50.300) of AQR (P0DTD1/4pj3, 19% identity, E = 10^{-30}), of which 10 also showed AQR in complex with the spliceosome (P0DTD1/6id0, 20% identity, E = 10^{-30}); two structures showing mimicry of PIF1 (P0DTD1/6hpt, 20% identity; E = 10^{-12}); eight structures showing mimicry of UPF1 (P0DTD1/2xzo, 24% identity; E = 10^{--32}), of which two also showed UPF1 in complex with UPF2 (P0DTD1/2wjy, 23% identity; E = 10^{-34}); and, finally, two structures showing mimicry of IGHMBP2 (P0DTD1/4b3f, 23% identity; E = 10^{-35}), of which one also showed IGHMBP2 in complex with RNA (P0DTD1/4b3g, 25% identity; E = 10^{-33}). An additional 41 structures showed matches to viral proteins (best match to SARS-CoV nsp13; P0DTD1/6jyt, 100% identity), with four also showing these viral proteins in complex with DNA (P0DTD1/4n0o, 22% identity, E = 10^{-19}). For nsp13, two regions were identified as Rossman fold motifs (CATH 3.40.50.300); one of these (CATH functional family 001105) spanned most of the region matching AQR, the other (CATH functional family 001139) spanned the region matching PIF1. The regions matching UPF1 and IGHMBP2 spanned both of these Rossman fold motifs. The N-terminal region of nsp14 (pp1ab 5325–5577) had no matches in CATH.

Nsp14. For nsp14 (pp1ab 5926–6452), Aquaria found four matching structures (P0DTD1/5nfy; 100% identity); all matched to SARS-CoV nsp14 and all were in complex with nsp10. There are currently no matches for nsp14 in CATH.

Nsp15. For nsp15 (pp1ab 6453–6798), Aquaria found 19 matching structures (P0DTD1/6wxc; 99% identity). The N-terminal region matched a known superfamily (CATH 2.20.25.360), while most of the nsp15 sequence (pp1ab 6516–6798) had no matches in CATH.

Nsp16. For nsp16 (pp1ab 6799–7096), Aquaria found 26 matching structures (P0DTC1/6w61; 100% identity), all in complex with nsp10. The full length nsp16 sequence matched to a single Rossman fold motif (CATH 3.40.50.150).

Viron and accessory proteins

The remaining 12 proteins encoded by the 3' end of the genome ultimately assemble together to form the mature viral capsid. Remarkably, however, our analysis of all related 3D structures found no structures showing interactions between any of these proteins.

Spike glycoprotein. For this protein, Aquaria found 136 matching structures (P0DTC2/6vxx from SARS-CoV-2, 99% identity) clustered in two regions. One region comprised 15 structures and matched to TM1, the C-terminal transmembrane helix (P0DTC2/2fxp from SARS-CoV, 98% identity, E = 10^{-14}); four of the TM1 matching structures showed a complex with antibodies (P0DTC2/6pxh from MERS-CoV, 22% identity, E = 10^{-11}). The second region comprised 121 matching structures, of which 34 spanned nearly the full length of the spike glycoprotein sequence, and were assembled into a homotrimer. Of the 121 structures in this region, 68 were in complex with antibodies (P0DTC2/6w41 from SARS-CoV-2, 100% identity) and two structures were in complex with inhibitory peptides (P0DTC2/5zvm from SARS-CoV, 88% identity, E = 10^{-32}; P0DTC2/5zvk from SARS-CoV, 55% identity, E = 10^{-32}). Of the remaining structures, 18 showed the spike glycoprotein in complex with human proteins: 15 were in complex with ACE2 (P0DTC2/6acg from SARS-CoV, 77% identity, E = 10^{-321}); one was in complex with both ACE2 and SLC6A19 (P0DTC2/6m17 from SARS-CoV-2, 100% identity); and, finally, three were in complex with DPP4 (P0DTC2/4qzv from BtCoV-HKU4, 21% identity, E = 10^{-45}). Spike glycoprotein was the only capsid protein with matching structures showing binding to human proteins; this is consistent with the belief that much of the assembly of the capsid occurs in cellular compartments, largely shielded from interaction from host proteins, host RNA, or host DNA. Finally, our analysis found four short dark regions in the spike glycoprotein, including one at each of the N- and C-terminals.

ORF3a protein. The ORF3a protein had no matching structures, i.e., was a dark protein.

Envelope protein. The envelope protein (a.k.a. E protein) had only two matching structures, both from SARS-CoV. One structure was a monomer (P0DTC4/2mm4, 91% identity, E = 10^{-26}), while in the other structure the envelope protein was assembled into a homopentamer (P0DTC4/5x29, 86% identity, E = 10^{-27}), forming a membrane-spanning ion channel.

Matrix glycoprotein. The matrix glycoprotein (a.k.a. M protein) had no matching structures, i.e., was a dark protein.

ORF6 protein. The ORF6 protein had no matching structures, i.e., was a dark protein.

ORF7a protein. The ORF7a protein had only three matching structures: one from SARS-CoV-2 (P0DTC7/6w37, 100% identity) and two from SARS-CoV (P0DTC7/1yo4, 90% identity, E = 10^{-59}; P0DTC7/1xak, 88% identity, E = 10^{-59}). Our analysis also found short dark regions at the N- and C-terminals.

ORF7b protein. The ORF7b protein had no matching structures, i.e., was a dark protein.

ORF8 protein. The ORF8 protein had no matching structures, i.e., was a dark protein.

Nucleocapsid protein. The nucleocapsid protein (a.k.a. N or NP protein) had 35 matching structures clustered in two regions. The region closest to the N-terminus had 24 matching structures (P0DTC9/6yi3 from SARS-CoV-2, 100% identity), most of which comprised a single monomer of the nucleocapsid protein; however five structure comprised a dimer, and one comprised a tetramer. The region closest to the C-terminus had 13 matching structures (P0DTC9/6wji from SARS-CoV-2, 98% identity), all of which comprised a dimer. Our analysis also found three short dark regions, two of which covered the N- and C-terminals.

ORF9a protein. This protein had only two matching structural models, both derived from the same PDB entry (P0DTD2/2cme from SARS-CoV, 78% identity, E = 10^{-48}), which included a lipid analog for this lipid-binding protein.

ORF10 protein. The ORF10 protein had no matching structures, i.e., was a dark protein.

ORF14 protein. The ORF14 protein had no matching structures, i.e., was a dark protein.

Related resources

Table 1 summarizes key statistics on our resource, and provides a comparison with related resources. The key difference in our approach is in the depth of structural modeling, which aims to find all related 3D models, including many remote homologs. As a result, our resource captures a wealth of detailed information about structural states and interactions not readily available from other resources.

DISCUSSION

Figure 1 provides an immediate and comprehensive visual overview of what is known - and not known - about the 3D structure of the viral proteome, thereby helping direct future research. The graph also documents all available structural evidence of viral mimicry or hijacking of human proteins, as well as all evidence of interactions between viral proteins. Somewhat remarkably, however, our analysis found so few cases of viral mimicry, hijacking, or self-assembly with structural evidence that all cases could be conveyed via a single, fairly simple graph (Fig. 1). This highlights the relatively poor state of current knowledge about the structural biology of SARS-CoV-2. Based on the results presented in Fig. 1, we could divide the 27 SARS-CoV-2 proteins into four distinct categories we called teams, hijackers, mimics, and suspects - these categories are outlined in more detail below.

Teams

In our analysis, only six SARS-CoV-2 proteins had direct structural evidence of binding to other viral proteins (Fig. 1). These could be divided into two distinct teams, each with three proteins.

Below, we outline how proteins in each of these teams may interact, based on the oligomeric states observed across all 54 matching structures found in this analysis, for these six proteins.

Team 1. This team comprised nsp7, nsp8, and nsp12, which assemble to form the viral RNA synthesis complex. For nsp7, it occurred as a monomer in only two of the 15 available matching structures. For nsp8, it did not occur alone as a monomer in any of the 14 matching structures, but always in complex with either nsp12 only (1 structure), nsp7 only (5 structures), or both nsp7 and nsp12 (8 structures). By contrast, nsp12 had 38 matching structures in which it was not bound to either nsp7 or nsp8. This is consistent with previous studies on SARS-CoV that have established that nsp12 alone has RNA-dependent RNA polymerase activity, but this activity is greatly stimulated by cooperative interactions with nsp7 and nsp8 (Kirchdoerfer and Ward, 2019).

Team 2. This team comprised nsp10, nsp14, and nsp16. Interestingly, of the 30 structures found to match either nsp14 or nsp16, all were heterodimers containing exactly one other viral protein, which always matched to nsp10. This is consistent with previous studies on SARS-CoV showing that the RNA-cap (nucleoside-2′-O-)-methyltransferase activity of nsp14 requires the presence of nsp10 (Decroly et al., 2011). Similarly, SARS-CoV nsp10 has also been shown to enhance both the 2′-O–-methyltransferase and N-terminal exoribonuclease activities of SARS-CoV nsp16 (Ma et al., 2015). However, for nsp10, we found four additional matching structures where it occurred by itself as a homooligomer. From these matching structures, we observed that a common region of nsp10 was used in binding to either itself, nsp14, or nsp16. This suggests that interactions between nsp10, nsp14, and nsp16 are competitive - in contrast to the cooperative interactions seen for team 1 - and thus we speculate that nsp10 availability could be rate limiting for infection.

Finally, it is noteworthy that no interactions were seen between any of the 12 viral proteins known to form part of the mature virus assembly (i.e., all proteins on the bottom third of Fig. 1). This, again, highlights the overall poor state of knowledge about SARS-CoV-2 structural biology.

Hijackers

Our analysis found only two SARS-CoV-2 proteins with direct structural evidence of hijacking of human proteins. These hijackers are indicated in Fig. 1 as dark gray nodes connected via dark gray, dashed lines to green nodes (representing viral proteins), and are described in detail below.

Nsp3. Our analysis showed that the nsp3 PL-Pro domain may hijack the ubiquitin precursors UBB and UBC - both separately and in combination - as well as the ubiquitin-like ISG15. The evidence for these interactions is quite strong, as it is based on experimentally derived 3D structures for the closely related SARS-CoV nsp3 (~82% sequence identity, E ~ 10^{-70}-10^{-62}). While it is useful that our resource makes clear the evidence for these hijackings, the deubiquitinating activity of coronavirus nsp3 is well known as a mechanism for innate immune suppression during infection (Barretto et al., 2005). A less obvious outcome of our analysis was finding evidence that the nsp3 PL-Pro domain may also hijack the ubiquitin-60S ribosomal

protein L40, encoded by the UBA52 gene, thus suggesting a potentially novel mechanism by which host ribosomal functions may be hijacked during COVID-19 infection. Here, the evidence is based on a more remote match to MERS-CoV nsp3 (P0DTC1/4rf0, 29% identity), but with a comparable level of significance ($E = 10^{-65}$).

Spike glycoprotein. Our analysis found 16 structural states capturing the well-known hijacking of ACE2 by the spike glycoprotein (P0DTC2/6acg from SARS-CoV, 77% identity, $E = 10^{-321}$); in one of these structures, ACE2 is also bound to SLC6A19 (P0DTC2/6m17 from SARS-CoV-2, 100% identity). A less obvious outcome of our analysis was finding evidence that the spike glycoprotein may also hijack the cell surface glycoprotein receptor DPP4, which plays a key role in T-cell activation, thus suggesting a potential mechanism by which SARS-CoV-2 may defend itself against the host immune system. Here again, the evidence is based on a remote match to the spike glycoprotein from BtCoV-HKU4 (P0DTC1/4rf0, 21% identity), but nonetheless is assessed to be highly significant ($E = 10^{-45}$). Finally, a potentially useful feature of our analysis is the identification of 68 matching structures showing the spike glycoprotein in complex with antibodies; by providing easy access to this wealth of structural detail on potential therapeutic agents, our resource may help researchers in developing antibody-based interventions for COVID-19.

Mimics

We found direct structural evidence for mimicry of human proteins in only two SARS-CoV-2 proteins. These mimics are indicated in Fig. 1 using orange-colored nodes, and are described in detail below.

Nsp3. Our analysis found that the macro domain of nsp3 may mimic the macro domains found in nine human proteins that each perform specific roles associated with the post-translational modification of proteins via ADP-ribosylation (Hottiger, 2015; O'Sullivan et al., 2019). For example, PARP14, PARP15, and possibly PARP9 are ADP-ribose writers; GDAP2 is a reader, while MACROD1, MACROD2, and OARD1 completely remove ADP-ribose from D and E amino acids on proteins. Interestingly, in our analysis, the macro domains for all nine proteins had no structural evidence of direct interactions with any other host proteins. Nonetheless, some of these macro domains are part of multidomain proteins and have extensive interactions with other proteins. Two particularly interesting examples are the core histone macro-H2A.1 and macro-H2A.2 proteins (encoded by the MACROH2A1 and MACROH2A2 genes, respectively); we speculate that the nsp3 macro domain may mimic functions of these core histone proteins, and thereby hijack regulation of histone modifications as part of COVID-19 infection. An ability to directly affect a cell's epigenomic state could shed light on the highly variable response seen in patients, currently one of the key unanswered questions in COVID research (Callaway et al., 2020). In addition, PARP9 and PARP14 are known to cross-regulate macrophage activation (Iwata et al., 2016), one of the key steps in vascular disease. Given emerging evidence that COVID-19 often progresses from a respiratory illness to deadly vascular disorders (Varga et al., 2020), we speculate that the nsp3 macro domain mimicry of PARP9 and PARP14 may be a potential mechanism that contributes to these vascular disorders.

Nsp13. We found evidence that nsp13 (a.k.a. viral helicase) may mimic four human helicase proteins, all associated with DNA repair, but no evidence for mimicry of the ~100 other human helicases (Umate et al., 2011). One of the potentially mimicked helicases, encoded by AQR, was most often found bound to the spliceosome complex (P0DTD1/6id0) where it is implicated in exon ligation (Zhang et al., 2019) and DNA recombination (Sakasai 2017). A second potentially mimicked helicase, encoded by PIF1, has an intrinsic strand annealing activity and is implicated in telomere maintenance. IAQR and PIF1 match to two separate domains on nsp13, while the other potentially mimicked proteins (encoded by IGHMBP2 and UPF1) both match to both of these domains. IGHMBP2 encodes immunoglobulin mu-binding protein 2 (a.k.a. SMBP2) which, in turn, binds IGMH, the protein comprising the constant region of immunoglobulin heavy chains; this suggests another potential mechanism by which infection may hijack the cell's immune response. Like AQR, IGHMBP2 is also implicated in DNA recombination; together with the strand annealing activity of PIF1, this suggests that nsp13 mimicry of these proteins may hijack other host proteins to assist in viral recombination, which is a key driver of coronavirus evolution (Graham and Baric, 2010). UPF1 encodes a protein known as the regulator of nonsense transcripts 1 (or RENT1), which, like PIF1, is involved in telomere maintenance. This potential mimicry of telomere-associated proteins suggests a possible mechanism underlying the connections seen between COVID-19 severity, age, and telomere length (Aviv, 2020). In addition, RENT1 is a key component of the nonsense-mediated mRNA decay pathway, which can act directly against coronavirus infection (Wada et al., 2018), so viral mimicry of this protein to hijack this pathway may be a critical step during COVID-19 infection.

Suspects

This leaves 17 of the 27 viral proteins in a final group we could call 'suspects': i.e., proteins believed to perform key roles in infection, but where there is currently no structural evidence showing the mechanism of hijacking, mimicry, or of interaction with other viral proteins. We further divided the suspects into two sub-groups, based on whether each protein had any matching structures; the sub-groups are described below.

Sub-group 1. These are proteins where our analysis found at least one matching structure, but also did not find any mimicry, hijacking, or interactions with other viral proteins. This sub-group comprised nsp1, nsp2, nsp4, nsp9, nsp15, E, ORF7a, and ORF9a. This sub-group includes some very well studied proteins with well documented functional roles. For example, nsp5 (a.k.a. 3CL-Pro or 3C-like protease) is believed to be the main protease responsible for cleaving the viral polyproteins. Currently, however, our study shows that there is no structural evidence revealing how any of the suspect proteins interact with any other proteins (viral or host). From our analysis, we can not only say that these interactions have not been determined by structural biology method: we can make the much stronger statement that no protein-protein interactions involving any proteins detectably similar have been observed to date by experimental structure determination methods - at least, based on the sequence-based homology modeling methods used in this analysis.

Sub-group 2. These are proteins where our analysis found no matching structures - i.e., structurally dark proteins - and did not find any evidence of mimicry, hijacking, or interactions with other viral proteins. This sub-group comprised nsp6, ORF3a protein, matrix glycoprotein (a.k.a. M protein), ORF6 protein, ORF8 protein, ORF10 protein, and ORF14 protein.

As above, the lack of matching structures in this study does not just mean that these particular proteins have not been determined by structural biology methods. From our analysis, we can conclude that the sequences of these proteins are not detectably similar to any protein that have been observed to date by experimental structure determination methods - at least, based on the sequence-based homology modeling methods used in this analysis. Thus, these proteins are ripe candidates for more sophisticated structure modelling methods, e.g., methods based on predicted residue-residue contacts combined with deep learning (e.g., Senior et al. 2020).

Conclusions

In summary, our resource provides researchers with a wealth of information on the molecular mechanisms of COVID-19; the information can easily be accessed, and, to the best of our knowledge, is currently not available at other resources. The resource provides an immediate visual overview of what is known - and not known - about the 3D structure of the viral proteome, thereby helping direct future research. An accompanying video (https://youtu.be/J2nWQTlJNaY) explains how to use the resource and some of the novel insights gained into COVID infection. The COVID-19 models - together with 32,717 sequence features - are available at https://aquaria.ws/covid19.

METHODS

SARS-CoV-2 Sequences

This study was based on the 14 protein sequences provided in UniProtKB/Swiss-Prot version 2020_03 (released April 22, 2020; https://www.uniprot.org/statistics/) as comprising the SARS-CoV-2 proteome. Swiss-Prot provides polyproteins 1a and 1ab (a.k.a. pp1a and pp1ab) as two separate entries, both identical for the first 4401 residues; pp1a then has four additional residues ('GFAV') not in pp1ab, which has 2695 additional residues not in pp1a. Swiss-Prot also indicates residue positions at which the polyproteins become cleaved in the cell, resulting in 16 protein fragments, named nsp1 though nsp16. The nsp11 fragment, which comprises the last 13 residues of pp1a (4393–4405), concatenates with the first 919 residues of pp1b to form nsp12. Thus, following cleavage, the proteome comprises a final total of 27 separate proteins.

Sequence-to-Structure Alignments

The 14 SARS-CoV-2 sequences were then systematically compared with sequences derived from all known 3D structures from all organisms, based on using all available PDB by May 30, 2020.

At its core, Aquaria (O'Donoghue et al., 2015) relies on aligning sequences of unknown structure to sequences with known structure. Below, we describe the steps used to create the underlying database of protein sequence-to-structure homologies, PSSH2, which is based on HHblits (Steinegger et al., 2019), an alignment method employing iterative comparisons of hidden Markov models (HMMs). HHblits is the key method used in HHpred (Zimmermann et al., 2018), a fully automated server for template-based structure prediction, that was ranked best out of 79 similar servers at the CASP9 competition in 2009 (http://bit.ly/hhblits-casp9). In later years the method has been integrated into other prediction tools and is now part of the CASP evaluation process (Abriata et al., 2019). Thus, we selected HHblits as it combines both speed and reliable detection of structural templates. Since the development of Aquaria, HHblits has seen a few updates (Steinegger et al., 2019) which have further increased its speed and sensitivity. Also, the non-redundant protein database provided by HHsuite, which is used to build sequence profiles, has been changed from UniProt20 to UniClust30, for which clusterings show a high consistency of functional annotation, owing to an optimised clustering pipeline (Mirdita et al., 2017). In order to ensure that PSSH2 has maintained its specificity and sensitivity compared to results published previously, we ran a validation of the alignments. Since the COPS database (Suhrer et al., 2009) used previously has unfortunately been discontinued, we used CATH (Dawson et al., 2017) instead. In particular, we used a test data set comprising 23,028 sequences from the CATH nr40 data set[1], built individual sequence profiles against UniClust30 and used these profiles to search against "PDB_full", a database of HMMs for all PDB sequences. We then evaluated how many false positives were retrieved at an E-value of 10^{-10}, where a false positive was seen to be a structure with a different CATH code at the level of Homologous superfamily (H) or Topology (T). We compared the ratio of false positives received with HH-suite3 and UniClust30 with a similar analysis for data produced in 2017 with HH-suite2 and UniProt20, and found that in both cases the false positive rate was at 2.5% at the homology level (H), and 1.9% at the topology level (T). The recovery rate, i.e. the ratio of proteins from the CATH nr40 data (with less than 40% sequence identity) found by our method that have the same CATH code, was slightly higher with HH-suite3 (20.8% vs. 19.4%).

For each sequence-to-structure, the Aquaria interface gives a pairwise sequence identity score, thus providing an intuitive indication of how closely related the given region of SARS-CoV-2 is to the sequence of the matched structure. However, to more accurately assess the quality of the match, Aquaria also gives an E-value, calculated by comparing two HMMs, one generated for each of these two sequences.

PredictProtein Features

To facilitate analysis of SARS-CoV-2 sequences, we enhanced the Aquaria resource to include PredictProtein features (Yachdav et al., 2014), thus providing a very rich set of predicted

[1] The non-redundant data sets contain a non-redundant subset of CATH domains that: * have no pair of domains (according to BLAST) with >= 20 or 40% sequence identity (depending on the data set chosen), over 60% overlap (over the longer sequence * is as big as we could make it otherwise. -- see https://www.cathdb.info/download

features for all Swiss-Protein protein sequences. In Aquaria, the PredictProtein feature collection is fetched directly by the browser via:

```
https://api.predictprotein.org/v1/results/molart/:uniprot_id
```

The PredictProtein feature sets used in the analysis presented in this work are specified below.

Conservation. This feature set is generated by ConSurf (Ashkenazy et al., 2010; Celniker et al., 2013) and estimates the evolutionary rate in protein families, based on evolutionary relatedness between the query protein and its homologues from UniProt using empirical Bayesian methods (Mayrose et al., 2004). The strength of these methods is that they rely on the phylogeny of the sequences and thus can accurately distinguish between conservation due to short evolutionary time and conservation resulting from importance for maintaining protein foldability and function.

Disordered Regions. This feature set gives consensus predictions generated by Meta-Disorder (Schlessinger and Rost, 2005), which combines outputs of several structure-based disorder predictors (Schlessinger et al., 2009, 2007a, 2007b, 2006).

Relative B-values. This feature set predicts, for each residue, the B-factor (a.k.a. Debye–Waller factor (Debye, 1913; Waller, 1923)) that would be observed in an X-ray-derived structure. The predictions were generated by PROFbval (Schlessinger et al., 2006). Large B-factors are generally believed to indicate parts of a protein that are very flexible.

Topology. This feature set is generated by TMSEG (Bernhofer et al., 2016), a machine learning model that uses evolutionary-derived information to predict regions of a protein that traverse membranes, as well as the subcellular locations of the complementary (non-transmembrane) regions.

SNAP2 Features

We further enhanced Aquaria to include SNAP2 features, which provides details on the mutational propensities for each residue position (Hecht et al., 2015). In Aquaria, the SNAP2 feature collection for each protein sequence is fetched directly by the browser via:

```
https://rostlab.org/services/aquaria/snap4aquaria/json.php?uniprotAcc=:uniprot_id
```

Two SNAP2 feature sets were used in this work:

Mutational Sensitivity. For each residue position, this feature set provides 20 scores indicating the predicted functional consequences of the position being occupied by each of the 20 standard amino acids. Large, positive scores (up to 100) indicate substitutions likely to have deleterious changes, while negative scores (down to -100) indicate no likely functional change. From these 20 values a single summary score is calculated based on the total fraction of substitutions predicted to have a deleterious effect on function, taken to be those with a score > 40. The summary scores are used to generate a red to blue color map, indicating residues with highest to least functional importance, respectively.

Mutational Score. This feature set is based on the same 20 scores above, but calculates the single summary score for each residue as the average of the individual scores for each of the 20 standard amino acids.

UniProt Features

UniProt features are curated annotations, and therefore largely complement the automatically generated PredictProtein features. In Aquaria, for each protein sequence, the UniProt feature collection is fetched directly by the browser via:

```
https://www.uniprot.org/uniprot/:uniprot_id.xml
```

CATH Features

For this work, we further enhanced Aquaria to include CATH domain annotations (Dawson et al., 2017). For most protein sequences, Aquaria fetches these annotations directly from the browser via APIs given at:

```
https://github.com/UCLOrengoGroup/cath-api-docs
```

For SARS-CoV-2 proteins, however, CATH annotations are not yet fully available via the above APIs. In this work, we used a pre-release version of these annotations, derived by scanning the UniProt pre-release sequences against the CATH-Gene3D v4.3 FunFams HMM library (Dawson et al., 2017; Lewis et al., 2018) using HMMsearch with inclusion thresholds cut-offs (Mistry et al., 2013)**.** Domain assignments were obtained using cath-resolve-hits and curated manually (Lewis et al., 2019). For the SARS-CoV-2 sequences, these data are fetched directly from the browser via:

```
https://aquaria.ws/covid19cath/P0DTC2
```

Two CATH feature sets were used in this work:

Superfamilies. These identify regions of protein sequences across a wide variety of organisms that are expected to have very similar 3D structure and to have general biological functions in common.

Functional Families. Also known as FunFams, these domains partition each superfamily into subsets expected to have more specific biological functions in common (Dawson et al., 2017; Lewis et al., 2018).

When examining a specific superfamily or functional family domain in Aquaria, the browser uses additional CATH API endpoints (see link above) to create compact, interactive data visualizations that give access to detailed information on the biological function and phylogenetic distribution of proteins containing this domain.

Matrix Layout

We created a web page featuring a matrix layout that summarizes the total number of structural matches found for each viral protein sequence, and allows navigation to the corresponding

Aquaria page for each protein. The structures shown on the Matrix page are served as static, two-dimensional images to optimize page load time. They are created at three sizes: 2000 pixels wide for extreme high-resolution displays, 1000, and 500 pixels wide. Images are saved in JPEG and WEBP formats to minimize file sizes, and the smallest appropriate versions are served based on the visitor's display resolution and browser.

Genome Layout

We create an additional web page with a layout derived from the organization of the viral genome. Each viral protein or domain typically has many matching structures; from these, we selected one representative structure to highlight in Fig. 1 and in the *Results* section. This selection was primarily based on which structure had highest identity to the SARS-CoV-2 sequence, or, in the case of matching identity, which structure had highest resolution. However, in some cases, representatives were chosen that did not have the highest identity, but best illustrated the consensus biological assembly seen across all related matching structures or showed the simplest assembly (e.g., for nsp7 in Fig. 1, we selected P0DTC1/3ub0, which showed two copies of nsp7 in complex with two copies of nsp8).

Under the name of each viral protein, the total number of matching structures found in PSSH2 (O'Donoghue et al., 2015), is indicated. The image of a single structure is displayed, which was manually selected as best representing the overall fold and oligomeric assembly. Below each structure, a tree is drawn to show the number of structures in which the viral sequence aligns onto human proteins (via PSSH2). Below that, another tree shows the number of structures showing interactions between the viral protein and human proteins, DNA, RNA, antibodies, or inhibitory factors. When these trees are missing, none of the matching structures meet these criteria. These criteria were used to highlight structures that potentially show the viral protein mimicking human proteins and hijacking their interactions.

ACKNOWLEDGEMENTS

Thanks to Tim Mercer (Garvan Institute, Australia) and Lucy van Dorp (UCL Genetics Institute, UK) for useful discussions, and Ian Sillitoe (UCL, UK) or helpful advice regarding the CATH API. We are very grateful to Max Ott (CSIRO, Australia) for detailed advice on improving the performance and reliability of the Aquaria web application, and to Tim Karl (TU Munich, Germany) for contributing towards this same goal, even during parental leave.

REFERENCES

Abriata, L.A., Tamò, G.E., Peraro, M.D., 2019. A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments. Proteins Struct. Funct. Bioinforma. 87, 1100–1112. https://doi.org/10.1002/prot.25787

Almeida, M.S., Johnson, M.A., Herrmann, T., Geralt, M., Wüthrich, K., 2007. Novel β-Barrel Fold in the Nuclear Magnetic Resonance Structure of the Replicase Nonstructural Protein 1

from the Severe Acute Respiratory Syndrome Coronavirus. J. Virol. 81, 3151–3161. https://doi.org/10.1128/JVI.01939-06
Ashkenazy, H., Erez, E., Martz, E., Pupko, T., Ben-Tal, N., 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. 38, W529–W533. https://doi.org/10.1093/nar/gkq399
Aviv, A., 2020. Telomeres and COVID‐19. FASEB J. 34, 7247–7252. https://doi.org/10.1096/fj.202001025
Barretto, N., Jukneliene, D., Ratia, K., Chen, Z., Mesecar, A.D., Baker, S.C., 2005. The papain-like protease of severe acute respiratory syndrome coronavirus has deubiquitinating activity. J. Virol. 79, 15189–15198. https://doi.org/10.1128/JVI.79.24.15189-15198.2005
Bernhofer, M., Kloppmann, E., Reeb, J., Rost, B., 2016. TMSEG: Novel prediction of transmembrane helices. Proteins Struct. Funct. Bioinforma. 84, 1706–1716. https://doi.org/10.1002/prot.25155
Callaway, E., Ledford, H., Mallapaty, S., 2020. Six months of coronavirus: the mysteries scientists are still racing to solve. Nature 583, 178–179. https://doi.org/10.1038/d41586-020-01989-z
Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., Ben‐Tal, N., 2013. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. Isr. J. Chem. 53, 199–206. https://doi.org/10.1002/ijch.201200096
Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G., Lee, D., Ashford, P., Orengo, C.A., Sillitoe, I., 2017. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 45, D289–D295. https://doi.org/10.1093/nar/gkw1098
Debye, P., 1913. Interferenz von Röntgenstrahlen und Wärmebewegung. Ann. Phys. 348, 49–92. https://doi.org/10.1002/andp.19133480105
Decroly, E., Debarnot, C., Ferron, F., Bouvet, M., Coutard, B., Imbert, I., Gluais, L., Papageorgiou, N., Sharff, A., Bricogne, G., Ortiz-Lombardia, M., Lescar, J., Canard, B., 2011. Crystal Structure and Functional Analysis of the SARS-Coronavirus RNA Cap 2′-O-Methyltransferase nsp10/nsp16 Complex. PLoS Pathog. 7, e1002059. https://doi.org/10.1371/journal.ppat.1002059
Graham, R.L., Baric, R.S., 2010. Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. J. Virol. 84, 3134–3146. https://doi.org/10.1128/JVI.01394-09
Hecht, M., Bromberg, Y., Rost, B., 2015. Better prediction of functional effects for sequence variants. BMC Genomics 16, S1. https://doi.org/10.1186/1471-2164-16-S8-S1
Heinrich, J., Kaur, S., O'Donoghue, S., 2015. Evaluating the Effectiveness of Color to Convey Alignment Quality in Macromolecular Structures. Presented at the Symposium on Big Data Visual Analytics, IEEE, Hobart, Australia.
Heo, L., Feig, M., 2020. Modeling of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Proteins by Machine Learning and Physics-Based Refinement. bioRxiv 2020.03.25.008904. https://doi.org/10.1101/2020.03.25.008904
Hottiger, M.O., 2015. SnapShot: ADP-Ribosylation Signaling. Mol. Cell 58, 1134-1134.e1. https://doi.org/10.1016/j.molcel.2015.06.001
Iwata, H., Goettsch, C., Sharma, A., Ricchiuto, P., Goh, W.W.B., Halu, A., Yamada, I., Yoshida, H., Hara, T., Wei, M., Inoue, N., Fukuda, D., Mojcher, A., Mattson, P.C., Barabási, A.-L., Boothby, M., Aikawa, E., Singh, S.A., Aikawa, M., 2016. PARP9 and PARP14 cross-regulate macrophage activation via STAT1 ADP-ribosylation. Nat. Commun. 7, 12849. https://doi.org/10.1038/ncomms12849

Jaimes, J.A., André, N.M., Chappie, J.S., Millet, J.K., Whittaker, G.R., 2020. Phylogenetic Analysis and Structural Modeling of SARS-CoV-2 Spike Protein Reveals an Evolutionary Distinct and Proteolytically Sensitive Activation Loop. J. Mol. Biol. 432, 3309–3325. https://doi.org/10.1016/j.jmb.2020.04.009

Johnson, M.A., Jaudzems, K., Wüthrich, K., 2010. NMR Structure of the SARS-CoV Nonstructural Protein 7 in Solution at pH 6.5. J. Mol. Biol. 402, 619–628. https://doi.org/10.1016/j.jmb.2010.07.043

Kawabata, T., 2016. HOMCOS: an updated server to search and model complex 3D structures. J. Struct. Funct. Genomics 17, 83–99. https://doi.org/10.1007/s10969-016-9208-y

Kirchdoerfer, R.N., Ward, A.B., 2019. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. Nat. Commun. 10, 2342. https://doi.org/10.1038/s41467-019-10280-3

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., Moult, J., 2019. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins Struct. Funct. Bioinforma. 87, 1011–1020. https://doi.org/10.1002/prot.25823

Lei, J., Kusov, Y., Hilgenfeld, R., 2018. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. Antiviral Res. 149, 58–74. https://doi.org/10.1016/j.antiviral.2017.11.001

Lewis, T.E., Sillitoe, I., Dawson, N., Lam, S.D., Clarke, T., Lee, D., Orengo, C., Lees, J., 2018. Gene3D: Extensive prediction of globular domains in proteins. Nucleic Acids Res. 46, D435–D439. https://doi.org/10.1093/nar/gkx1069

Lewis, T.E., Sillitoe, I., Lees, J.G., 2019. cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly. Bioinformatics 35, 1766–1767. https://doi.org/10.1093/bioinformatics/bty863

Ma, Y., Wu, L., Shaw, N., Gao, Y., Wang, J., Sun, Y., Lou, Z., Yan, L., Zhang, R., Rao, Z., 2015. Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. Proc. Natl. Acad. Sci. 112, 9436–9441. https://doi.org/10.1073/pnas.1508686112

Mayrose, I., Graur, D., Ben-Tal, N., Pupko, T., 2004. Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior. Mol. Biol. Evol. 21, 1781–1791. https://doi.org/10.1093/molbev/msh194

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., Söding, J., Steinegger, M., 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 45, D170–D176. https://doi.org/10.1093/nar/gkw1081

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., Punta, M., 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 41, e121. https://doi.org/10.1093/nar/gkt263

O'Donoghue, S.I., Baldi, B.F., Clark, S.J., Darling, A.E., Hogan, J.M., Kaur, S., Maier-Hein, L., McCarthy, D.J., Moore, W.J., Stenau, E., Swedlow, J.R., Vuong, J., Procter, J.B., 2018. Visualization of Biomedical Data. Annu. Rev. Biomed. Data Sci. 1, 275–304. https://doi.org/10.1146/annurev-biodatasci-080917-013424

O'Donoghue, S.I., Sabir, K.S., Kalemanov, M., Stolte, C., Wellmann, B., Ho, V., Roos, M., Perdigão, N., Buske, F.A., Heinrich, J., Rost, B., Schafferhans, A., 2015. Aquaria: simplifying discovery and insight from protein structures. Nat. Methods 12, 98–99. https://doi.org/10.1038/nmeth.3258

O'Sullivan, J., Tedim Ferreira, M., Gagné, J.-P., Sharma, A.K., Hendzel, M.J., Masson, J.-Y., Poirier, G.G., 2019. Emerging roles of eraser enzymes in the dynamic control of protein ADP-ribosylation. Nat. Commun. 10, 1182. https://doi.org/10.1038/s41467-019-08859-x

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K.S., Buckley, M.J., Tabor, B., Signal, B., Gloss,

B.S., Hammang, C.J., Rost, B., Schafferhans, A., O'Donoghue, S.I., 2015. Unexpected features of the dark proteome. Proc. Natl. Acad. Sci. 112, 15898–15903. https://doi.org/10.1073/pnas.1508380112

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., Baker, D., 2004. Protein Structure Prediction Using Rosetta, in: Methods in Enzymology, Numerical Computer Methods, Part D. Academic Press, pp. 66–93. https://doi.org/10.1016/S0076-6879(04)83004-0

Schafferhans, A., O'Donoghue, S.I., Heinzinger, M., Rost, B., 2018. Dark Proteins Important for Cellular Function. PROTEOMICS 18, 1800227. https://doi.org/10.1002/pmic.201800227

Schlessinger, A., Liu, J., Rost, B., 2007a. Natively Unstructured Loops Differ from Other Loops. PLOS Comput. Biol. 3, e140. https://doi.org/10.1371/journal.pcbi.0030140

Schlessinger, A., Punta, M., Rost, B., 2007b. Natively unstructured regions in proteins identified from contact predictions. Bioinformatics 23, 2376–2384. https://doi.org/10.1093/bioinformatics/btm349

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., Rost, B., 2009. Improved Disorder Prediction by Combination of Orthogonal Approaches. PLOS ONE 4, e4433. https://doi.org/10.1371/journal.pone.0004433

Schlessinger, A., Rost, B., 2005. Protein flexibility and rigidity predicted from sequence. Proteins Struct. Funct. Bioinforma. 61, 115–126. https://doi.org/10.1002/prot.20587

Schlessinger, A., Yachdav, G., Rost, B., 2006. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 22, 891–893. https://doi.org/10.1093/bioinformatics/btl032

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D.T., Silver, D., Kavukcuoglu, K., Hassabis, D., 2020. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710. https://doi.org/10.1038/s41586-019-1923-7

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., Söding, J., 2019. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20, 473. https://doi.org/10.1186/s12859-019-3019-7

Su, D., Lou, Z., Sun, F., Zhai, Y., Yang, H., Zhang, R., Joachimiak, A., Zhang, X.C., Bartlam, M., Rao, Z., 2006. Dodecamer Structure of Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein nsp10. J. Virol. 80, 7902–7908. https://doi.org/10.1128/JVI.00483-06

Suhrer, S.J., Wiederstein, M., Gruber, M., Sippl, M.J., 2009. COPS—a novel workbench for explorations in fold space. Nucleic Acids Res. 37, W539–W544. https://doi.org/10.1093/nar/gkp411

Umate, P., Tuteja, N., Tuteja, R., 2011. Genome-wide comprehensive analysis of human helicases. Commun. Integr. Biol. 4, 118–137. https://doi.org/10.4161/cib.4.1.13844

Varga, Z., Flammer, A.J., Steiger, P., Haberecker, M., Andermatt, R., Zinkernagel, A.S., Mehra, M.R., Schuepbach, R.A., Ruschitzka, F., Moch, H., 2020. Endothelial cell infection and endotheliitis in COVID-19. The Lancet 395, 1417–1418. https://doi.org/10.1016/S0140-6736(20)30937-5

Wada, M., Lokugamage, K.G., Nakagawa, K., Narayanan, K., Makino, S., 2018. Interplay between coronavirus, a cytoplasmic RNA virus, and nonsense-mediated mRNA decay pathway. Proc. Natl. Acad. Sci. 115, E10157–E10166. https://doi.org/10.1073/pnas.1811675115

Waller, I., 1923. Zur Frage der Einwirkung der Wärmebewegung auf die Interferenz von Röntgenstrahlen. Z. Für Phys. 17, 398–408. https://doi.org/10.1007/BF01328696

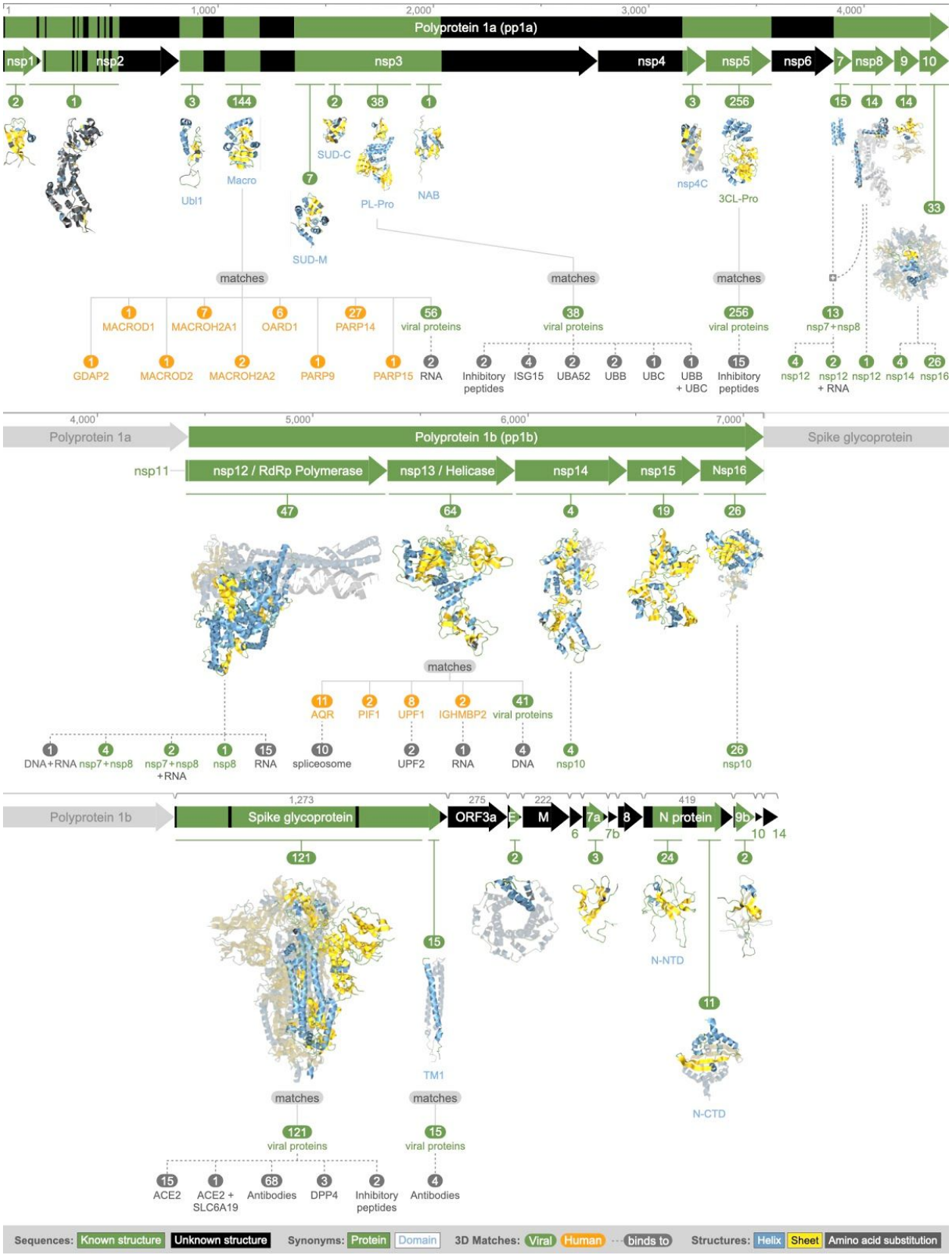
Wang, L., Hu, W., Fan, C., 2020. Structural and biochemical characterization of SADS-CoV

papain-like protease 2. Protein Sci. 29, 1228–1241. https://doi.org/10.1002/pro.3857
Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46, W296–W303. https://doi.org/10.1093/nar/gky427
Xu, X., Lou, Z., Ma, Y., Chen, X., Yang, Z., Tong, X., Zhao, Q., Xu, Y., Deng, H., Bartlam, M., Rao, Z., 2009. Crystal Structure of the C-Terminal Cytoplasmic Domain of Non-Structural Protein 4 from Mouse Hepatitis Virus A59. PLoS ONE 4, e6217. https://doi.org/10.1371/journal.pone.0006217
Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., Hönigschmid, P., Schafferhans, A., Roos, M., Bernhofer, M., Richter, L., Ashkenazy, H., Punta, M., Schlessinger, A., Bromberg, Y., Schneider, R., Vriend, G., Sander, C., Ben-Tal, N., Rost, B., 2014. PredictProtein—an open resource for online prediction of protein structural and functional features. Nucleic Acids Res. 42, W337–W343. https://doi.org/10.1093/nar/gku366
Zhang, X., Zhan, X., Yan, C., Zhang, W., Liu, D., Lei, J., Shi, Y., 2019. Structures of the human spliceosomes before and after release of the ligated exon. Cell Res. 29, 274–285. https://doi.org/10.1038/s41422-019-0143-x
Zheng, W., Li, Y., Zhang, C., Pearce, R., Mortuza, S.M., Zhang, Y., 2019. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins 87, 1149–1164. https://doi.org/10.1002/prot.25792
Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., Alva, V., 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J. Mol. Biol., Computation Resources for Molecular Biology 430, 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007

FIGURES

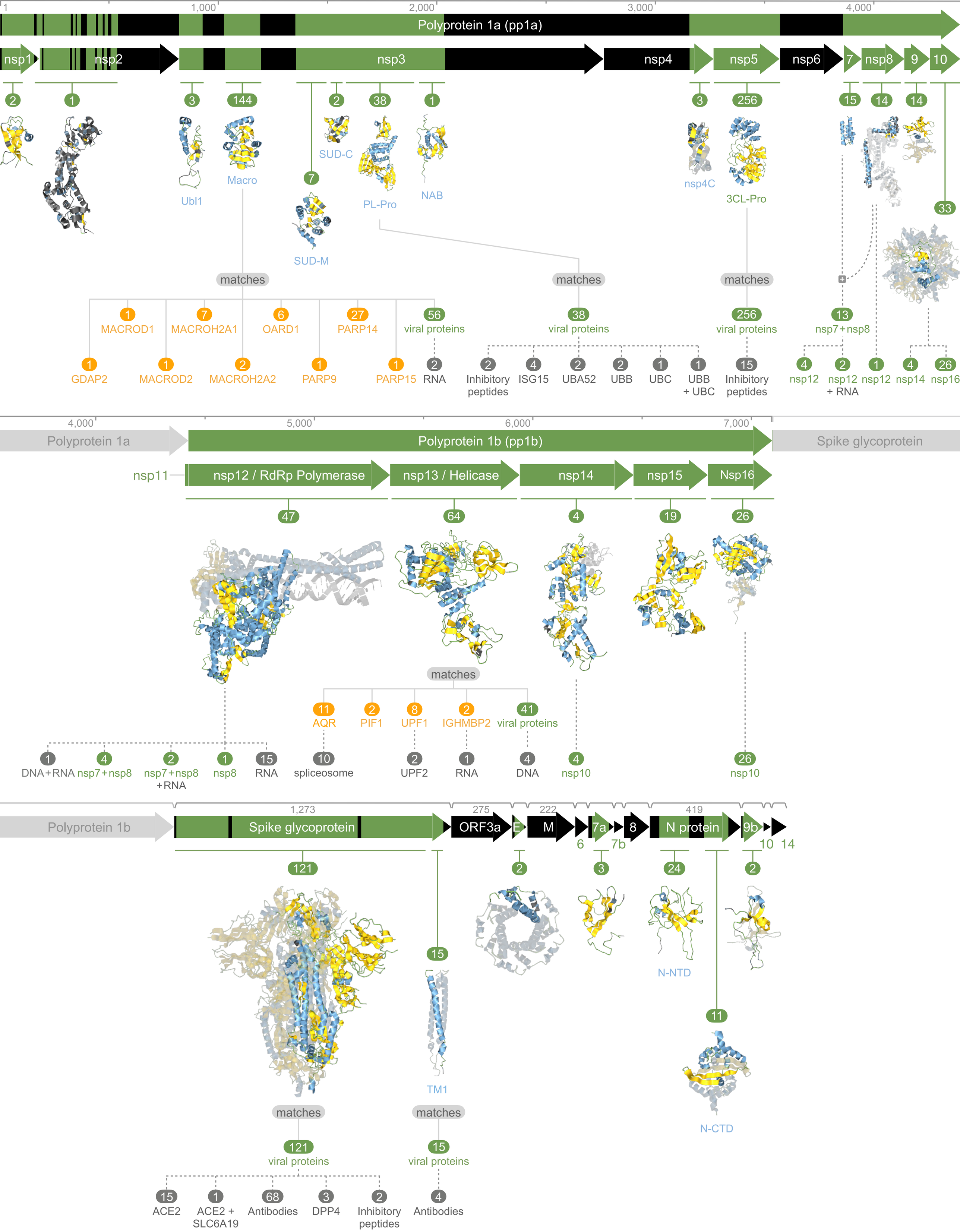
Figure 1 | Overview

Summary of all available 3D molecular structural knowledge for the viral proteome, as well as derived mimicry, hijacking, and protein interactions.


Polyprotein 1a (pp1a)
nsp1
nsp2
nsp3
nsp4
nsp5
nsp6
7
nsp8
9
10
Ubl1
Macro
SUD-C
SUD-M
PL-Pro
NAB
nsp4C
3CL-Pro
matches
MACROD1
MACROH2A1
OARD1
PARP14
viral proteins
GDAP2
MACROD2
MACROH2A2
PARP9
PARP15
RNA
Inhibitory peptides
ISG15
UBA52
UBB
UBC
UBB + UBC
nsp7+nsp8
nsp12
nsp12 + RNA
nsp14
nsp16
Polyprotein 1a
Polyprotein 1b (pp1b)
Spike glycoprotein
nsp11
nsp12 / RdRp Polymerase
nsp13 / Helicase
nsp15
Nsp16
AQR
PIF1
UPF1
IGHMBP2
DNA+RNA
nsp7+nsp8 +RNA
nsp8
spliceosome
UPF2
DNA
nsp10
Polyprotein 1b
Spike glycoprotein
ORF3a
E
M
7a
8
N protein
9b
N-NTD
TM1
N-CTD
ACE2
ACE2 + SLC6A19
Antibodies
DPP4
Sequences: Known structure Unknown structure
Synonyms: Protein Domain
3D Matches: Viral Human binds to
Structures: Helix Sheet Amino acid substitution

Protein	3D Models (grouped by sequence identity)				Interaction partners		Features
	Database	*Identical (100%)*	*Close (≥ 70%)*	*Remote (< 70%)*	*Type*	*Number*	
Envelope protein (a.k.a. E protein, vemp, envelope small membrane protein)	*PDB:*	0			*Proteins:*	0	197
	Swiss-Model:	0			*DNA/RNA:*	0	
	Aquaria:	0	1	0			
			2	0			
Matrix glycoprotein (a.k.a. M protein, vme1)	*PDB:*	0					
	Swiss-Model:	0	0	4			
	Aquaria:	0	0	0			
Nucleocapsid protein (a.k.a. N protein, ncap)	*PDB:*	4			*Proteins:*	0	937
	Swiss-Model:	2	2	0	*DNA/RNA:*	0	
	Aquaria:	4	9	22			
ORF9b protein (a.k.a. accessory protein 9b)	*PDB:*	0			*Proteins:*	0	236
	Swiss-Model:	0	1	0	*DNA/RNA:*	0	
	Aquaria:	0	4	0			
ORF7a protein (a.k.a. accessory protein 7a)	*PDB:*	1			*Proteins:*	0	290
	Swiss-Model:	1	2	0	*DNA/RNA:*	0	
	Aquaria:	1	2	0			
ORF3a protein (a.k.a. accessory protein 3a)	*PDB:*	0					
	Swiss-Model:	0	0	2			
	Aquaria:	0	0	0			
ORF6 protein (a.k.a. accessory protein 6)	*PDB:*	0					
	Swiss-Model:	0	0	3			
	Aquaria:	0	0	0			
ORF8 protein (a.k.a. accessory protein 8)	*PDB:*	0					
	Swiss-Model:	0	0	2			
	Aquaria:	0	0	0			
ORF7b protein (a.k.a. accessory protein 7b)	*PDB:*	0	0	0			
	Swiss-Model:						
	Aquaria:						
ORF10 protein	*PDB:*	0	0	0			
	Swiss-Model:						
	Aquaria:						
ORF14 protein (a.k.a. y14, uncharacterized protein)	*PDB:*	0	0	0			
	Swiss-Model:						
	Aquaria:						
Polyprotein 1a (a.k.a. replicase 1a, r1a)	*PDB:*	9			*Proteins:*	45 (38)	10,936
	Swiss-Model:	11	18	8	*DNA/RNA:*	0	
	Aquaria:	159	150	219			
Polyprotein 1ab (a.k.a. replicase 1ab, r1ab)	*PDB:*	137			*Proteins:*	163 (24)	16,656
	Swiss-Model:	14	29	12	*DNA/RNA:*	13 (0)	
	Aquaria:	179	166	340			
Spike glycoprotein (a.k.a. S, spike)	*PDB:*	14			*Proteins:*	60 (23)	3,465
	Swiss-Model:	0	4	0	*DNA/RNA:*	0	
	Aquaria:	19	42	76			

Table 1 | Summarizes structure models and sequence feature information available in Aquaria COVID-19 resource, compared with other related resources. In general, the strength of Aquaria is in providing a wealth of structures, especially for distant homologs.


1
1,000
2,000
3,000
4,000
Polyprotein 1a (pp1a)
nsp1
nsp2
nsp3
nsp4
nsp5
nsp6
7
nsp8
9
10
2
1
3
144
2
38
1
3
256
15
14
14
33
Ubl1
Macro
SUD-C
7
SUD-M
PL-Pro
NAB
nsp4C
3CL-Pro
matches
1
MACROD1
7
MACROH2A1
6
OARD1
27
PARP14
56
viral proteins
38
viral proteins
256
viral proteins
13
nsp7+nsp8
1
GDAP2
1
MACROD2
2
MACROH2A2
1
PARP9
1
PARP15
2
RNA
2
Inhibitory peptides
4
ISG15
2
UBA52
2
UBB
1
UBC
1
UBB + UBC
15
Inhibitory peptides
4
nsp12
2
nsp12 + RNA
1
nsp12
4
nsp14
26
nsp16
4,000
5,000
6,000
7,000
Polyprotein 1a
Polyprotein 1b (pp1b)
Spike glycoprotein
nsp11
nsp12 / RdRp Polymerase
nsp13 / Helicase
nsp14
nsp15
Nsp16
47
64
4
19
26
matches
11
AQR
2
PIF1
8
UPF1
2
IGHMBP2
41
viral proteins
1
DNA+RNA
4
nsp7+nsp8
2
nsp7+nsp8 +RNA
1
nsp8
15
RNA
10
spliceosome
2
UPF2
1
RNA
4
DNA
4
nsp10
26
nsp10
Polyprotein 1b
1,273
Spike glycoprotein
275
ORF3a
E
222
M
6
7a
7b
8
419
N protein
9b
10
14
121
2
3
24
2
15
TM1
N-NTD
11
N-CTD
matches
121
viral proteins
matches
15
viral proteins
15
ACE2
1
ACE2 + SLC6A19
68
Antibodies
3
DPP4
2
Inhibitory peptides
4
Antibodies
Sequences: Known structure Unknown structure
Synonyms: Protein Domain
3D Matches: Viral Human binds to
Structures: Helix Sheet Amino acid substitution