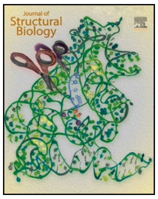
Journal of Structural Biology 212 (2020) 107617

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi



A graph-based approach identifies dynamic H-bond communication networks in spike protein S of SARS-CoV-2

Konstantina Karathanou[a,1], Michalis Lazaratos[a,1], Éva Bertalan[a], Malte Siemers[a], Krzysztof Buzar[a], Gebhard F.X. Schertler[b,c], Coral del Val[d,e,f], Ana-Nicoleta Bondar[a,*]

[a] Freie Universität Berlin, Department of Physics, Theoretical Molecular Biophysics, Arnimallee 14, D-14195 Berlin, Germany
[b] Paul Scherrer Institut, Department of Biology and Chemistry, Laboratory of Biomolecular Research, CH-5303 Villigen-PSI, Switzerland
[c] ETH Zürich, Department of Biology, 8093 Zürich, Switzerland
[d] University of Granada, Department of Computer Science and Artificial Intelligence, E-18071 Granada, Spain
[e] Instituto de Investigación Biosanitaria ibs.GRANADA, 18012 Granada, Spain
[f] Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI Institute), 18014 Granada, Spain

ARTICLE INFO

Keywords:
Bimolecular structure
Hydrogen bonding
Dynamic hydrogen-bond clusters
Conformational plasticity
SARS-CoV-2 protein S
ACE2

ABSTRACT

Corona virus spike protein S is a large homo-trimeric protein anchored in the membrane of the virion particle. Protein S binds to angiotensin-converting-enzyme 2, ACE2, of the host cell, followed by proteolysis of the spike protein, drastic protein conformational change with exposure of the fusion peptide of the virus, and entry of the virion into the host cell. The structural elements that govern conformational plasticity of the spike protein are largely unknown. Here, we present a methodology that relies upon graph and centrality analyses, augmented by bioinformatics, to identify and characterize large H-bond clusters in protein structures. We apply this methodology to protein S ectodomain and find that, in the closed conformation, the three protomers of protein S bring the same contribution to an extensive central network of H-bonds, and contribute symmetrically to a relatively large H-bond cluster at the receptor binding domain, and to a cluster near a protease cleavage site. Markedly different H-bonding at these three clusters in open and pre-fusion conformations suggest dynamic H-bond clusters could facilitate structural plasticity and selection of a protein S protomer for binding to the host receptor, and proteolytic cleavage. From analyses of spike protein sequences we identify patches of histidine and carboxylate groups that could be involved in transient proton binding.

1. Introduction

The surface of the Severe Acute Respiratory Syndrome (SARS)-CoV-2 virion is decorated with large membrane-anchored spike proteins S (Fig. 1) that bind to Angiotensin Converting Enzyme 2 (ACE2) receptor of the host cell (Briefing, 2020; Hoffmann et al., 2020; Li et al., 2003; Xiao et al., 2003; Zhou et al., 2020). Interactions between the spike protein and the host receptor, and large-scale structural rearrangements of the spike protein, are essential for virus entry (Belouzard et al., 2012).

Protein conformational plasticity, as required for large-scale rearrangements, may originate in weak hydrogen(H) bonds (Joh et al., 2008) and dynamic H-bond clusters (Bondar and White, 2012). Here, we present a graph-based methodology to characterize hydrogen(H)-bond clusters in protein S and in protein S – ACE2 interaction complexes, and use this methodology to identify interaction networks with potential role in selecting spike S protein conformations for receptor binding and proteolytic activation.

Corona spike glycoprotein S (Fig. 1) is arranged as a homotrimer (Fig. 1B), each protomer consisting of a large N-terminus ectodomain that contains domains S1 and S2 (Fig. 1A). S1 contains the Receptor Binding Domain (RBD) (Babcock et al., 2004; Xiao et al., 2003) that binds to ACE2 via an extended loop denoted as the Receptor Binding Motif (RBM) (Graham and Baric, 2010; Li, 2013). The S2 domain includes the fusion peptide (FP), the two Heptad Repeats HR1 and HR2 (also denoted as helical regions HR-A and HR-B (Colman and Lawrence, 2003)), a TM segment that anchors the protein into the viral envelope, and a relatively short C-terminus segment exposed to the interior of the virion (Fig. 1A). Exposure of the fusion peptide requires two proteolytic cleavage events, first at the boundary between S1 and S2 (the S1/S2

[*] Corresponding author.
E-mail address: nbondar@zedat.fu-berlin.de (A.-N. Bondar).
[1] Equal contribution.

https://doi.org/10.1016/j.jsb.2020.107617
Received 18 May 2020; Received in revised form 3 September 2020; Accepted 6 September 2020
Available online 10 September 2020

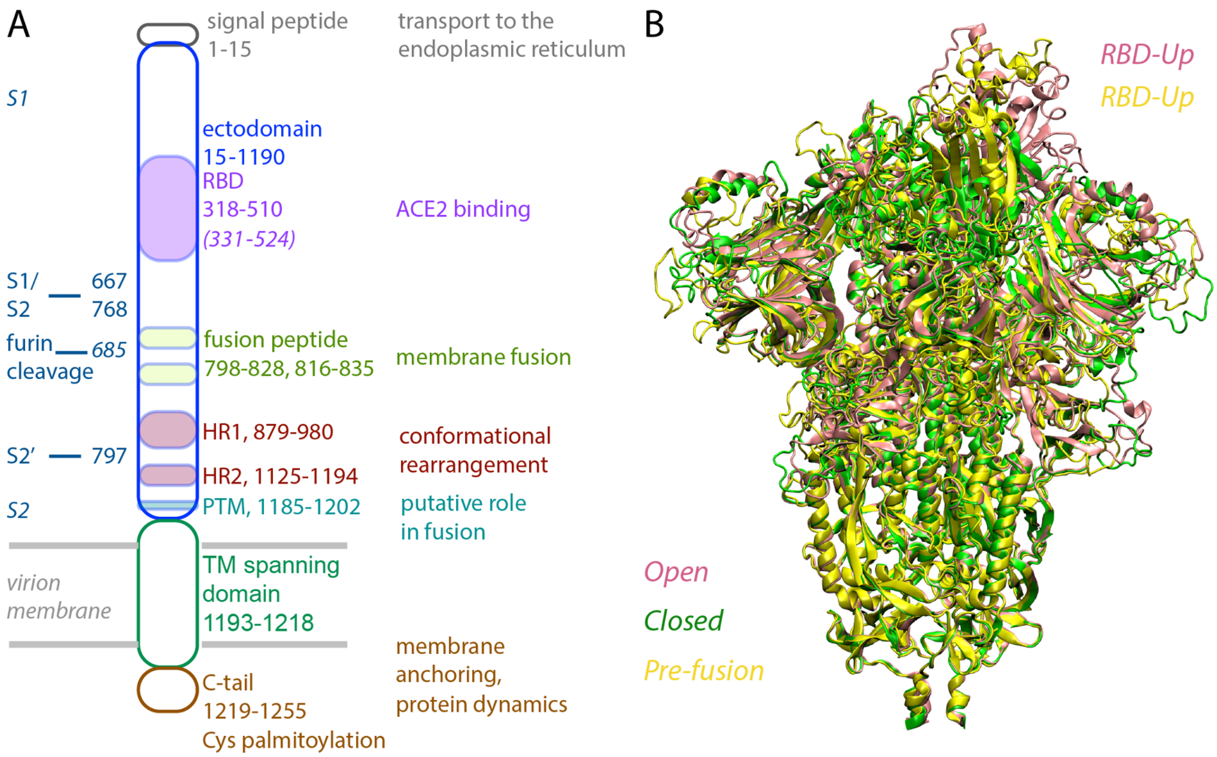


Fig. 1. Schematic representation of the domain organization of the spike protein S. (A) Protein S consists of the N-terminus ectodomain, the TM helical segment, and a C-tail. Approximate amino acid residue ranges for functional regions of SARS-CoV protein S, and proposed functional roles, are based on refs. (Babcock et al., 2004; Bosch et al., 2004; Hofmann and Pöhlmann, 2004; Li, 2013; Petit et al., 2007; Tai et al., 2020; Xiao et al., 2003). The putative location of fusion peptide regions is from ref. (Lai et al., 2017). Numbers in italics give approximate domain ranges for SARS-CoV-2 based on ref. (Tai et al., 2020). Trypsin-mediated cleavage of SARS-CoV occurs at R667 (Belouzard et al., 2009 1464), a group that is also required for cleavage by the TM serine protease TMPRSS2 (Reinke et al., 2017); cathepsin L-mediated cleavage of SARS-CoV occurs at T678 (Millet and Whittaker, 2015). R797 is at the S2′ cleavage site of SARS-CoV (Belouzard et al., 2009; Millet and Whittaker, 2015; Reinke et al., 2017). R685 at the furin cleavage site of SARS-CoV-2 (Walls et al., 2020), and R815 at the S2′ site (Jaimes et al., 2020). In ref. (Wrapp et al., 2020) the RBD is indicated as amino acid residues 330–521 for SARS-CoV-2, and 317–507 for SARS-CoV. (B) Overlap of the ectodomain of SARS-CoV-2 protein S in the open (pink), closed (green), and pre-fusion conformation (yellow). Unless specified otherwise, molecular graphics were prepared using Visual Molecular Dynamics, VMD (Humphrey et al., 1996). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

site, Fig. 1A), and subsequently at the S2′ site within the S2 domain (Belouzard et al., 2009; Millet and Whittaker, 2014; Walls et al., 2017).

Main conformations of spike proteins are typically described in terms of a native state (prior to proteolytic cleavage), pre-hairpin (after receptor binding), hairpin, and post-fusion (Eckert and Kim, 2001; Xu et al., 2004). Upon binding to the host receptor, protein S undergoes a conformational change that reduces the height of the spike by ~10 Å (Beniac et al., 2007). In the pre-hairpin intermediate of the HIV envelope protein, the transmembrane subunit gp41 spans the membranes of both the virion and host cell (Eckert and Kim, 2001). In post-fusion, structural rearrangements and refolding of the HRs bring about massive structural rearrangement of the S2 region, which lengthens from ~88 Å to ~185 Å (Walls et al., 2017).

Coordinate snapshots of the ectodomain of SARS-CoV-2 protein S captured with cryo-electron microscopy (cryo-EM) were assigned to conformations denoted as open, closed, and pre-fusion. The open and closed conformations of the ectodomain (Walls et al., 2020) are largely similar to each other, except for the RBD of one of the protomers having an *up* orientation (Fig. 1B, S2). Likewise, in the pre-fusion conformation (Wrapp et al., 2020) one of the protomers with the RBD is in the *up* conformation (Fig. 1B, S2).

Similar structural dynamics, whereby the inactive conformation of protein S is three-fold symmetrical with all three RBDs in a *down* orientation incompatible with receptor binding, and the active-like conformation is asymmetric, with one RBD in *up* orientation compatible with receptor binding, were observed before for SARS-CoV (Gui et al., 2017). In a cryo-EM study of a trimeric SARS-CoV protein S ectodomain with mutations that stabilize the pre-fusion conformation, the majority of the protein was captured with one RBD *up*, a lesser amount with two RBDs *up*, and a minority of the proteins with all three RBDs in the *up* conformation (Kirchdoerfer et al., 2018).

The binding interface between SARS-CoV-2 RBD and ACE2 has a major contribution from polar interactions (Wang et al., 2020a; Yan et al., 2020), and the more extensive interactions between ACE2 and SARS-CoV-2 RBD as compared with SARS-CoV RBD might contribute to 4-fold higher binding affinity measured for the former complex (Wang et al., 2020a). Binding of the RBD to the receptor, and the RBD *up* conformation, may facilitate transition to the post-fusion conformation of protein S, for which weak interactions between protomers of protein S at the S2 domain, mediated largely by polar groups, might be important (Li et al., 2019). For example, downstream the sequence of the RBD, the D614G mutation became prevalent during the corona pandemic (Korber et al., 2020), and lack of inter-protomer H-bonding between D614 and T859, as observed in the cryo-EM structure of the pre-fusion conformation (Wrapp et al., 2020), was suggested could impact local flexibility and glycosylation of N616 (Korber et al., 2020). Since the D614 site is non-critical for antibody binding, an impact of the D614G mutation on vaccine development was suggested to be unlikely (Grubaugh et al., 2020).

Studies with simpler protein models suggest dynamic H-bonds and H-bond clusters shape protein conformational plasticity. Individual inter-helical H-bonds bring modest contributions to protein stability, on the order of ~0.6 kcal/mol on the average (Joh et al., 2008), though more significant contributions of up to 3.5–5.6 kcal/mol, were measured for salt-bridges (Hong et al., 2006). When located close to each other in a cluster, H-bonding protein sidechains can engage in clusters of dynamic H-bonds that stabilize protein conformations and shift populations during the reaction cycle of the protein (Bondar and White, 2012).

A protein complex as large as the spike protein S (Fig. 1), and for which experimental information about functional roles of specific amino acid residues is somewhat scarce, brings about the challenge of how to identify intra-molecular interactions that could shape conformational dynamics. To tackle this challenge, here we compute H-bond interaction networks, and then rank H-bonding groups according to centrality measures. The usefulness of using centrality measures for interaction networks was illustrated by earlier works indicating high centrality values for amino acid residues that are conserved and/or are located at enzyme active sites (Amitai et al., 2004) and a negative correlation between centrality and evolutionary rate (Fokas et al., 2016); more recently, centrality measures were used to identify functional interactions in a membrane transporter (Harris et al., 2020).

To derive clues about intra-molecular interactions with potential role in shaping structural dynamics of protein S, we computed two-dimensional graphs of all H-bonds of protein S in structures proposed for the closed, open, and pre-fusion conformation, and for ACE2 bound to an RBD fragment. We then used centrality measures to identify amino acid residues with central roles for the connectivity within local H-bond clusters. As some of the H-bond clusters we identified are large, we refined centrality computations by considering unique paths between all H-bonding groups of each cluster.

We find that the closed conformation of the ectodomain of protein S hosts an extensive, central cluster of H-bonds, contributed symmetrically by the three protomers. Relatively close to the ACE2 binding interface, each protomer of the closed conformation has H-bond clusters contributed by the same groups. In the open and, even more in the pre-fusion conformation, this symmetry of H-bond clusters is largely perturbed. In structures of the ACE2-RBD complex, four clusters of H-bonds mediate the binding interface. Taken together, the analyses presented here suggest the reaction coordinate of protein S includes rearrangements of extensive H-bond clusters, such that a three-fold compositional symmetry that characterizes the closed conformation of protein S is lost in the open and pre-fusion conformations. A H-bond network that extends deeply across the interface between protein S and ACE2 potentially contributes to the strong binding affinity.

2. Methods

2.1. Datasets of spike-like proteins used for bioinformatics analyses

We prepared two sets of sequences of spike protein S. *Set-A* consists of protein sequences from various organisms, and thus can show relatively large variations in the amino acid sequence; *Set-B* contains sequences of SARS-CoV-2 isolated from human hosts. We extracted and curated the sequences as summarized below.

Protein S sequences for *Set-A* were extracted by performing blastp and blastx (Altschul et al., 1997) database searches against the SARS-CoV-2 Database hosted at the NCBI (National Center for Biotechnology Information, accessed March 26, 2020). For *Set-B* sequences we used the Virus Variation Resource, VVR (Hatcher et al., 2017) to extract proteins available for SARS-CoV-2 genomes, selected manually sequences of S proteins according to the database annotation, and removed partial hits.

For the both *Set-A* and *Set-B*, redundant sequences were removed automatically using a threshold of 100% similarity such that, when two sequences were identical, only one was kept. The resulting datasets included 48 (*Set-A*) and 14 (*Set-B*) protein sequences.

We aligned sequences of S proteins from *Set-A* and *Set-B* separately using MAFFT (Katoh and Toh, 2008; Katoh and Standley, 2013). Each alignment was manually inspected and curated, and figures of sequence alignments were prepared using Easy Sequencing in PostScript, ESPript 3.x (Robert and Gouet, 2014). These software packages were used for all sequence alignments described below. Likewise, we inspected and hand curated all sequence alignments.

2.2. Sequence region corresponding to SARS-CoV-2 RBD

To inspect the sequence variation in the region corresponding to the RBD of SARS-CoV-2 protein S we started with the sequence alignment for *Set-A* generated as described above, and extracted the sequences corresponding to SARS-CoV-2 RBD groups R319 to F541 (Yan et al., 2020). Sequence regions corresponding to SARS-CoV-2 RBD were then combined into a single multifasta file; identical sequences were removed. The resulting sequences were realigned with MAFFT using SARS-CoV-2 as a reference.

2.3. Dataset of ACE2 protein sequences

We prepared two sets of ACE2 sequences. *Set-C* consists of orthologue protein sequences from various organisms, which we extracted by performing blastp (Altschul et al., 1997) database searches against the NCBI non-redundant protein database (accessed March 29, 2020). *Set-D* contains human sequences of ACE2 from the 1000 human Genome project (Auton and Brooks, 2015); for this set, we used the Ensembl project (Hunt et al., 2018) and extracted the different protein haplotypes existing in the 1000 Genome Project for ACE2 using the GRCh38 human genome assembly as reference. For both sets, we removed redundant sequences according to a threshold of 100% similarity such that, when two sequences were identical, only one was kept. The resulting *Set-C* and *Set-D* datasets included 46 and 22 ACE2 sequences, respectively.

2.4. Analyses of the length of S protein sequences and amino acid residue composition

The length of a protein sequence is given by the total number of amino acid residues. To characterize the amino acid composition of S proteins we used the program pepstats from the Open Source suite EMBOSS (Rice et al., 2000) to calculate, for each sequence in *Set-A*, the total number of charged and polar groups grouped into *i)* Asp and Glu, which are negatively charged at standard protonation; *ii)* positively charged Arg and Lys; *iii)* His groups, which can be neutral or protonated; *iv)* polar groups Asn, Gln, Ser, Thr, Trp, and Tyr.

2.5. Motif searches for protein S

We used *Set-A* of protein S sequences to identify motifs that include Asp, Glu, and His sidechains, *i.e.*, motifs with groups that could change protonation depending on pH. To identify motifs of interest we first inspected the distribution of Asp, Glu, and His in the sequence of SARS-CoV-2 protein S. Based on this, we chose for analysis motifs that consist of *i)* HE and HD; *ii)* D[HDE], which includes DH, DD and DE; *iii)* E[HDE]; *iv)* D(3,4), where (3,4) indicates that we searched for DDD and DDDD; *v)* [DE][GAVLIPFMW][DE]; *vi)* [DE](2,3)[GAVLIPFMW][DE]; *vii)* [DE][ST][DE]. Motif searches were performed using fuzzpro, a program from the Open Source suite EMBOSS (Rice et al., 2000) and analyzed using own scripts.

2.6. Structures of the SARS-CoV-2 spike protein S ectodomain in the closed, open, and pre-fusion conformations

Spike protein S is a homotrimer of three polypeptide chains, or protomers, which for simplicity we label here as *A*, *B*, and *C*. For the starting coordinates of protein S ectodomain in the open- and closed-conformations we used, respectively, the cryo-EM structures PDB ID:6VYB (3.2 Å resolution) and PDB:ID 6VXX (2.8 Å resolution) (Walls et al., 2020). For the pre-fusion conformation we used the cryo-EM structure PDB ID:6VSB (3.5 Å resolution) (Wrapp et al., 2020).

Table 1
Spike protein structures of SARS-Co-V-2 used to compute H-bond graphs and centrality values.

Protein Conformation[a)]	Length/chain[b)]	PDB	Resolution	Reference
pre-fusion	1120	6VSB	3.5 Å	(Wrapp et al., 2020)
open	1121	6VYB	3.2 Å	(Walls et al., 2020)
closed	1121	6VXX	2.8 Å	(Walls et al., 2020)

[a)] Conformation of the protein as proposed in the publication referenced here. [b)] Number of amino acid residues in each chain after constructing coordinates for missing internal groups. Relative to the structure of the pre-fusion conformation, the cryo-EM structures of protein S in open and closed conformations report coordinates for one additional amino acid residue, S1147.

Disulfide bridges were included as set in the Protein Data Bank (Berman et al., 2000) entry. In the closed conformation, each protomer (chain) has 12 disulfide bridges. In the open conformation, chains A and C have 12 disulfide bridges each, and chain B has 11 disulfide bridges. In the pre-fusion conformation, chains A and B have 12 disulfide bridges each, and chain C has 11. The number of amino acid residues for each of the structures used for analyses is reported in Table 1.

All structures were prepared by considering standard protonation for titratable amino acid residues, *i.e.*, Asp/Glu are negatively charged, Arg/Lys are positively charged, and His groups are singly protonated on the Nε atom. For simplicity, and since our analyses focus on internal H-bond networks of the spike protein, sugar moieties were not included.

The three cryo-EM structures used for analyses (see below) lack coordinates for water molecules. Coordinates for missing internal amino acid residues and H atoms of the proteins were generated using CHARMM-GUI (Jo and Kim, 2008; Lee et al., 2016) and CHARMM (Brooks et al., 1983). Molecular graphics illustrating the protein segments for which we constructed coordinates are presented in Fig. S3.

2.7. Electrostatic potential surface

Computations of the electrostatic potential surface were performed with the Adaptive Poisson Boltzmann Solver, APBS (Baker et al., 2001), in PyMol 2.0 (Schrödinger, 2015). We used the PDB2PQR web interface (Dolinsky et al., 2004) to assign partial atomic charges and atom radii according to the CHARMM force field (MacKerell et al., 1998). Electrostatic potential computations were restricted to protein atoms.

2.8. Dataset of structures of ACE2 bound to protein S fragments

We analyzed H-bond networks in three structures of ACE2-protein S complexes as summarized in Tables 2 and S2. As computations of average H-bond graphs require the same number of amino acid residues in the graphs to be averaged, where needed we used Modeller 9.21 (Marti-Renom et al., 2000) to construct coordinates for missing amino acid residues. Lists of amino acid residues whose coordinates were constructed for each structure are presented in Table S1. For all computations of graphs of H-bonds for ACE2-protein S complexes we used regions S19 to D615 of ACE2, and T333 to P527 for protein S.

2.9. Criteria for H-bonding

To identify H-bonds, we used the same geometry-based criteria as in our previous work (Siemers et al., 2019), whereby we consider that two groups are H-bonded when the distance between heavy atoms of the H-bonding groups is ≤3.5 Å, and the angle between the acceptor heavy atom, the H atom, and the heavy donor atom is ≤60°. For all structures included here in analyses we computed H-bonds between protein sidechains, and between protein sidechains and backbone groups. Interactions between charged groups (Asp, Glu, Arg, Lys) and between polar sidechains are treated with the same H-bonding criteria.

2.10. Tests for H-bond criteria

Analyses of H-bonding based on experimental electron densities used a set of H-heavy atom distances ranging from 1.56 Å to 2.59 Å (Espinosa et al., 1999), which is compatible with the 3.5 Å heavy-atom to heavy-atom distance and 60° H-bond angle criteria we used.

To illustrate how the distance criteria may influence details of H-bond cluster, we recalculated interactions of three selected H-bond clusters by considering only heavy atoms, and used as H-bond criterion a distance 2.8 Å, 3.0 Å, 3.5 Å, or 4.0 Å (Fig. S4). As anticipated, the stricter the distance criterion used, the fewer the H-bonds in a cluster (Fig. S4). At a heavy-atom distance criterion of 3.5 Å, H-bond clusters are slightly larger than clusters we reported with both distance and angle criterion (Fig. S4).

We further used 10 coordinate snapshots from a simulation on the soluble protein SecA (Karathanou and Bondar, 2019) to calculate the total number of protein H-bonds with *i)* the combined criteria of ≤3.5 Å distance between heavy atoms and ≤60° H-bond angle; *ii)* the criterion of a ≤2.5 Å distance between the H atom and the heavy atom. We obtained for SecA a total of 220–236 H-bonds in test *i)*, and 223–240 H-bonds in test *ii)*, suggesting the combined 60° angle and 3.5 Å heavy-atom distance criterion is largely equivalent to using a distance of 2.5 Å between the heavy atom and the H atom. In two additional tests on SecA simulations that start from two different protein conformations (Karathanou and Bondar, 2019), we used the combined 3.5 Å distance and 60°-angle criterion to compute the frequency for all H-bonds present in starting crystal structures. We found that about 90–95% of H-bonds that appear as weak in a protein crystal structure, with the distance between the heavy atoms from 3.0 Å to 3.5 Å, are nevertheless sampled over the course of the two prolonged simulations, and almost 60–75% of the weak H-bonds are sampled during at least 10% of the two simulations.

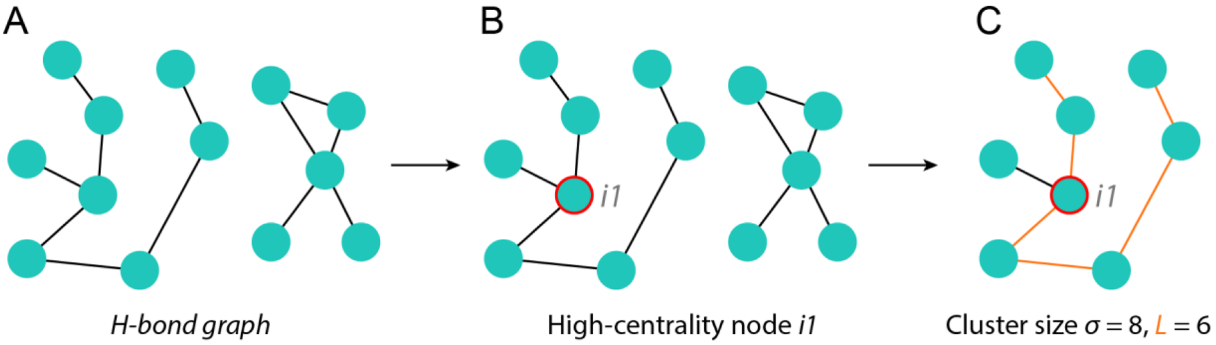
Pursuant to the considerations above, we suggest the combined ≤3.5 Å distance and ≤60° angle H-bond criterion we used here is reasonable.

2.11. Graphs of H-bond networks: nodes, edges, H-bond paths, and shortest path length

A graph, or network, of H-bonds, consists of a collection of *vertices* (also denoted as *nodes* or *points*) and *edges* that inter-connect the vertices. In graph theory, each vertex can be assigned a *degree* that gives the number of edges incident on that vertex, such that in an undirected graph a vertex of degree 1 (usually denoted as a *leaf*) has only one edge (Bender and Williamson, 2010; Sadavare and Kulkarni, 2012; West, 1996). The definition of leafs is related to that of a *tree*, which is a connected, acyclic graph whose edges are called *branches*, and whose nodes can be internal or leafs (Bender and Williamson, 2010). In a tree graph, a node is denoted as a leaf when it has degree of 1 or in the case of a rooted tree, when it has no child nodes (Bender and Williamson, 2010). The definition of leaf nodes can be used in the context of more general graphs that are not trees. Two nodes are denoted as *adjacent* if they directly connect to a common edge; two edges are denoted as *adjacent edges* if they share a node (Bender and Williamson, 2010).

Here we report graphs of H-bond interactions. These graphs have as nodes amino acid residues or heavy atoms that H-bond. The edges of the graph represent H-bonds, i.e., non-covalent interactions (Scheme 1).

An *H-bond path* is defined as a continuous chain of H-bonds that



Scheme 1. H-bond paths and H-bond clusters. H-bond paths are identified with the Connected Component tool of Bridge (Siemers et al., 2019). (A) Schematic representation of a H-bond graph. Cyan dots indicate amino acid residues that H-bond, and thin lines, H-bonding. (B) We compute and report the centrality of each node. Here, node i_1 has high *BC* value relative to all nodes of the entire graph. (C) The H-bond cluster of node i_1 is extracted for analysis. A cluster size $\sigma = 8$ means this cluster contains 8 nodes. The shortest path highlighted yellow passes via node i_1 and it includes 6 H-bonds, thus its path length is $L = 6$.

connects two nodes (amino acid residues) of the graph. The *shortest path* between two nodes is the path that connects these two nodes via the least number of intermediate nodes (Cormen et al., 2009). The path length *L* of a shortest-distance path is given by the number of H bonds that are inter-connected in that path (Scheme 1).

2.12. Connected Components graph searches, root nodes, and H-bond clusters

Bridge (Siemers et al., 2019) uses *Connected Component* searches (Cormen et al., 2009) to compute subgraphs of H-bonds in which all nodes (H-bonding groups) inter-connect with each other. Such a Conected Component search starts from a specific node of the graph, denoted as a *root node*, and identifies all H-bond paths starting from that node. The result of the Connected Component search is thus a subnetwork, or sub-graph, of H-bonds, which we denote as *H-bond cluster*.

2.13. Centrality measures and a new measure of the unique shortest paths

To assess connectivity within the network and rank the relative importance of nodes in a graph of H bonds, we used centrality measures as defined below and illustrated in Scheme 1.

The *Degree Centrality* (*DC*) of a node (or vertex) n_i gives the number of edges (H-bonds) of the node (Freeman, 1979) (Scheme 1). The normalized *DC* value of node n_i is computed by dividing its *DC* by the maximum possible edges to n_i (which is *N*-1, where *N* is the number of nodes in the graph). In the case of H-bond graphs, the *DC* value of a protein group indicates the number of its unique direct H-bonds, thus providing information on the local H-bond environment of that protein group.

The *Betweenness Centrality* (*BC*) of a node n_i gives the number of shortest-distance paths between any two other nodes n_j and n_k that pass via node n_i divided by the total number of shortest paths that connect n_j and n_k irrespective of whether they pass via node n_i (Brandes, 2001; Freeman, 1977, 1979) (Scheme 1). The normalized *BC* value of node n_i is computed by dividing its *BC* by the number of pairs of nodes not including n_i. *BC* values of protein groups that are part of an interaction network can thus be used to assess the topology of the network (Fokas et al., 2016).

Some of the H-bond clusters we identified for the spike protein are rather large, with up to 33 amino acid residues. For such large clusters, the standard definition of the *BC* as introduced above can lead to large *BC* values difficult to interpret intuitively. To tackle this issue we introduce here a new, more restrictive definition, which we denote as Unique Shortest Paths, *USP*. By contrast to *BC*, which accounts for all shortest paths between all pairwise nodes *j, k* that pass through node *i*, *USP* accounts only for the *longest unique shortest paths*. A unique shortest path is not included in any other shortest path (Scheme 2). The *USP* calculation is performed for all nodes n_i (i.e., for all H-bond groups).

2.14. Cluster size and histograms of H-bond path lengths

Our searches of H-bond clusters indicated protein S contains several H-bond clusters that are rather large in, e.g., the closed state of the protein, whereas in the open and/or pre-fusion conformation only some of the amino acid residues remain inter-connected in a common cluster of H-bonds. As a coarse measure of the changes in the number of H-bonding groups of a cluster, we use here *i)* the size of a cluster *σ*; *ii)* histograms of the length of shortest-distance paths of selected H-bond networks.

The cluster size *σ* is given by the total number of nodes (H-bonding amino acid residues) of that cluster (Scheme 1C). A value of the cluster size $\sigma = 2$ is thus equivalent with a H-bond.

Histograms of the distribution of path lengths in a H-bond cluster give an overview of the H-bond connections within that cluster. We compute all shortest paths between all pairs of H-bonding groups (nodes) in all clusters of H-bonds. For each path, we store its length. We then select high-*BC* groups (nodes) n_i of particular interest, and the shortest-distance H-bond paths that pass through that node n_i. We calculate from these selected H-bond paths histograms that we use to illustrate the size of the network of node n_i.

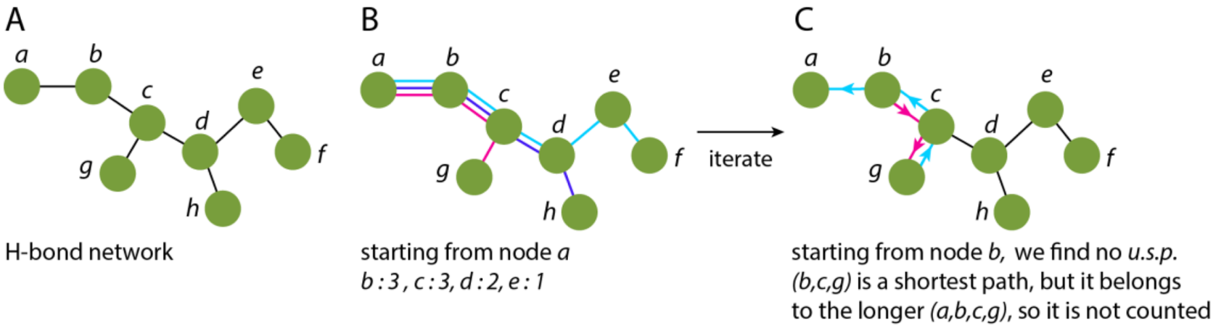
2.15. Detection of anchors of H-bond clusters from directed graphs, and cluster density

Changes in nodes that delineate the periphery of a large H-bond cluster indicate rearrangements in that cluster. We denote these peripheral nodes as *anchors* or leafs (Scheme 3).

For a given H-bond cluster identified from a Connected Components search for a root node n_i, we compute all shortest paths from node n_i to all other nodes in the H-bond cluster (Scheme 3A), construct a directed graph using all shortest paths whose edges point away from root node n_i, and query this directed graph to identify all nodes that have only inward edges (Scheme 3B). Nodes N_A marked as anchor nodes of the cluster are illustrated in Scheme 3C. The *node density of a cluster*, ρ, is given by the number of anchor nodes divided by the total number of nodes of the graph. AP_L, the distance between two anchors of the cluster, is given by the number of H-bonds that constitute the path that connects two anchor points passing via N_c (Scheme 3C).

2.16. Conserved H-bond networks in ACE2-RBD complexes

We consider that nodes and edges of a graph are conserved when present in all chains of the four ACE2-protein S structures we analyzed


A
a
b
c
d
e
f
g
h
H-bond network
B
a
b
c
d
e
f
g
h
starting from node a
b:3, c:3, d:2, e:1
iterate
C
a
b
c
d
e
f
g
h
starting from node b, we find no u.s.p.
(b,c,g) is a shortest path, but it belongs
to the longer (a,b,c,g), so it is not counted

Scheme 2. Unique Shortest Path computations, *USP*. (A) H-bond network with amino acid residues represented as nodes *a-f*. (B) There are 3 unique shortest paths starting from node *a*: *a-b-c-d-e-f*, *a-b-c-d-h*, and *a-b-c-g*. (C) Path *b-c-g* belongs to the longer path *a-b-c-g*, and thus it is discarded, path *a-b-c-g* being reported as *USP* that includes nodes *b* and *g*.

(Table 2). Thus, we excluded from the computations of conserved graphs 21 amino acid residues that are different between the chimera RBD from PBD ID:6VW1 (Shang et al., 2020), and the wild-type RBD from the other three structures we analyzed (Table 2). For conserved networks of H-bonds we report average centrality values computed for all chains of the structures used for analyses.

2.17. H-bond clusters and cluster labels for the ectodomain of protein S

Amino acid residues found to have high centrality values were used as root nodes in Bridge (Siemers et al., 2019) for Connected Components searches to identify all H-bond paths that connect to these groups. To facilitate comparison of H-bond clusters computed for different protein conformations, labels of H-bond clusters give the protein conformation, the protomer, and the amino acid residue with high centrality value in the cluster. Thus, label 'PA_D663′ indicates a cluster identified by searching for all H-bonds of the high-centrality group D663 in protomer A of the pre-fusion conformation. The open and closed conformations of protein S are indicated by the letters 'O' and 'C', respectively.

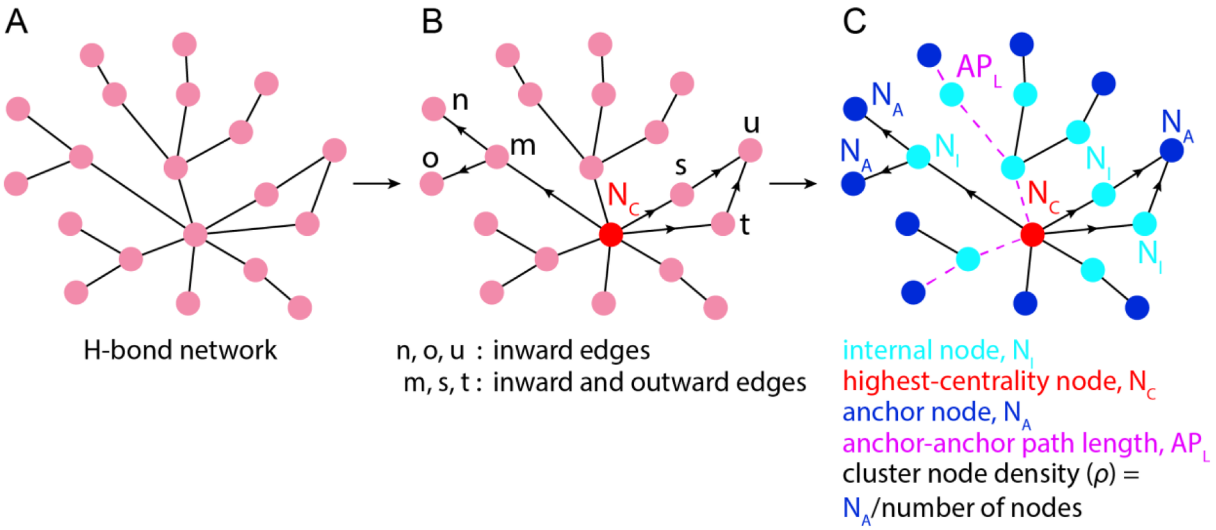
Table 2
Structures of SARS-CoV-2 S protein fragments bound to ACE2. We report the resolution (Res.) in Å.

PDB ID	Method	Res.	Length/chain		Reference
			ACE2	Spike	
6M0J	X-ray	2.45	603	229	(Lan et al., 2020)
6VW1[a)]		2.68	597	217	(Shang et al., 2020)
6LZG		2.5	597	229	(Wang et al., 2020a)
6 M17	Cryo-EM	2.90[b)]	814	223	(Yan et al., 2020)

[a)] The protein S fragment is a chimera of the SARS-CoV-2 RBM, except for a RBM loop, and SARS-CoV for the remaining of the amino acid residues (Shang et al., 2020). [b)] The local resolution at the interface between ACE2 and the RBD is 3.5 Å (Yan et al., 2020).

2.18. Labels of H-bond clusters at the interface between ACE2 and the RBD of protein S

For each of the four ACE2-RBD structures (Table 2) we first identified all connected components of the H-bond network using the Networkx package (Hagberg et al., 2008). In the second step, we selected those components that include direct H-bonds between amino acid



Scheme 3. Anchor points and H-bond cluster density. (A) H-bond cluster with nodes (graph vertices) shown as pink dots, and H-bonds (graph edges) as black lines. (B) Node N_c with the highest centrality value from the cluster is identified and used as starting point for a directed search to identify nodes *n*, *o* that are connected only with inward edges, and nodes *m* that have both inward and outward edges. Similarly, nodes *s*, *t* are connected with both inward and outward edges, while node *u* is connected only with inward edges. Paths *n-m-*N_c*-s* and *n-m-*N_c*-t* are examples of longest shortest paths; these paths include paths *m-*N_c*-s* and *m-*N_c*-t*. (C) Nodes that connect only to inward nodes, *i.e.*, leafs of the graph, are denoted here as anchors of the cluster.

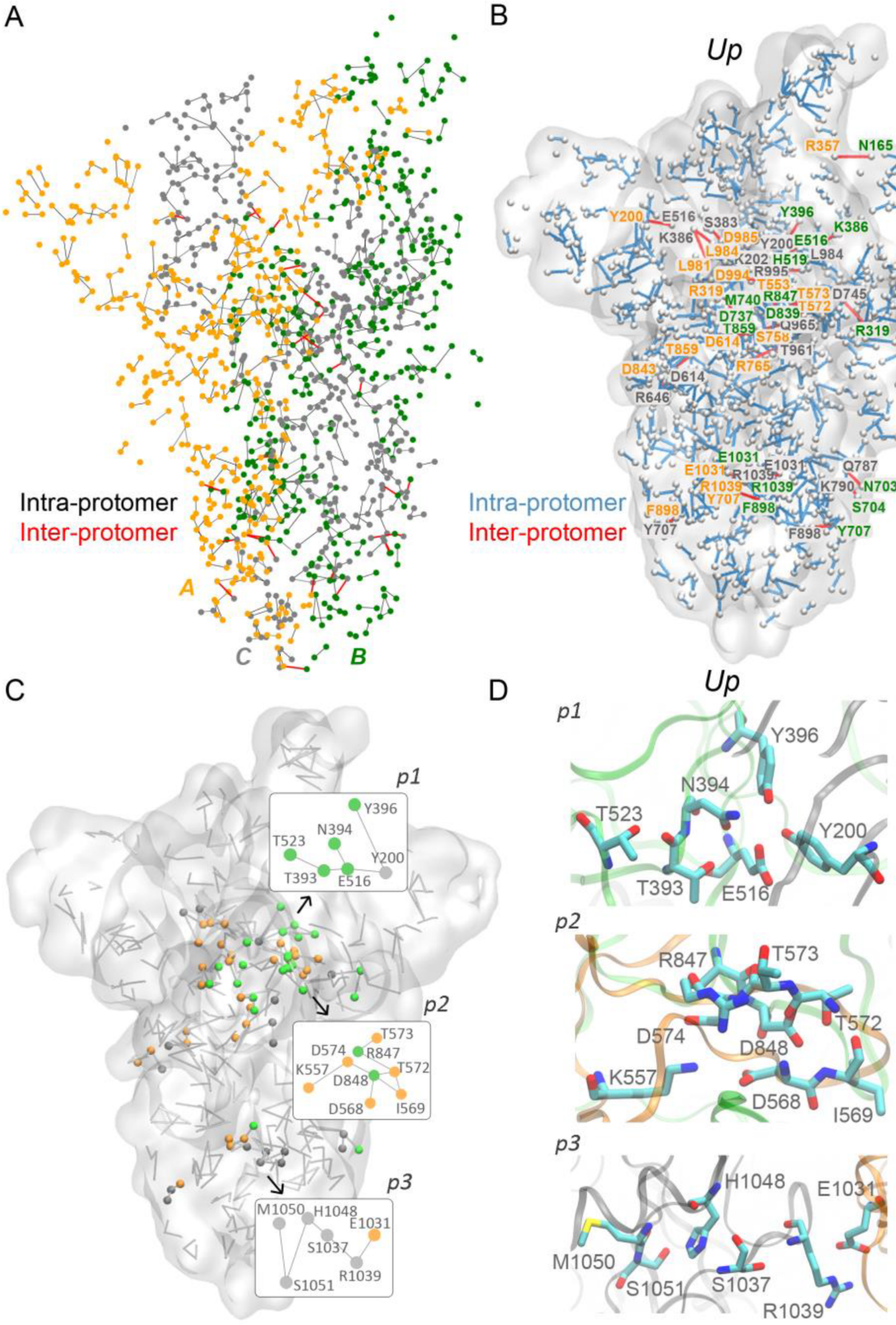


Fig. 2. H-bond network of the ectodomain of SARS-Co-V-2 protein S in pre-fusion conformation. (A) Complete H-bond graph of protein S. Filled circles representing H-bonding amino acid residues are colored red, green or dark gray, according to the protein chain. Gray and red lines indicate intra- and inter-protomer H-bonds, respectively. (B) H-bond clusters that inter-connect protomers of protein S. The protein is represented as a surface, and Cα atoms of H-bonding groups are shown as small white spheres. Labels for amino-acid residues involved in inter-domain H-bonds are colored according to the protomer. (C) Schematic representation of H-bond clusters that involve inter- and intra-protomer H bonding with $\sigma \geq 6$. (D) Illustration of interactions in H-bond clusters with $\sigma \geq 6$. H-bonds for the open and closed conformations are presented in Figure S5.

residues of ACE2 and of the RBD. This led, for each structure, to 3–4 H-bond clusters that we labeled as *a*, *b*, *c*, and *d*. The assignment of specific amino acid residues to clusters *a* - *d* is the same for the four structures.

3. Results and discussion

We implemented a protocol that relies on computations of H-bond graphs and centrality measures to characterize H-bonding in large proteins and protein complexes. With this protocol, we computed graphs of H-bonds for the closed, open, and pre-fusion conformations of the ectodomain of the SARS-CoV-2 protein S, and for structures of the RBD of protein S bound to ACE2, and identified H-bond clusters that characterize conformations of protein S, and H-bond clusters that mediate the binding interface with ACE2. From analyses of sequences of spike proteins we identified conserved motifs that could enable pH sensitivity.

3.1. The ectodomain of protein S has an extensive network of H-bonds

Each of the protein chains, or protomers, of a protein S trimer, has numerous H-bonds: there are, for each protomer, between 238 and 286 H-bonds between groups of the same protomer (Table S2), and 22–47 H-bonds between groups between protomer pairs (Fig. 3A, 3B, S5, Table S2). In total, there are ~798–902 H-bonds for each conformation of protein S (Fig. 2A, Table S2). These H-bonds have a non-uniform distribution across protein S: In the pre-fusion conformation there are two regions with extensive inter-domain H-bonds: one central region in which most of the inter-domain H-bonding belongs to two clusters (clusters p_1 and p_2 in Fig. 2C, 2D), and a region close to the stalk part of the protein (cluster p_3, Fig. 2C, 2D); each of these three clusters includes at least one carboxylate group. A qualitatively similar picture of the distribution of inter-protomer H-bonds is also observed for the open and closed conformations of protein S (Fig. S5).

3.2. H-bonding groups with high DC values and extensive local H-bond connectivity

The large size of protein S makes it challenging to characterize its structural dynamics, and we wondered whether H-bond graphs could aid by identifying sites where interactions change. Our strategy was to first use centrality measures to rank H-bonding groups according to their involvement in H-bond connections, and then inspect interactions of groups with high centrality values.

We found that most of the H-bonding amino acid residues of protein S have rather small centrality values (Fig. 3C, 3D), suggesting most of these groups lack extensive local connections or participation in large numbers of H-bond paths. Importantly, in all three conformations of protein S we analyzed, *BC* values ≥ 15 belong to groups with coordinates solved experimentally (Table S3, Fig. S6). To avoid uncertainties about modeling of missing protein group, in the discussion below we focus on clusters that are centered at groups with *BC* ≥ 15.

The majority of the H-bonding sidechains have $DC < 2$, i.e., most of the H-bonding groups have at most one H-bond, and only a few groups have 3–5 H-bonds (Fig. 3C, 3D). For example, in pre-fusion conformation protomer *A* (Fig. 3B,E,H) N437 ($DC = 5$) has three H-bonds at the carboxyamide sidechain group, and two H-bonds at its backbone carbonyl group (Fig. S7); near N437, the sidechain of R509 ($DC = 4$) has 4 H-bonds with nearby protein sidechain and backbone groups (Fig. S7). Both N437 and R509 have large *BC* values (Fig. 3A,H), indicating they are part of H-bond clusters. In the closed conformation R509 has high *BC*, but N437 has only one H-bond, with N439, thus $DC = 1$ (Fig. S7 A-C). This suggests structural rearrangements in the local environment of N437, such that its number of H-bonds changed.

Local interactions of high-*DC* groups might be energetically costly to break, suggesting H-bond clusters that include high-*DC* groups could have somewhat reduced local dynamics. This suggestion is compatible with previous analyses of static crystal structures suggesting high flexibility for groups with low *DC* (Fokas et al., 2016). But analyses of static protein coordinate snapshots, as reported here, lack direct information about motions of the protein along its reaction path and about time scales for transitions between intermediate states of protein S. We thus use centrality values to identify regions of the protein in which H-bonding groups are part of extended H-bond clusters in static coordinate snapshots, and suggest these clusters could be used as structural signatures to characterize of metastable conformations of protein S.

In a static protein structure, the *DC* value of an amino acid residue is limited by the number of H-bonds permissible for its sidechain and backbone. A high-*DC* group may also have high *BC,* indicative of a central role in a H-bond cluster –examples here are N437 and R509, which have the highest *DC* and *BC* values in their H-bond cluster in pre-fusion (Fig. 3A,B,E,H, S8). A group with relatively small, but nonzero, *DC* value may, however, also have small *BC* (Fig. S8). In what follows, for H-bond clusters of interest we use centrality values and close inspection of the H-bond connectivity to identify sites with large H-bond networks, and sites where local structure rearranges when protein S changes conformation.

3.3. Using cluster size and centrality values to identify rearrangements within H-bond clusters

The size and shape of a H-bond cluster, i.e., how many H-bonding groups are in a cluster and how these groups locate relative to each other, impact centrality values and H-bond paths of specific H-bonding groups of that cluster. Thus, H-bonding groups in the center of the H-bond cluster would have larger centrality values than groups at the periphery (see node N_c in Scheme 3B). When a particular H-bonding group has markedly different *BC* values in two different protein conformations, the H-bond cluster that contains this group should have different size and/or different shape.

To evaluate the relationship between *BC* value, cluster size, and cluster shape, we analyzed H-bond clusters that contain high-*BC* groups (Figs. 3, 4, S9), the distribution of *BC* values in clusters with $\sigma \geq 2$ (Figs. 2, 4, 5, S10), and the distribution of *USP* values as a function of the cluster size (Fig. S13).

Highest *BC* values tend to be obtained for H-bonding groups that are part of large clusters (Figs. S10, S11, S12). In a cluster with many H-bonding groups (large σ) a group located centrally can be part of many shortest-distance H-bond paths, and hence have large *BC*. Low *BC* values, including $BC = 0$, can also be observed in clusters with relatively large σ (Figs. S10, S11); these are H-bonding groups located at the periphery of the cluster.

In the case of large clusters, computations of *BC* values by accounting for all shortest paths can lead to very large *BC* values, e.g., in the closed conformation R1039 is part of a cluster with $\sigma = 33$ and has $BC = 218$. The refined centrality computation we introduced here, *USP*, based on unique shortest paths, provides a qualitatively similar, but somewhat more intuitive picture of the connectivity within a cluster (Figs. 4, S13). The largest *UBC* value, of 30, is obtained for R1039 in the closed conformation (Fig. 4A, 4C). Except for the PA_R509 cluster of pre-fusion, where R509 has $UBC = 19$ (Fig. 4E), all other H-bond clusters have *UBC* values within 10, i.e., maximum 10 shortest paths can include one H-bond group (Fig. 4).

The size of H-bond clusters (Fig. 5) and *BC* values of specific H-bonding groups (Fig. 3) can vary significantly among the three conformations of protein S, indicating that conformational changes of protein S associate with structural rearrangements within clusters of H-bonds. We inspected closely H-bond clusters located at functionally important regions of protein S.

3.4. H-bond clusters near proteolytic cleavage sites

D663 is relatively close in the sequence to the R685 S1/S2 cleavage site (Fig. 1A). The pre-fusion protomer A, whose RBD is in the *up* conformation (Fig. 3A), has a relatively large H-bond cluster ($\sigma = 9$) centered at D663 ($BC = 20$, Fig. 5E, S14G); this H-bond cluster includes K310, E661, and Y695. In protomer B, the D663 cluster is smaller ($\sigma = 6$), but it still includes E661 and Y695 (Fig. S14H), whereas in protomer C, D663 H-bonds only to K310 (Fig. S14I). All protomers of

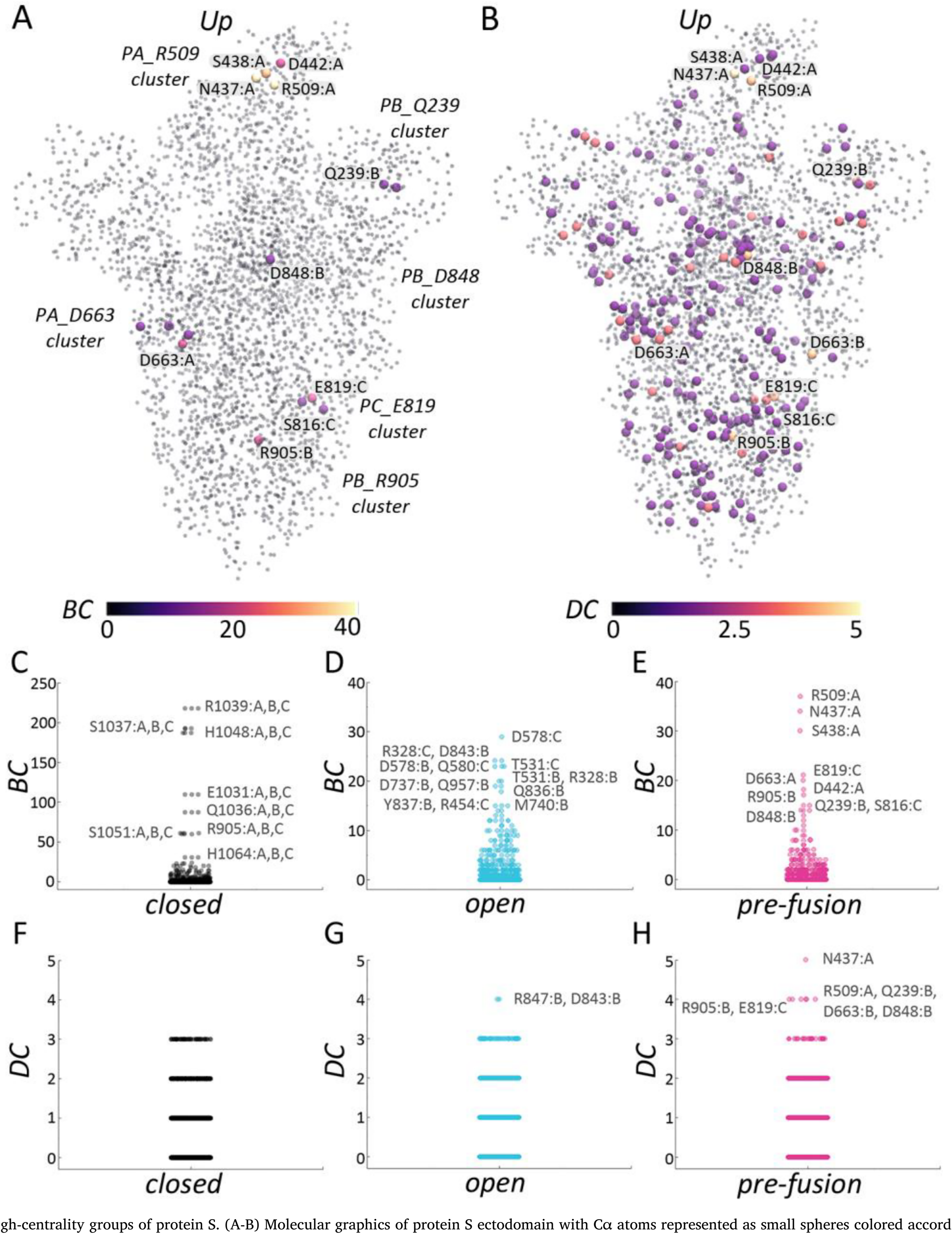


Fig. 3. High-centrality groups of protein S. (A-B) Molecular graphics of protein S ectodomain with Cα atoms represented as small spheres colored according to *BC* (panel A) vs. *DC* values (panel B). For selected high-centrality groups, we label H-bond clusters by the highest-centrality amino acid residue of that cluster, and by the protomer to which that amino acid residue belongs. Color bars indicate centrality values. Here and in Figs. 4, 5, 9, and S9, we use a linear color scheme in which colors range from black (low value) to yellow (high value). (C–H) Jitter plots of the distribution of *BC* (panels C-E) and *DC* values (panels F-H). Centrality values and jitter plots were computed and prepared with MATLAB R2017b (The MathWorks, 2017). *BC* and *DC* values for selected protein groups are presented in Table S3. A scatter plot of the *DC* vs. *BC* values of selected amino acid residues in pre-fusion is presented in Figure S8, and H-bond clusters of the open and closed conformations, in Figure S9. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

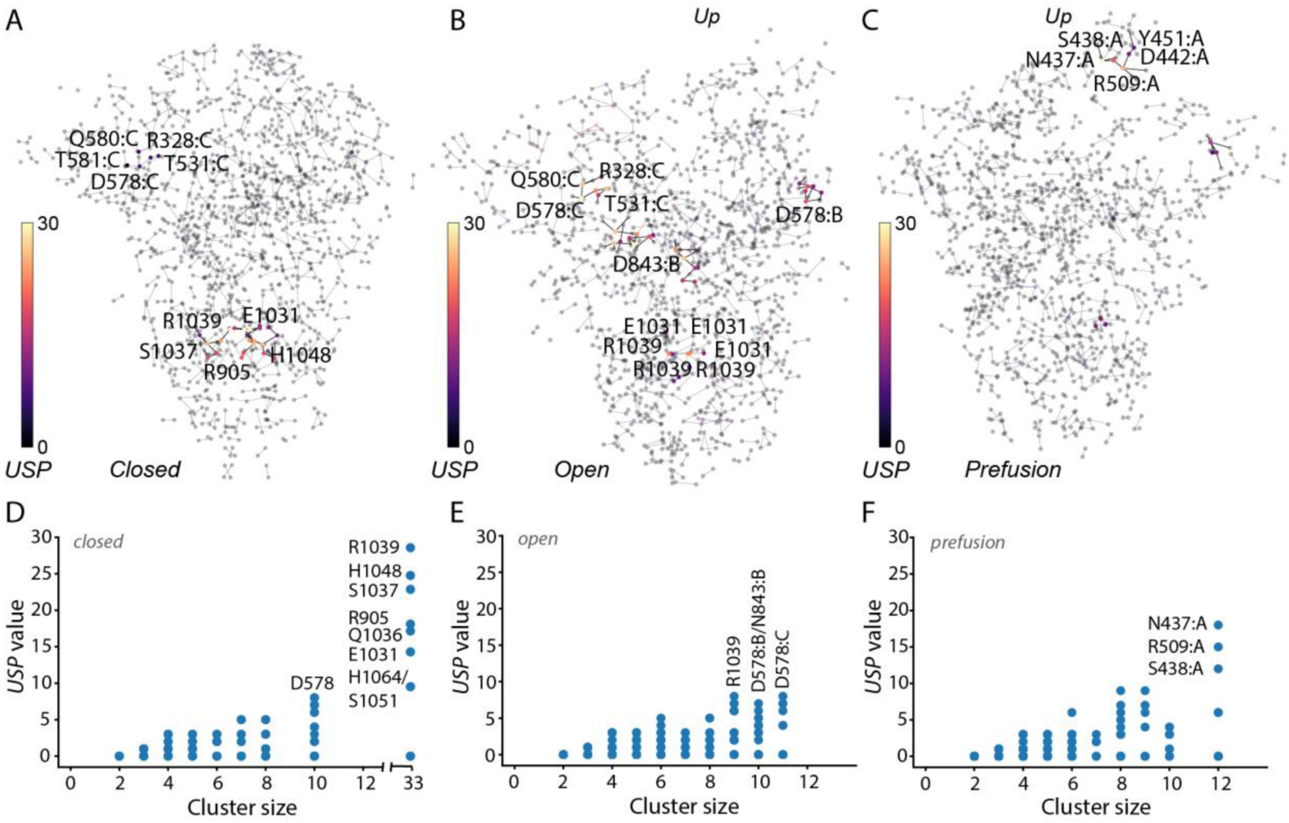


Fig. 4. High-centrality groups of protein S in closed vs. open conformations. Small filled circles indicate amino acid residues that participate in H-bond clusters, color-coded according to unique shortest path computations. (A, B) Unique shortest path values computed for the closed (panel A) and open (panel B) conformations of the ectodomain of protein S. (C-E) Unique shortest path and σ values for all H-bond clusters identified for the closed (panel C), open (panel D), and pre-fusion conformations (panel E) of the ectodomain of protein S. Additional labels of H-bond clusters in panels D-F are presented in Figure S13.

the closed conformation, and two protomers of the open conformation, have D663 H-bonded only to K310 (Fig. S14A-F). That is, the interaction between K310 and D663 appears to be a common structural motif, and the K310-D663 can become part of a more extensive H-bond cluster in the pre-fusion conformation of protein S.

E819 is close to the R815 S2′ proteolytic cleavage site (Fig. 1A). In pre-fusion, each protomer has a different arrangement of the H-bond cluster centered at E819: The RBD *up* protomer A has E819 H-bonded to S816 and part of a relatively small H-bond cluster with three other amino acid residues (Fig. S15 G). In protomer C, the S816-E819 H-bond is part of a large cluster with five protein groups, and protomer B has the S816-E819 H-bond without a cluster (Fig. S15 H, I). By contrast, each protomer of the closed and open conformations has the S816-E819 H-bond part of small clusters ($\sigma = 4$) arranged symmetrically (Fig. S15 A-F). Thus, H-bond clusters centered at E819 and D663 are similar in that they maintain a H-bond (with S816 and, respectively, K310) in all three conformations of protein S, and tend to have similar or the same compositions in the closed and open conformations; in pre-fusion, one of the protomers has significantly larger D663 or E819 clusters (Figs. S14, S15).

Taken together, the analyses above indicate that conformational transitions of protein S associate with changes in H-bonding near the S1/S2 and S2′ proteolytic cleavage sites. At these sites, compositional symmetry of H-bond clusters D663 and E819 is lost when the protein changes conformation from closed to pre-fusion.

3.5. The R509 H-bond cluster of the RBD rearranges drastically during conformational dynamics of protein S

In SARS-CoV mutating to Ala the R495 (corresponding to SARS-CoV-2 R509) decreases binding to ACE2 (Chakraborti et al., 2005) (Table S4). As summarized below, we find that R509 is part of three-fold symmetrical clusters in the closed conformation; smaller, largely symmetrical R509 H-bond clusters are present in two protomers of the open conformation; in pre-fusion, the symmetrical composition of the H-bond cluster is lost, with an extensive H-bond cluster observed only for the RBD up conformer (Figs. 5, S16).

Except for one protomer of pre-fusion, R509 is within H-bond distance of D442 (Figs. 6, S16 G&H); mutating to Ala D442 (which corresponds to D429 in SARS-CoV-2) abolishes ACE2 binding of SARS-CoV protein S (Chakraborti et al., 2005). In each protomer of the closed conformation, the D442-R509 H-bond is part of H-bond clusters constituted by the same groups including N448 and N450 (Figs. 6, S16 A-C). Adjacent in the sequence, Y449 H-bonds to ACE2 (Lan et al., 2020). In the open conformation, two protein conformers have the D442-R509 H-bond as part of relatively large and similar clusters (Fig. 5), whereas in the third protomer this H-bond is singular (Fig. S16E).

A similar observation of a three-fold symmetry present in the closed conformation, but not in open and pre-fusion, can be made for H-bonding at D578, which is relatively close to the RBD. Each protomer of the closed conformation has large, symmetrical D578 H-bond clusters whereby D578 H-bonds to R328, Q580, and T581 (Figs. 6, S17 A-C, J-L). This core H-bonding of D578 is preserved in the open conformation, but only two of the protomers maintain larger H-bond clusters centered at D578 (Figs. 6, S17 D-F). In pre-fusion, D578 is part of a relatively small H-bond cluster on one of the protomers (Fig. S17H), and has just singular H-bonds in the other two protomers (Figs. 6, S17I, S18).

As the two protomers with RBD *down* in pre-fusion lack the extended R509 cluster of the RBD *up* protomer, we inspected the H-bond clusters to find out whether the *down* protomers might have other clusters of H-bonds. We identified, in each of the RBD *down* protomers, a relatively large H-bond cluster of 6–8 protein groups that include

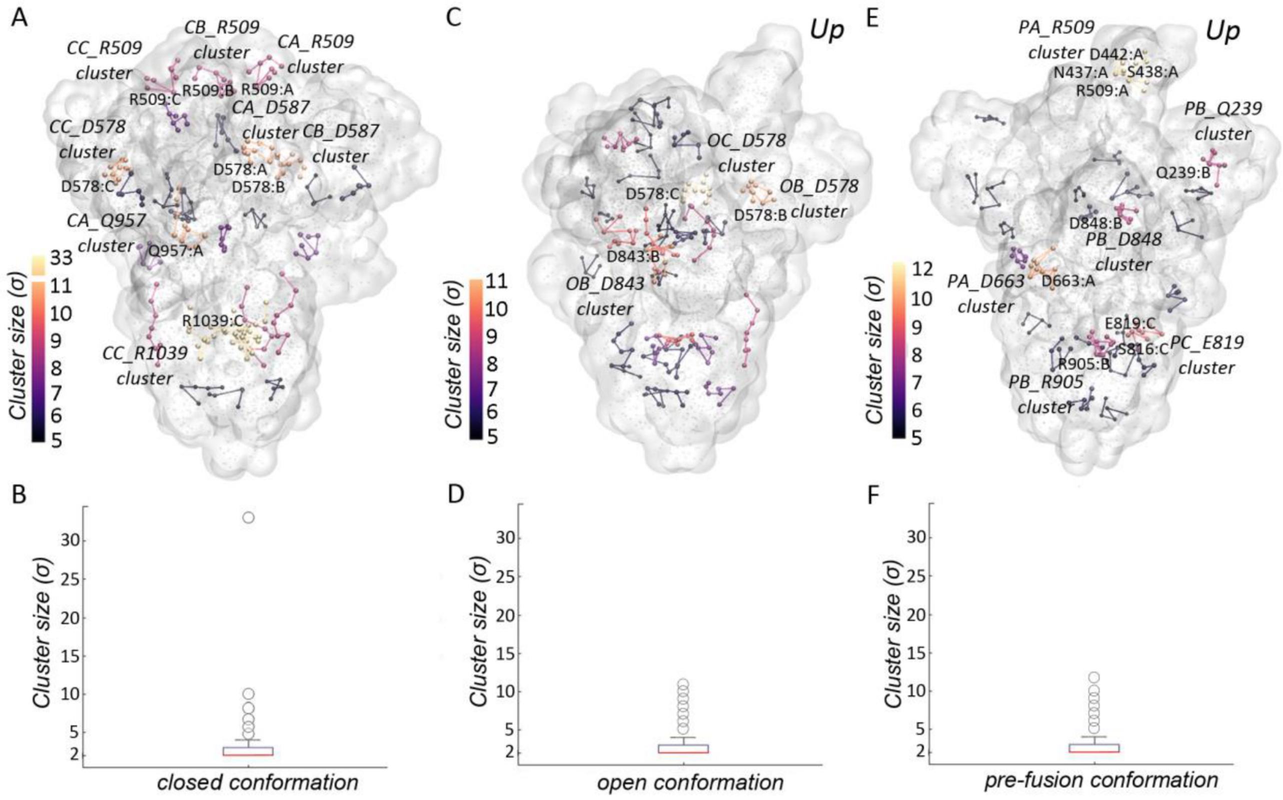


Fig. 5. H-bond cluster size in the closed, open, and pre-fusion conformations of the ectodomain of protein S. The protein is shown as transparent white surface. Cα atoms of amino acid residues involved in H-bond clusters are shown as small spheres if $\sigma < 5$, and color-coded according to cluster size if $\sigma \geq 5$. For clarity, we show explicitly only H-bond clusters with $\sigma \geq 5$. (A, B) Molecular graphics of the closed conformation (panel A) and boxplot of the distribution of all H-bond clusters computed for this conformation (panel B). (C, D) Molecular graphics of the open conformation (panel C) and corresponding distribution of all H-bond clusters (panel D). (E, F). Molecular graphics of the pre-fusion conformation (panel E) and corresponding distribution of all H-bond clusters (panel F). Cluster analysis was performed with MATLAB script, Network Components (Larremore, 2014), MATLAB Central File Exchange retrieved April 28, 2020.

Q239 (Fig. 5E, S12 F&G). In the RBD *up* protomer, Q239 is H-bonded to N81 without being part of a larger H-bond cluster (Fig. S12E); H-bonding between Q239 and N81 is present in protomers of the closed conformation, and open conformations. It thus appears that re-arrangements of H-bonding groups leading to sampling of the R509 cluster in the RBD *up* conformer associate with less extensive H-bonding at the Q239 site, where interactions remain restricted to the core N81-Q239 H-bond.

The loss of three-fold symmetry of the R509 and D578 H-bond clusters in pre-fusion conformation could be important to enable binding of protein S to ACE2, and/or for proteolytic activation of protein S. Structural asymmetry in protomers was also observed in the crystal structure of the trimeric membrane transporter AcrB, being thought that each protomer was captured in a different conformation sampled during the reaction cycle of the transporter (Seeger et al., 2006).

3.6. The central H-bond cluster of protein S

The most prominent H-bond cluster of the closed conformation is located close to the stalk of the ectodomain: The R1039 cluster includes 33 groups, 11 from each protomer (cluster CC_R1039 in Fig. 4A, 6). This is the largest H-bond cluster we identified for all three protein conformations (Figs. 3, 4). We denote the R1039 cluster of the closed conformation as *the central H-bond cluster* of protein S (Fig. 6, S19, S20).

The R1039 H-bond cluster has a broad distribution of path lengths in the close conformation, as compared to a narrow range of short path values in the open and pre-fusion conformations (Fig. 7). Protomers have largely similar cluster densities at R1039 in the closed and open conformations, but largely different values in pre-fusion (Figs. 8, S21), suggesting that in pre-fusion the R1039 cluster changes size and shape, and looses three-fold compositional symmetry. In the closed and open conformations of protein S, R1039 of each protomer bridges to E1031 (Figs. 6, S20 A,B). This core network of R1039-E1031 branches out via S1037 (Figs. 6, S19 A-B). In the closed conformation, each S1037 connects via H1048 to a local, intra-protomer H-bond cluster that includes R509 (Fig. 6). These core and local H-bond clusters are present, but disconnected, in the open conformation (Figs. 6, S19). In pre-fusion, the three-fold compositional symmetry of the R1039 H-bond cluster is lost, as each R1039 interacts with E1031 of a different protomer (Figs. 6, S19). In two of the protomers, R1039 further connects to S1037 and, in just one of the protomers, S1037 still connects to H1048 (Figs. S19 I-K, S20D).

We thus found that the closed conformation of protein S has a remarkable three-fold symmetrical central R1039 cluster and symmetrical R509 clusters at each of the three RBDs. These two clusters, and other H-bond clusters we identified, are altered in the open and pre-fusion conformations, indicating rearrangement of H-bond networks is part of the reaction coordinate of protein S. Rearrangements with loss of three-fold symmetry within H-bond clusters could be important for conformational selection for the binding of protein S to ACE2, and/or for proteolytic activation.

3.7. H-bond clusters extend deep across the ACE2-RBD binding interface

Infection with coronavirus largely depends on the affinity with which protein S binds to ACE2 (Li, 2013). Structures of ACE2 bound to SARS-CoV-2 protein S fragments (Lan et al., 2020; Yan et al., 2020), or to a chimera protein S fragment (Shang et al., 2020), identified H bonds and salt bridges between the RBD and ACE2 (Lan et al., 2020; Shang

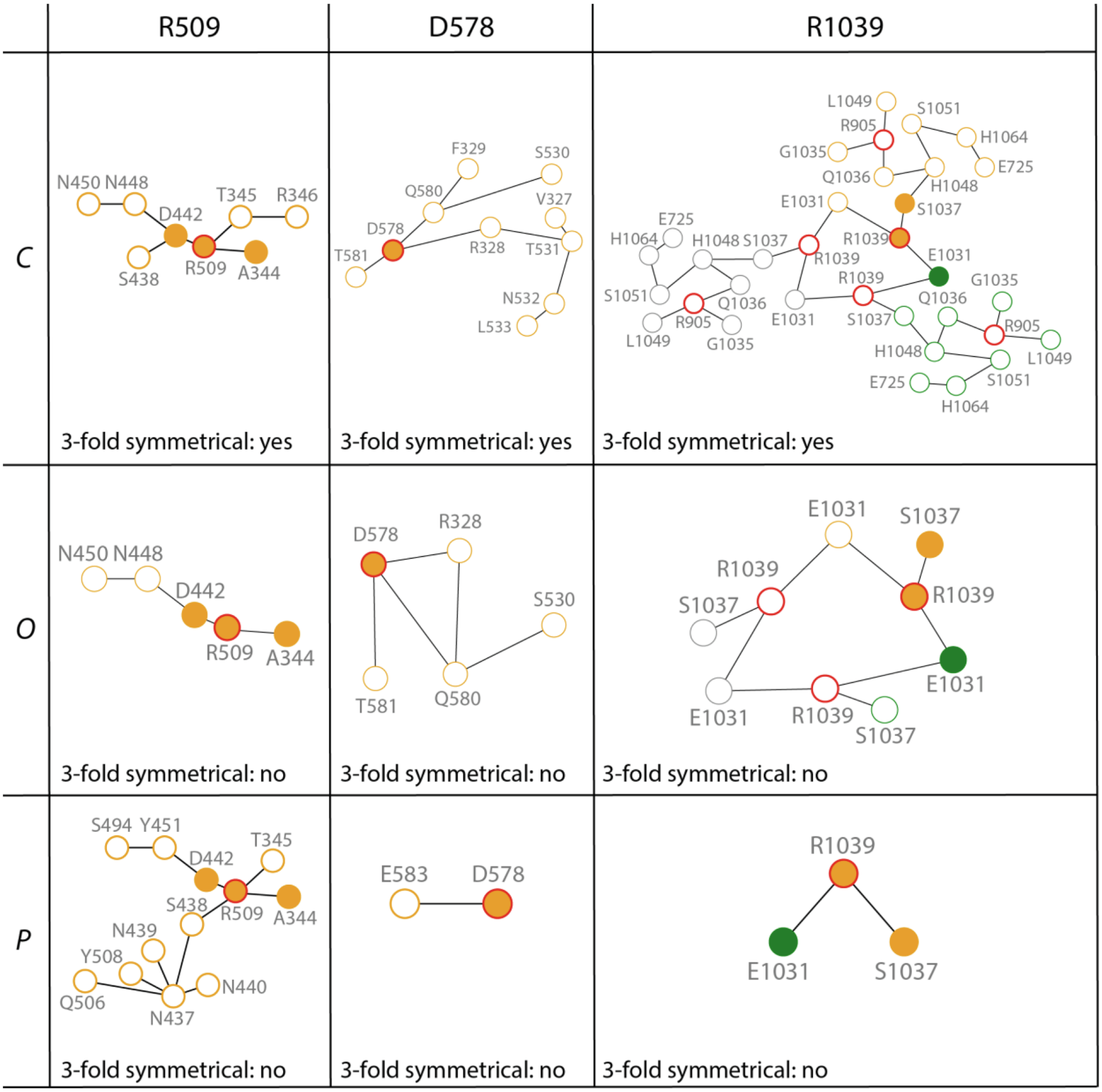


Fig. 6. Examples of clusters of spike protein S that have altered symmetry in open and/or pre-fusion conformations as compared to the closed conformation. H-bond clusters R509, D578, and R1039 are shown schematically for conformations *C*, *O*, and *P*. A cluster is considered as 3-fold symmetrical when the same groups are contributed by all three protomers of protein S.

et al., 2020; Yan et al., 2020). We wondered whether, instead of single H-bonds and salt-bridges, entire clusters of H-bonds, as we identified above for protein S, could mediate binding of protein S to ACE2.

From graphs of H-bonds for ACE2-RBD complexes (Fig. 9, S24-S34) and centrality values (Tables S4, S5), we found that the ACE2-RBD binding interface has 3–4 H-bond clusters, which we denote as the local interface clusters as *a*, *b*, *c*, and *d* (Fig. 9, S24-S34). In Fig. 9 we present H-bonds of the full-length ACE2 bound to the RBD, that are present in all ACE2-RBD structures we analyzed; each of these structures can have more H-bonds contributing to the interface clusters (Figs. S24-S29).

In full-length ACE2 bound to RBD (PDB ID:6M17, chains B, E) (Yan et al., 2020) cluster *a* has 16 RBD groups, and 10 ACE2 groups (Fig. 9C). In structure PDB ID:6M0J (Yan et al., 2020) cluster *a* is significantly larger, with 31 and 14 groups contributed by the RBD and ACE2, respectively (Figs. S29, S33B). The larger size of the interface cluster *a* in the latter structure could be due the resolution being slightly higher (Table 2), or to the protein conformation being different.

The four interface clusters include groups whose functional role has been probed experimentally (Table S4), groups that have relatively high centrality values (Tables S5, S6), and groups that are highly conserved (Fig. 9C, Tables S4, S7). For example, N501 of SARS-COV-2 protein S is T487 in SARS-CoV protein S, in which the methyl group of the Thr is thought important for the binding of the RBD to ACE2 (Li et al., 2005); T487 is among the amino acid residues essential for the binding of SARS-CoV to ACE2. Y449 corresponds to Y436 of SARS-CoV protein S, and is conserved in SARS spike proteins that use ACE2 (Hoffmann et al., 2020). N437 and D442 of the interface cluster *a* (Fig. 9C) are also part of the high-centrality cluster R509 identified in the *up* protomer of the pre-fusion conformation of isolated protein S (Fig. 3A, 3B, 6). In cluster *b*, Q493 is part of a H-bond network that includes ACE2 K31 (Fig. 6C) a group considered a virus binding hot spot (Wan et al., 2020). This cluster contains two other ACE2 groups, and two from the RBD.

Given the extensive H-bond network we observe for ACE2 (Fig. 9A), and the large number of ACE2 groups that participate in interface H-bonding (Fig. 9C,D), we wondered whether ACE2 groups that participate in interface H-bond clusters might reach the vicinity of the

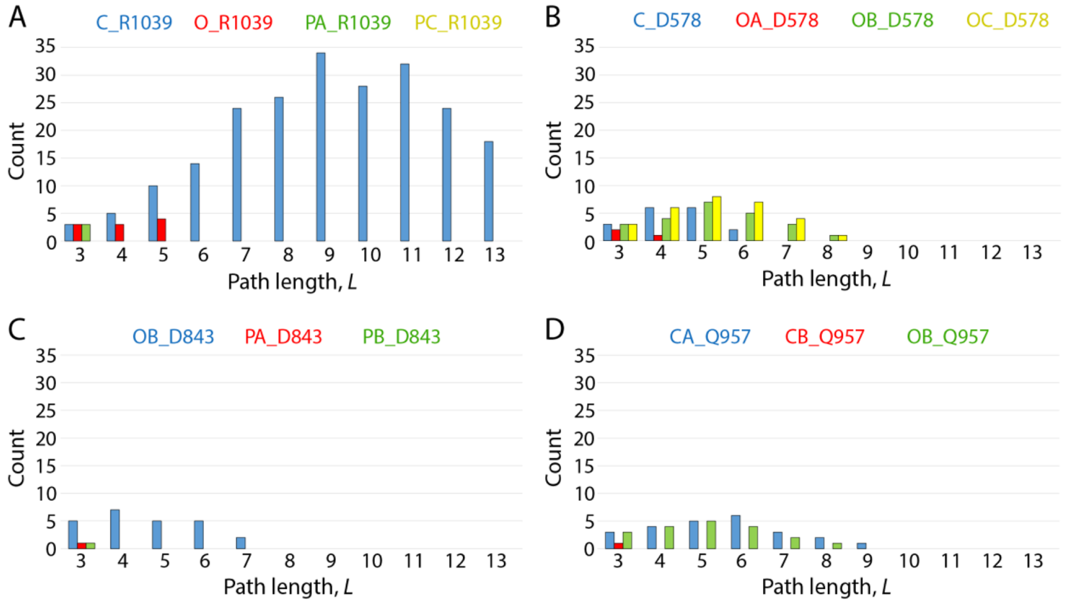


Fig. 7. Shortest path length for selected H-bond clusters in the closed vs. open conformations. We consider, for each cluster, all continuous shortest paths that include the H-bonding group with highest-centrality. (A-D) Distribution of shortest path lengths computed for selected H-bond clusters. Illustrations of the connectivity within the D578, D843, Q957, and R1039 H-bond clusters are presented in, respectively, Figs. S17, S22, S23, and S19.

catalytic site of the enzyme, or groups known to be otherwise important for the functioning of ACE2.

R273, H345, E375, H505, and Y515 delineate an inhibitor-binding site of ACE2 (Guy et al., 2005), whereas H374, H378, and E402 coordinate the Zn^{2+} ion (Towler et al., 2004). According to our centrality computations, R273 is within 4 amino acid residues of the high-centrality group D269 (Fig. 9A,B), and H378 has relatively high average *DC* value.

H-bond clusters of R273, H345, H505, H378, H374, and E402 (Fig. S34) lack common H-bonds with the conserved interface clusters presented in Fig. 9A. Nevertheless, the H378 H-bond cluster (Fig. 9B) is relatively close to the interface (Fig. 9B, S34C), N394 can be part of interface cluster *a* (Fig. S33B), and N397 is part of the H405 cluster of ACE2 (Fig. S34B). Thus, depending on the protein conformation, interface H bonding of ACE2 can extend deep into the protein, towards the active site. Together with the interface H-bond clusters (Fig. 9), the overall picture that emerges is that the binding interface between the RBD and ACE2 is mediated by extensive H-bonding, which could

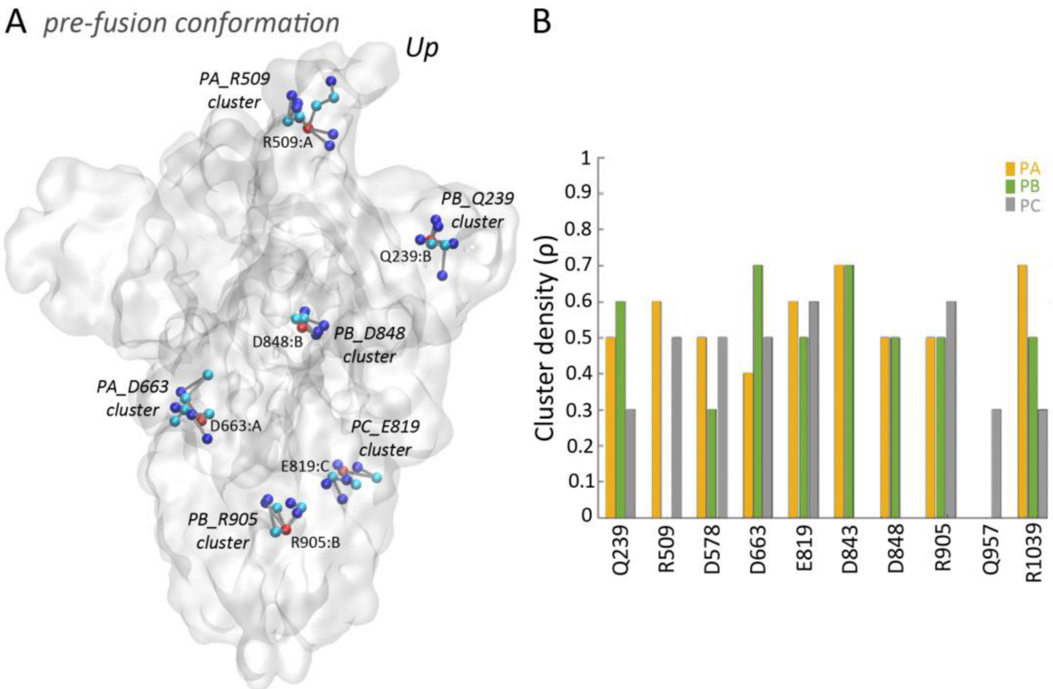


Fig. 8. H-bond clusters in protomers of the pre-fusion conformation can have different composition. (A) Selected H-bond clusters with their anchor, internal, and high-*BC* nodes colored blue, cyan, and red, respectively. (B) Cluster density ρ for high-*BC* groups found in the three protein conformations. The cluster density ρ is given by the number of anchor nodes divided by the total number of nodes of the graph. Analyses of cluster density for the closed and open conformations are presented in Figure S21.

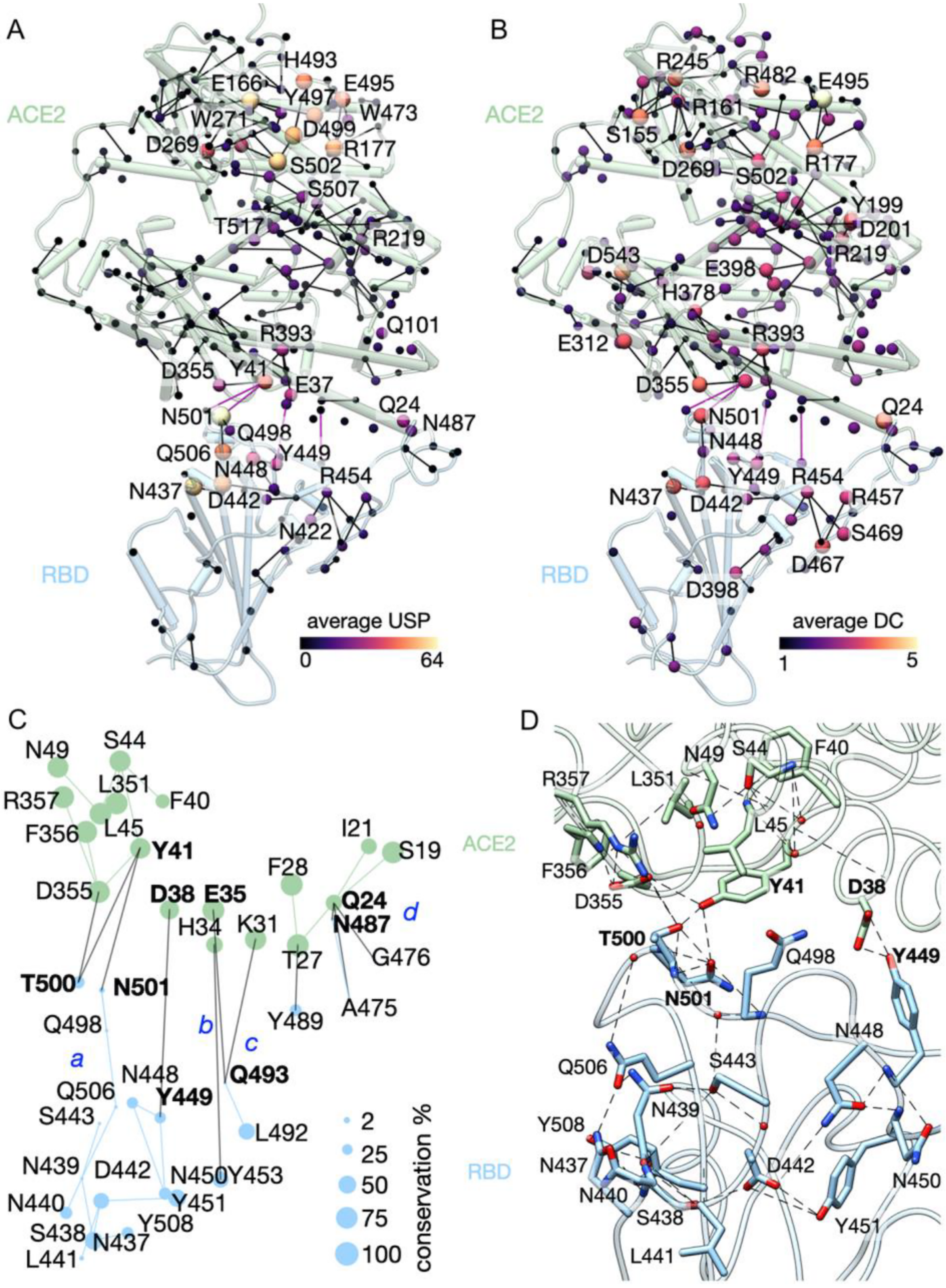


Fig. 9. H-bond clusters at the interface between ACE2 and the RBD of SARS-CoV-2. (A, B) Graph of H-bonds with labels for high-centrality groups that participate in conserved networks. Small spheres indicate Cα atoms, and are colored according to average centrality values. Pink lines represent inter-protein interactions. We label here only groups with high centrality, and give additional labels in Figure S29; the complete H-bond graph is presented in Figure S30. Molecular graphics are based on structure PDB ID:6M0J (Lan et al., 2020). (C) H-bond clusters at the binding interface between ACE2 (green dots) and the RBD (blue dots). Clusters *a*, *b*, *c*, and *d*, include amino acid-residues inter-connected via H-bonds shown as gray for inter-protein interactions. The size of the node represents the sequence conservation of the amino acid residue. Labels in bold indicate H-bonding between ACE2 and the RBD in all structures we analyzed. (D) Molecular graphics of cluster *a*. Dash lines indicate H-bonding. Panels C and D are based on the structure of full-length ACE2 bound to RBD, PDB ID:6M17 (Yan et al., 2020), chains B and E. Additional images and analyses of H-bond clusters at the interface between ACE2 and the RBD are presented in Figs. S24–S28, S31–S34. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

explain the strong binding observed in experiments (Tai et al., 2020).

That an entire local, dynamic H-bond network, could be important for protein binding, was observed before for arrestin (Ostermeier et al., 2014), and molecular dynamics simulations of an extrinsic subunit of the photosystem II complex indicated participation of protein sidechains in dynamic interaction clusters could be beneficial for protein binding (del Val and Bondar, 2017): In a dynamic H-bond cluster, multiple H-bonds that break and reform rapidly provide structural plasticity, such that a protein sidechain essential for protein–protein binding can sample orientations compatible with binding; by contrast, when in the unbound state, i.e., prior to protein–protein binding, that sidechain is engaged in a stable interaction whose perturbation is energetically costly, the availability of the sidechain for new interactions needed to mediate binding may be reduced.

3.8. *Corona spike protein S sequences carry a significant net negative charge and contain patches with carboxylate and carboxylate-histidine motifs*

Knowledge of amino acid residues that could transiently bind protons during the functioning of protein S is important, as protonation and changes in protonation can alter protein dynamics (del Val et al., 2014; Lazaratos et al., 2020). *In vitro* experiments on the recombinant ectodomain of SARS-CoV suggested that, at pH between 5 and 6, monomers form trimers irreversibly (Li et al., 2006). As fusion of protein S with the host cell can occur at neutral pH (Xiao et al., 2003), and conformational changes of the ectodomain can occur independent of the pH (Walls et al., 2017), it remains unclear whether or not protein S binds protons, and whether proton binding shapes conformational dynamics relevant to viral infectivity.

In the future, high-resolution structures might inform on likely protonation states of specific protein groups and facilitate experiments and computations to probe pH sensitivity. To evaluate whether the sequence of protein S contain clues about sites where a proton might bind, here we analyzed sequences of spike proteins to calculate an estimated net charge and identify conserved sequence motifs of carboxylate and histidine groups. We found that the 47 protein S sequences included in Set-A have full length in the range of 1235–1363 amino acid residues (Fig. 10A, see also Supplementary Sequence Analyses). SARS-CoV and SARS-CoV-2 proteins S are among the shorter sequences of Set-A. All protein sequences have an estimated net negative charge between −3*e* and −25*e* (Fig. 10B); for SARS-CoV-2 protein S, the net estimated charge is −7*e*.

The shortest spike protein sequences, 1236 amino acid residues, are of the mouse hepatitis virus; this sequence also carries the smallest estimated charge. The four longest sequences, 1363 amino acid residues, are of corona viruses isolated from rabbit, deer, bovine, and horse; these are also the sequences with significant negative charges of −20*e* or −23*e* (Fig. 10B). The large estimated net charges arise from the numerous charged and polar groups carried by the sequences (Figs. S35, S36), with ~ 100–118 Asp/Glu (Fig. S35A), and ~ 85–105 Arg/Lys (Fig. 35B). Protein S is also subject to mutations that can affect H-bonding groups (Korber et al., 2020) (Table S8), and thus its overall polarity.

A wide range for the net charge carried by sequences of a protein from different organisms, as found here for spike proteins S, was observed before for other proteins (del Val and Bondar, 2017, 2020), suggesting the net estimated charge might be related to details of the cellular environments and/or protein interactions. The actual net

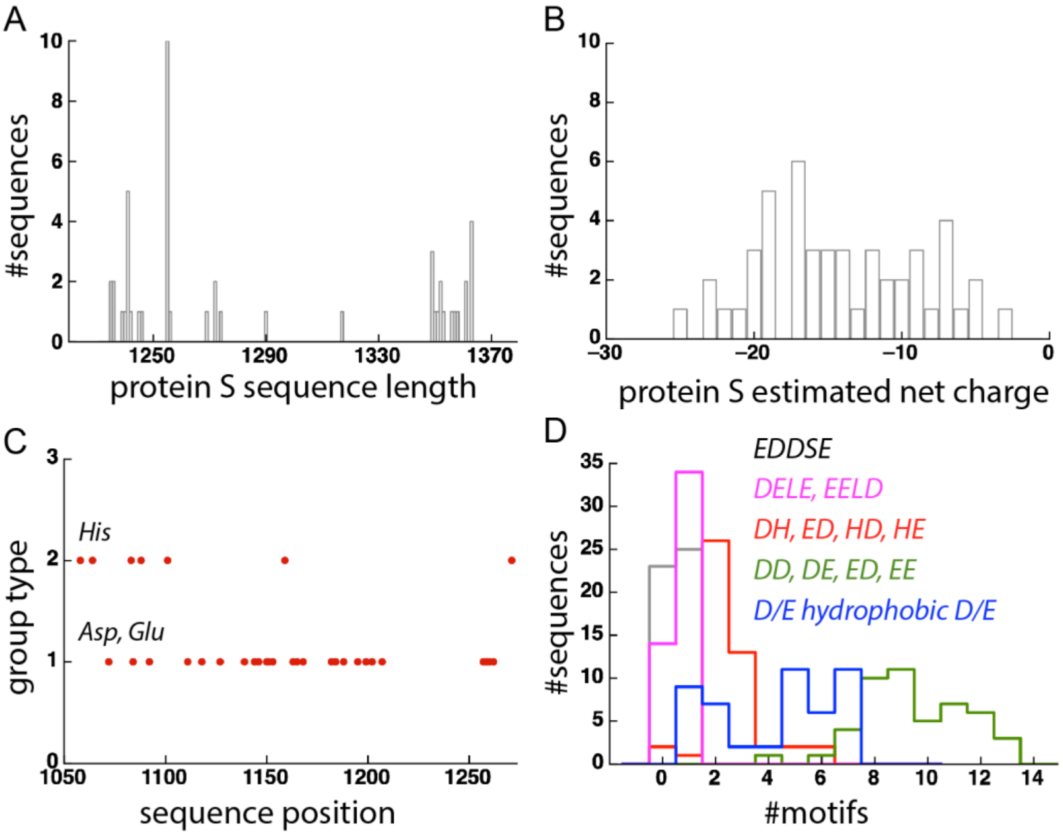


Fig. 10. Sequence length and charges of corona proteins S. Analyses were performed for the 48 sequences from *Set-A*. (A, B) Histogram of the full length (panel B) and of the net estimated charge (panel B) of corona proteins S. (C) Position of Asp, Glu, and His groups along the amino acid sequence of SARS-CoV-2 protein S; for clarity, only the C-terminal region is shown. Group type = 1 indicates Asp or Glu, and group type = 2 indicates His. (D) Histogram of the number of selected motifs identified in *Set-A*. Gray indicates EDDSE; green, DD, DE, ED, or EE; red, DH, EH, HD, or HE; blue, DFD, DGD, DID, DLD, DEL, DVD, DVE, EAE, EID, ELD, ELE, or EVD. Additional sequence analyses of *Set-A* proteins are presented in Figs. S35-S38.

charge of a protein will depend on protonation in the cellular environment in which the protein is found.

When aligning the RBD of SARS-CoV-2 with the sequences of the other 47 protein S sequences from *Set-A*, we obtained regions with 142 – 241 amino acid residues (Fig. S37A) and an estimated net charge that tends to be positive (Fig. S37B). The variation in the length of the region corresponding to SARS-CoV-2 RBD, and the associated variation in estimated charge, is due to numerous deletions and insertions of amino acid residues in other protein S sequences from Set-A (see Supporting Information Sequence Alignments).

A slightly positive net charge of the RBD of protein S could be important for the binding of protein S to ACE2: The electrostatic potential surfaces of both SARS-CoV-2 protein S and ACE2 have patches that are predominantly negative or positive; at the binding interface, ACE2 exposes a predominantly negative surface, whereas the potential energy surface for the RBD is predominantly positive (Figs. S39-S43). The complementary electrostatic potential surfaces at the region where the RBD binds to ACE2 are compatible with the presence of multiple H-bonds and H-bond clusters we identified (Fig. 9).

Given the large number of carboxylate groups and net negative charge carried by protein S sequences, we sought to find out whether protein S has patches of carboxylate and histidine groups, as such patches are of potential interest for proton binding (Bondar and Lemieux, 2019; Checover et al., 2001; Gerland et al., 2020; Guerra and Bondar, 2015; Kemmler et al., 2019; Lorch et al., 2015; Shutova et al., 2007). We found that the sequence of SARS-CoV-2 protein S has multiple positions where Asp, Glu or His groups are adjacent or separated by 1–2 amino acid residues; of the 17 His groups of the protein, 7 are within the C-terminal ~220 residue fragment (Fig. 10C, S38A). Moreover, particular arrangements, or motifs, appear multiple times in the sequence of SARS-CoV-2 protein S (Fig. 10C, S38, S44A). There are several positions with 2–3 carboxylates consecutive in the sequence; close to the C-terminus, there is a DEDSE motif (Fig. S44B).

Such carboxylate and carboxylate-histidine motifs appear to be a more general feature of the spike proteins. Of the 48 sequences from Set-A, only two lack any carboxylate-histidine motif; 2 vs. 3 motifs are present in 26 vs. 13 sequences (Fig. 10D). Motifs with two consecutive carboxylate groups appear relatively frequently: 22 of the Set-A sequences have 8–9 such motifs (Fig. 10D). The more restrictive DELE or EELD, and EDDSE, motifs are present only once in 34 and 25 sequences, respectively (Fig. 10D). SARS-CoV-2 protein S has a (D)EDDSE motif at the C-tail, immediately upstream the Cys-rich region (Fig. S38) where palmitoylation occurs (Fig. 1A) (Petit et al., 2007). Such a motif is also observed in SARS-CoV protein S, whereas several other coronavirus proteins S have here two carboxylates adjacent to a highly conserved Cys group (Petit et al., 2007). It is unclear whether this carboxylate motif is related to palmitoylation or to another functional role. Further experiments and computations will be needed to ascertain whether the conserved carboxylate and carboxylate-histidine motifs we identified might be involved in transient proton binding, or whether they might instead contribute toshaping protein dynamics and protein interactions.

4. Conclusions

Binding of the coronavirus protein S to the host ACE2 receptor initiates a series of reactions that include large-scale structural rearrangements of protein S (Shulla and Gallagher, 2009), changes in the structure and expression of the receptor, proteolitic cleavage followed by conformational change thought to assist viral entry (Kam et al., 2009), local dehydration and ordering of the membrane upon interaction with the fusion peptide (Lai et al., 2017), and culminating with membrane fusion and virus entry (Wang et al., 2020b). This is a highly complex reaction coordinate whose molecular movie could facilitate the development of therapeutics.

As a first step towards deciphering interactions that govern structural plasticity of SARS-CoV-2 protein S, here we focused on H-bonding and H-bond clusters, as H-bonding and H-bond clusters shape protein conformational dynamics (Bondar and White, 2012; Joh et al., 2008), and H-bond networks are central to working models of long-distance conformational couplings in proteins (Karathanou and Bondar, 2018; Venkatakrishnan et al., 2019).

Using graph-based approaches to identify and characterize H-bond clusters of protein S (Schemes 1–3), we found that the open and pre-fusion conformations are distinguished by rearrangements of H-bond clusters at discrete sites of the protein, suggesting H-bond clusters could be used to characterize conformational dynamics of protein S.

The RBDs of the closed conformation host a H-bond cluster centered at R509, with the same groups contributing to the clusters of each protomer (Fig. 5A, 6, 11). In the vicinity of the S1/S2 cleavage site, each protomer has a H-bond cluster centered at D578 (Fig. 5A, 6, 11). Both the R509 and D578 clusters are relatively large, with 8 and, respectively, 10 H-bonding groups (Fig. 6).

Close to the stalk region of the ectodomain in closed conformation, the same 11 groups of each protomer participate in the central cluster, in which the core H-bond network of the E1031, S1037, and R1039 groups branches out via H1048 to local networks centered at R905 (Fig. 5A, 6, 11).

In the open and pre-fusion conformations, the R509, D578, and R1039 clusters have rearranged significantly. The R509 H-bond cluster of the open conformation is reduced to 5–6 H-bonds, or to even one H-bond (Figs. 6, 11, S16). In pre-fusion, only the RBD *up* protomer has a R509 H-bond cluster (Fig. 5A-C, 6, S16). This cluster includes N437 and N439, which are also part of the H-bond cluster that contributes to the binding interface between protein S and ACE2 (Figs. 9, 11). The D578 H-bond cluster is reduced by about half in one of the protomers of the open conformation, and replaced by singular H-bonds in two protomers of the pre-fusion conformation (Figs. 6, 11, S17, S18). The central R1039 cluster of the open conformation is separated into the central core and the R905 branch; in pre-fusion the core H-bond cluster is absent, though R1039 still makes bridges to E1031 of another protomer (Figs. 6, 11, S19).

Thus, three major H-bond clusters of protein S, which in the closed conformation are contributed by the same groups of each of the three protomers, rearrange drastically in the open and pre-fusion conformation, and are distinct in each of the conformers (Fig. 11). Such structural rearrangement, whereby the three protomers of protein S experience different H-bonding, could facilitate conformational selection of a protomer for the binding to ACE2, and/or for proteolytic cleavage for activation.

The three major H-bond clusters we identified for protein S, and the H-bond clusters at the RBD-ACE2 interface, include carboxylate and histidine groups, and sequences of protein S contain short patches of such groups (Fig. 10D). Patches of closely spaced carboxylate groups could be involved in transient binding of protons (Checover et al., 2001; del Val and Bondar, 2017; Shutova et al., 2007), and/or for enhanced local protein dynamics (del Val and Bondar, 2017). Whether carboxylate- and histidine-containing H-bond clusters change protonation during the functioning of protein S, making the conformational dynamics of SARS-CoV-2 protein S sensitive to the local pH, is unclear. A sensitivity to pH was noted before for the spike protein of HA, which is thought to undergo a conformational change at low pH in the endosome (Millet and Whittaker, 2018), low pH enabling transition of HA from the pre-fusion to the fusogenic conformation (Eckert and Kim, 2001), and the fusion peptide to approach the host membrane (Bullough et al., 1998). In the future, experiments and computations will be needed to derive accurate information on sites where protons might bind on the surface of protein S, and evaluate couplings between protonation change and protein dynamics.

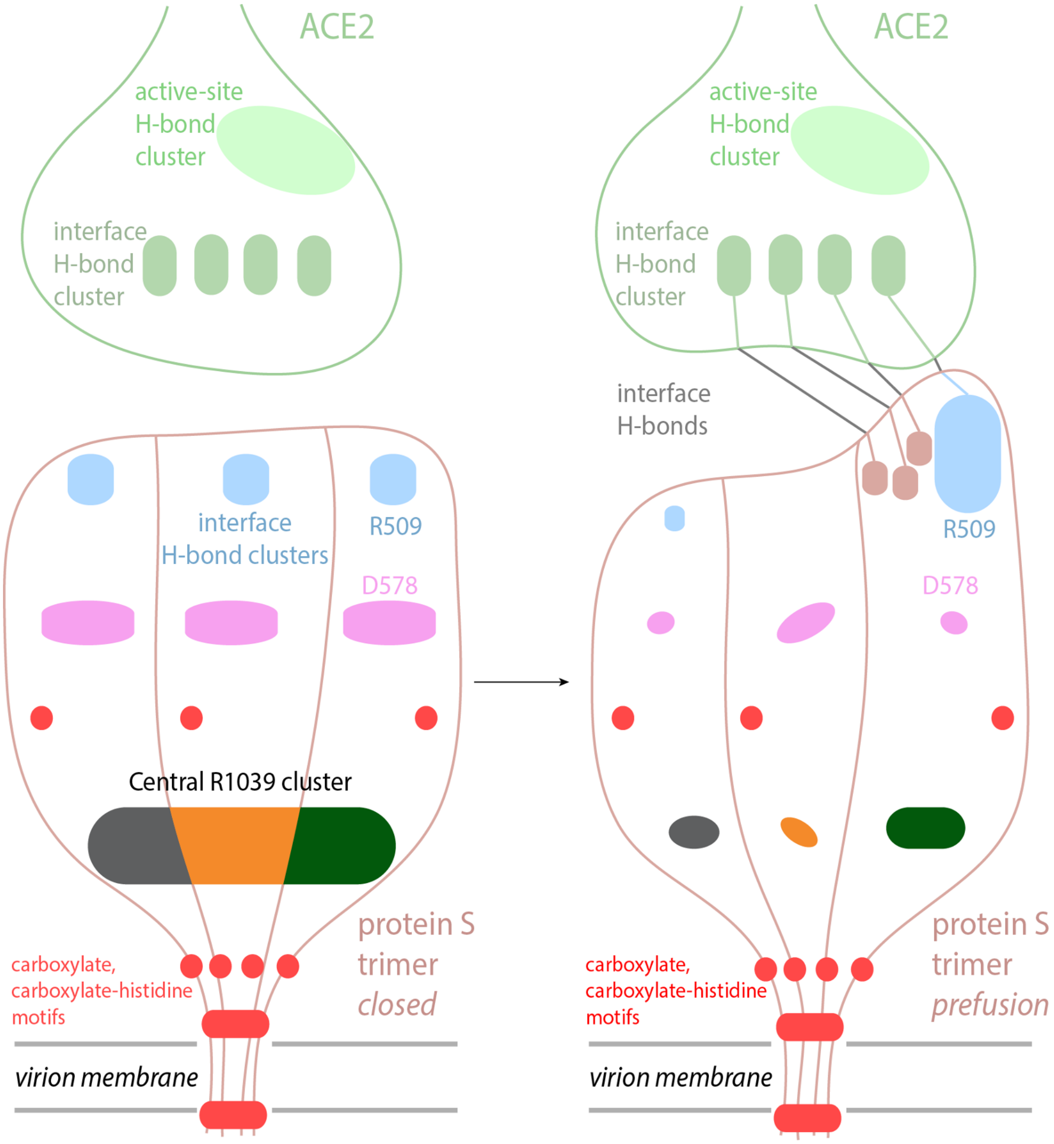


Fig. 11. Cartoon representation of main observations presented here. Protein S binds to ACE2 via H-bond clusters, such that binding between protein S and ACE2 is characterized by extensive H-bonding across the interface. The closed conformation of protein S has three H-bond clusters with three-fold symmetrical composition – the D578 cluster near the S1/S2 cleavage site, the R509 cluster at the RBD, and the central R1039 cluster. These three H-bond clusters loose symmetry in the open and/or pre-fusion conformations. In ACE2, the interface H-bond network extends to the vicinity of the active site. Protein S has numerous carboxylate motifs, some of which we suggest could bind protons and/or shape local protein dynamics.

A caveat of our approach is that it lacks description of the hydrophobic interactions, which could bring important contributions to protein stability, protein binding interfaces, and membrane fusion. Moreover, hydrophobic packing of H-bonding sidechains shapes H-bond dynamics (del Val et al., 2014). Our preliminary tests suggest, for example, that hydrophobic clusters are located close to the R509 and D578 H-bond clusters in the closed conformation of protein S. In the future, we plan to extend our graph-based approach to dissect the interplay between H-bonding and hydrophobic interactions in protein S.

The work presented here relied on static coordinate snapshots of protein S and of protein S fragments bound to ACE2. We anticipate that, as structures solved at increasingly higher resolution might become available, the conformational dynamics of protein S could be described more accurately, and that knowledge about the kinetics of transitions between intermediate conformations, and of protonation states of specific protein groups, will allow us to derive a more complete picture of the reaction coordinate of protein S in fluid, physiologic environments. The graph-based analyses presented here, combined with prolonged molecular dynamics simulations could enable estimations of the energetics associated with breaking and reforming of H-bonds within clusters of protein S, and of other large protein complexes.

CRediT authorship contribution statement

Konstantina Karathanou: Investigation, Methodology, Visualization, Writing - original draft, Writing - review & editing. **Michalis Lazaratos:** Investigation, Methodology, Visualization, Writing - original draft, Writing - review & editing. **Éva Bertalan:** Investigation, Methodology, Visualization, Writing - original draft. **Malte Siemers:** Methodology, Validation, Writing - original draft. **Krzysztof Buzar:** Investigation, Visualization, Writing - original draft. **Gebhard F.X. Schertler:** Conceptualization, Writing - review & editing. **Coral del Val:** Conceptualization, Investigation, Visualization, Writing - original draft, Writing - review & editing. **Ana-Nicoleta Bondar:** Conceptualization, Supervision, Visualization, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

GFXS acknowledges support in part from The PSI COVID19 Emergency Science Fund. CdV acknowledges support in part from the Spanish Ministry of Science, Innovation and Universities, grant RTI2018-098983-B-I00. A-NB acknowledges support in part by the Excellence Initiative of the German Federal and State Governments via the Freie Universität Berlin and from the German Research Foundation DFG SFB 1078, Project C4. A-NB and GFXS thank Dr. Gregor Chicchetti (PSI) for valuable comments on the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsb.2020.107617.

References

Altschul, S.F., Madden, T.L., Scäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acid Res. 25, 3389–3402.

Amitai, G., Shemesh, A., E., S., Shklar, M., Netanely, D., Venger, I., Pietrovski, S., 2004. Network analysis of protein structures identifies functional residues. J. Mol. Biol. 344, 1135-1146.

Auton, A., Brooks, L.D., 2015. A global reference for human genetic variation. Nature 526, 68–74.

Babcock, G.J., Esshaki, D.J., Thomas, W.D.J., Ambrosino, D.M., 2004. Amino acids 270 to 510 of the severe acute respiratory syndrome spike protein are required for interaction with the receptor. J. Virology 78, 4552–4560.

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., McCammon, A.J., 2001. Electrostatics of nanosystems: application to microtubules and the ribosomes. Proc. Natl. Acad. Sci. 98, 10037–10041.

Belouzard, S., Chu, V.C., Whittaker, G.R., 2009. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. Proc. Natl. Acad. Sci. 106, 5871–5876.

Belouzard, S., Millet, J.K., Licitra, B.N., Whittaker, G.R., 2012. Mechanism of coronavirus cell entry mediated by the viral spike protein. Viruses 4, 1011–1033.

Bender, E.A., Williamson, S.G., 2010. Lists, Decisions and Graphs. With an Introduction to Probability University of California, San Diego.

Beniac, D.R., deVarennes, S.L., Andonov, A., He, R., Booth, T.F., 2007. Conformational reorganization of the SARS coronavirus spike following receptor binding: implications for membrane fusion. PLoS ONE 2, e1082.

Berman, H.M., Westbrook, J., Feng, G., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucl. Acid Res. 28, 235–242.

Bondar, A.-N., White, S.H., 2012. Hydrogen bond dynamics in membrane protein function. Biochim. Biophys. Acta 1818, 942–950.

Bondar, A.-N., Lemieux, H.J., 2019. Reactions at membrane interfaces. Chem. Rev. 119, 6162–6183.

Bosch, B.J., Martina, B.E.E., van der Zee, R., Lepault, J., Haijema, B.J., Versluis, C., Heck, A.J.R., de Groot, R., Osterhaus, A.D.M.E., Rottier, P.J.M., 2004. Severe acute respiratory syndrome coronavirus (SARS-CoV) infection inhibition using spike protein heptat repeat-derived peptides. Proc. Natl. Acad. Sci. 101, 8455–8460.

Brandes, U., 2001. A faster algorithm for betweenness centrality. J. Mathematical Sociol. 25, 163–177.

Briefing, 2020. Anatomy of a killer. The Economist March 14, 14-16.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M., 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 4, 187–217.

Bullough, P.A., Hughson, F.M., Skehel, J.J., Wiley, D.C., 1998. Structure of influenza haemagglutinin at the pH of membrane fusion. Nature 371, 37–43.

Chakraborti, S., Prabakaran, P., Xiao, X., Dimitrov, D.S., 2005. The SARS coronavirus S glycoprotein receptor binding domain: The mapping and functional characterization. Virology J. 2, 1–10.

Checover, S., Marantz, Y., Nachliel, E., Gutman, M., 2001. Dynamics of the proton transfer reaction on the cytoplasmic surface of bacteriorhodopsin. Biochemistry 40, 4281–4292.

Colman, P.M., Lawrence, M.L., 2003. The structural biology of type I viral membrane fusion. Nature Rev. Mol. Cell Biol. 4, 309–319.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Sten, C., 2009. Introduction to Algorithms, third ed. Massachusetts Institute of Technology.

del Val, C., Bondar, A.-N., 2017. Charged groups at binding interfaces of the PsbO subunit of photosystem II: a combined bioinformatics and simulation study. Biochim. Biophys. Acta 1858, 432–441.

del Val, C., Bondar, A.-N., 2020. Diversity and sequence motifs of the bacterial SecA protein motor. Biochim. Biophys. Acta 1862, 183319.

del Val, C., Bondar, L., Bondar, A.-N., 2014. Coupling between inter-helical hydrogen bonding and water dynamics in a proton transporter. J. Struct. Biol. 186, 95–111.

Dolinsky, T.J., Nielsen, J.E., McCammon, J.A., Baker, N.A., 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucl. Acid Res. 32, W665–W667.

Eckert, D.M., Kim, P.S., 2001. Mechanisms of viral membrane fusion and its inhibition. Annu. Rev. Biochem. 70, 777–810.

Espinosa, E., Souhassou, M., Lachekar, H., Lecomte, C., 1999. Topological analysis of the electron density in hydrogen bonds. Acta Cryst. B55, 563–572.

Fokas, A.S., Cole, D.J., Ahnert, S.E., Chin, A.W., 2016. Residue geometry networks: a rigidity-based approach to the amini acid network and evolutionary rate. Sci. Rep. 6, 33213.

Freeman, L.C., 1977. A set of measures of centrality based on betweenness. Sociometry 40, 35–41.

Freeman, L.C., 1979. Centrality in social networks. Conceptual clarification. Social Networks 1, 215-239.

Gerland, L., Friedrich, D., Hopf, L., Donovan, E.J., Wallmann, A., Erdmann, N., Diehl, A., Bommer, M., Buzar, K., Ibrahim, M., Schmieder, P., Dobbek, H., Zouni, A., Bondar, A.-N., Dau, H., Oschkinat, H., 2020. pH-dependent protonation of surface carboxylates in PsbO enables local buffering and triggers structural changes. ChemBioChem 21, 1597–1604.

Graham, R.L., Baric, R.S., 2010. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J. Virol. 3134–3146.

Grubaugh, N.D., Hanage, W.P., Rasmussen, A.L., 2020. Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. Cell 182, 1–2.

Guerra, F., Bondar, A.-N., 2015. Dynamics of the plasma membrane proton pump. J. Membr. Biol. 248, 443–453.

Gui, M., Song, W., Zhou, H., Xu, J., Chen, S., CXiang, Y., Wang, X., 2017. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. Cell Res., 27, 119-129.

Guy, J.L., Jackson, R.M., Jense, H.A., Hooper, N.M., Turner, A.J., 2005. Identification of critical active-site residues in angiotensin-converting enzyme-2 (ACE2) by site-directed mutagenesis. FEBS J. 272, 3512–3520.

Hagberg, A.A., Schult, D.A., Swart, P.J., 2008. Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Phyton in Science Conference (SciPy 2008), 11-16.

Harris, A., Lazaratos, M., Siemers, M., Watt, E., Hoang, A., Tomida, S., Schubert, L., Saita, M., Heberle, J., Furutani, Y., Kandori, H., Bondar, A.-N., Brown, L.S., 2020. Mechanism of inward proton transport in an Antartic microbial rhodopsin. J. Phys. Chem. B 124, 4851–4872.

Hatcher, E.L., Zhdanov, S.A., Bao, Y., Blinkova, O., Nawrocki, E.P., Ostapchuck, Y., Schaffer, A.A., Brister, J.R., 2017. Virus Variation Resource - improved response to emergent viral outbreaks. Nucl. Acid Res. 45, D482–D490.

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichse, S., Schiergens, T.S., Herrler, T., Wu, N.-H., Nitsche, A., Müller, M.A., Drosten, C., Pöhlmann, S., 2020. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181, 1–10.

Hofmann, H., Pöhlmann, S., 2004. Cellular entry of the SARS coronavirus. Trends Microbiol. 12, 466–472.

Hong, H., Szabo, G., Tamm, L.K., 2006. Electrostatic couplings in OmpA ion-channel gating suggest a mechanism for pore opening. Nature Chem. Biol. 2, 627–633.

Humphrey, W., Dalke, W., Schulten, K., 1996. VMD: visual molecular dynamics. J. Mol. Graph. 14, 33–38.

Hunt, S.E., McLaren, W., Gil, L., Thorman, A., Schuilenburg, H., Sheppard, D., Parton, A., Armean, I.M., Trevanion, S.J., Flicek, P., Cunningham, F., 2018. Ensembl variation resources. Database 2018, 1–12.

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.

Jo, S., Kim, T., 2008. CHARMM-GUI Solvator.

Joh, N.H., Min, A., Faham, S., Whitelegge, J.P., Yang, D., Woods Jr., V.L., Bowie, J.U., 2008. Modest stabilization by most hydrogen-bonded side-chain interactions in membrane proteins. Nature 453, 1266–1270.

Kam, Y.-W., Okumura, Y., Kido, H., Ng, L.F.P., Bruzzone, R., Altmeyer, R., 2009. Cleavage of the SARS coronavirus spike glycoprotein by airway proteases enhances virus entry into human bronchial epithelial cells *in vitro*. PLoS ONE 4, e7870.

Karathanou, K., Bondar, A.-N., 2018. Dynamic hydrogen bonds in bacterial protein secretion. FEMS Microbiol. Lett. 365, fny124.

Karathanou, K., Bondar, A.-N., 2019. Using graphs of dynamic hydrogen-bond networks to dissect conformational coupling in a protein motor. J. Chem. Inf. Model. 59, 1882–1896.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT sequence alignment program. Brief. Bioinform. 9, 286–298.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Kemmler, L., Ibrahim, M., Dobbek, H., Zouni, A., Bondar, A.-N., 2019. Dynamic water bridging and proton transfer at a surface carboxylate cluster of photosystem II. Phys. Chem. Chem. Phys. 21, 25449–25466.

Kirchdoerfer, R.N., Wang, N., Pallesen, J., Wrapp, D., Turner, H.L., Cottrell, C.A., Corbett, K.S., Graham, B.S., McLellan, J.S., Ward, A.B., 2018. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. Sci. Rep. 8, 15701.

Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H.W., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., Hastie, K.M., Parker, M.D., Partridge, D.G., Evans, C.M., Freeman, T.M., de Silva, T.I., McDanal, C., Perez, L.G., Tang, H., Moon-Walker, A., Whelan, S.P., LabRanche, C.C., Saphire, E.O., Montefiori, D.C., 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182, 1–16.

Lai, A.L., Millet, J.K., Daniel, S., Freed, J.H., Whittacker, G.R., 2017. The SARS-CoV fusion peptide forms an extended bipartite fusion platform that perturbs membrane order in a calcium-dependent manner. J. Mol. Biol. 429, 3875–3892.

Lan, J., Ge, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., Wang, X., 2020. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature.

Larremore, D. 2014. Find Network Components. Version 1.2.0.0 MathWorks.

Lazaratos, M., Karathanou, K., Bondar, A.-N., 2020. Graphs of dynamic H-bond networks: from model proteins to protein complexes in cell signaling. Curr. Opin. Struct. Biol. 64, 79–87.

Lee, J., Chheng, X., Swails, J.M., Yeom, M.S., Eatsman, P.K., Lemkul, J.A., Wei, S., Buckner, J., Jeong, J.C., Qi, Y., Jo, S., Pande, V.S., Case, D.A., C.L., B., A.D., M.J., Klauda, J.B., Im, W., 2016. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM SImulations using the CHARMM36 additive force field. J. Chem. Theory Comput. 12, 405-413.

Li, F., 2013. Receptor recognition and cross-species infections of SARS coronavirus. Antiviral Res. 100, 246–254.

Li, F., Berardi, M., Li, W., Farzan, M., Dormitzer, P.R., Harrison, S.C., 2006. Conformational states of the severe acute respiratory syndrome coronavirus spike protein ectodomain. J. Virol. 80, 6794–6800.

Li, W., Moore, M.J., Vasilieva, N., Sui, J., Wong, S.K., Berne, M.A., Somasundaran, M., Sullivan, J.L., Luzuriaga, K., Greenough, T.C., Choe, H.-W., Farzan, M., 2003. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature 426, 450–454.

Li, W., Zhang, C., Sui, J., Kuhn, J.H., Moore, M.J., Luo, S., Wong, S.-K., Huang, I.-C., Xu, C., Vasilieva, N., Murakami, A., He, Y., Marasco, W.A., Guan, Y., Choe, H., Farzan, M., 2005. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO J. 24, 1634–1643.

Li, Z., Tomlinson, A.C.A., Wong, A.H.M., Zhou, D., Desforges, M., Talbot, P.J., Benlekbir, S., Rubinstein, J.L., Rini, J.M., 2019. The human coronavirus HCoV-229E S-protein structure and receptor binding. eLife 8, e51230.

Lorch, S., Capponi, S., Pieront, F., Bondar, A.-N., 2015. Dynamic carboxylate/water networks on the surface of the PsbO subunit of photosystem II. J. Phys. Chem. B 119, 12172–12181.

MacKerell Jr., A.D., Bashford, D., Bellot, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, W.E.I., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., Karplus, M., 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B 102, 3586–3616.

Marti-Renom, M.A., Stuart, A., Fiser, A., Sanchez, R., Melo, F., Sali, A., 2000. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biomol. Struct. 29, 291–325.

Millet, J.K., Whittaker, G.R., 2014. Host cell entry of Middle East respiratory syndrome after two-step, furin-mediated activation of the spike protein. Proc. Natl. Acad. Sci. 111, 15214–15219.

Millet, J.K., Whittaker, G.R., 2015. Host cell proteases: critical determinants of coronavirus tropism and pathogenesis. Virus Res. 202, 120–134.

Millet, J.K., Whittaker, G.R., 2018. Physiological and molecular triggers for SARS-CoV membrane fusion and entry into host cells. Virology 517, 3–8.

Ostermeier, M.K., Schertler, G.F.X., Standfuss, J., 2014. Molecular mechanisms of phosphorylation-dependent arrestin activation. Curr. Opin. Struct. Biol. 29, 143–151.

Petit, C.M., Chouljenko, V.N., Iyer, A., Colgrove, R., Farzan, M., Knipe, D.M., Kousoulas, K.G., 2007. Pamitoylation of the cysteine-rich ectodomain of the SARS-coronavirus spike protein is important for spike-mediated cell fusion. Virology 360, 264–274.

Reinke, L.M., Spoegel, M., Plegge, T., Hartleib, A., Nehlmeier, I., Gierer, S., Hoffmann, M., Hofmann-Winkler, H., Winkler, M., Pöhlmann, S., 2017. Different residues in the SARS-CoV spike protein determine cleavage and activation by the host cell protease TMPRSS2. PLoS ONE 12, e0179177.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. TIG 16, 276-277.

Robert, X., Gouet, P., 2014. Deciphering key features in protein structures with the new ENDscript server. Nucl. Acid Res. 42, W320–W324.

Sadavare, A.B., Kulkarni, R.V., 2012. A review of application of graph theory for network. Int. J. Computer Sci. Information Technol. 3, 5296–5300.

Schrödinger, L., 2015. The PyMol Molecular Graphics System, Version 1.8.

Seeger, M.A., Schiefner, A., Eicher, T., Verrey, F., Diederichs, K., Pos, K.M., 2006. Structural asymmetry of AcrB trimer suggests a peristaltic pump mechanism. Science 313, 1295–1298.

Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A., Li, F., 2020. Structural basis of receptor recognition by SARS-CoV-2. Nature 581, 221–224.

Shulla, A., Gallagher, T., 2009. Role of spike protein endodomains in regulating coronavirus entry. J. Biol. Chem. 284, 32725–32734.

Shutova, T., Klimov, V.V., Andersson, B., Samuelsson, G., 2007. A cluster of carboxylic groups in PsbO protein is involved in proton transfer from the water oxidizing complex of Photosystem II. Biochim. Biophys. Acta 1767, 434–440.

Siemers, M., Lazaratos, M., Karathanou, K., Guerra, F., Brown, L.S., Bondar, A.-N., 2019. Bridge: A graph-based algorithm to analyze dynamic H-bond networks in membrane proteins. J. Chem. Theory Comput. 15, 6781–6798.

Tai, W., He, L., Zhang, X., Pu, J.Z., Voronin, D., Jiang, S., Zhou, Y., Du, L., 2020. Characterization of the receptor-binding domain (RBD) of the 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. Cell. Mol. Immunol. https://doi.org/10.1038/s41423-41020-40400-41424.

Towler, P., Staker, B., Prasad, S.G., Menon, S., Tang, J., Parsons, T., Ryan, D., Fischer, M.S., Williams, D., Dales, N.A., Patane, M.A., Pantoliano, M.W., 2004. ACE2 X-ray structures reveal a large hinge-bending motion important for inhibitor binding and catalysis. J. Biol. Chem. 279, 17996–18007.

Venkatakrishnan, A.J., Ma, A.K., Fonseca, R., Latorraca, N.R., Kelly, B., Betz, R.M., Asawa, C., Kobilka, B.K., Dror, R.O., 2019. Diverse GPCRs exhibit conserved water networks for stabilization and activation. Proc. Natl. Acad. Sci. 116, 3288–3293.

Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T., Veesler, D., 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 180, 1–12.

Walls, A.C., Tortorici, M.A., Snijder, J., Xiong, X.-F., Bosch, B.-J., Rey, F.A., Veesler, D., 2017. Tectonic conformational changes of a coronavirus spike glycoprotein promote membrae fusion. Proc. Natl. Acad. Sci. 114, 11157–11162.

Wan, Y., Shang, J., Graham, R.L., Baric, R.S., Li, F., 2020. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. J. Virol. 94, e00127–00120.

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Yuen, K.-Y., Wang, Q., Zhou, H., Yan, J., Qi, J., 2020a. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell 181, 894–904.

Wang, X., Xu, W., Hu, G., Xia, S., Sun, Z., Liu, Z., Xie, Y., Zhang, R., Jiang, S., Lu, L., 2020b. SARS-CoV-2 infects T lymphocites through its spike-protein-mediated membrane fusion. Cell. Mol. Immunol.

West, D.B., 1996. Introduction to graph theory Upper Saddle River, NJ: Prentice Hall.

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

Xiao, X., Chakraborti, S., Dimitrov, A.S., Gramatikoff, K., Dimitrov, D.S., 2003. The SARS-CoV S glycoprotein: expression and functional characterization. Biochim. Biophys. Res. Commun. 312, 1159–1164.

Xu, Y., Lou, Z., Liu, Y., Pang, H., Tien, P., Gao, G.F., Rao, Z., 2004. Crystal structure of Severe Acute Respiratory Syndrome coronavirus spike protein fusion core. J. Biol. Chem. 279, 49414–49419.

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., Zhou, Q., 2020. Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. Science 367, 1444–1448.

Zhou, P., Yang, Y.-L., Wang, X.-G., Hu, B., Zhang, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Huang, C.-L., Chen, H.-D., Chen, J.-H., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, C.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., Shi, Z.-L., 2020. A pneumonia outbreak associated with a coronavirus of probable bat origin. Nature 579, 270–273.