frontiers in Microbiology

PERSPECTIVE
published: 17 September 2020
doi: 10.3389/fmicb.2020.02112

Flexible, Functional, and Familiar: Characteristics of SARS-CoV-2 Spike Protein Evolution

Dianita S. Saputri[1†], Songling Li[1†], Floris J. van Eerden[2], John Rozewicki[1], Zichang Xu[1], Hendra S. Ismanto[1], Ana Davila[1], Shunsuke Teraguchi[1,2], Kazutaka Katoh[1] and Daron M. Standley[1,2]*

[1] Department of Genome Informatics, Research Institute for Microbial Diseases, Osaka University, Suita, Japan,
[2] Immunology Frontier Research Center, Osaka University, Suita, Japan

The SARS-CoV-2 S protein is a major point of interaction between the virus and the human immune system. As a consequence, the S protein is not a static target but undergoes rapid molecular evolution. In order to more fully understand the selection pressure during evolution, we examined residue positions in the S protein that vary greatly across closely related viruses but are conserved in the subset of viruses that infect humans. These "evolutionarily important" residues were not distributed evenly across the S protein but were concentrated in two domains: the N-terminal domain and the receptor-binding domain, both of which play a role in host cell binding in a number of related viruses. In addition to being localized in these two domains, evolutionary importance correlated with structural flexibility and inversely correlated with distance from known or predicted host receptor-binding residues. Finally, we observed a bias in the composition of the amino acids that make up such residues toward more human-like, rather than virus-like, sequence motifs.

Keywords: flexibility, host like, molecular evolution, phylogenetics, SARS-CoV-2, spike protein, structural modeling, structure alignment

OPEN ACCESS

Edited by:
Hirokazu Kimura,
Gunma Paz University, Japan

Reviewed by:
Koo Nagasawa,
Eastern Chiba Medical Center, Japan
Masahiro Ito,
Ritsumeikan University, Japan
Yoshiyuki Suzuki,
Nagoya City University, Japan

***Correspondence:**
Daron M. Standley
standley@biken.osaka-u.ac.jp

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *03 July 2020*
Accepted: *11 August 2020*
Published: *17 September 2020*

Citation:
Saputri DS, Li S, van Eerden FJ, Rozewicki J, Xu Z, Ismanto HS, Davila A, Teraguchi S, Katoh K and Standley DM (2020) Flexible, Functional, and Familiar: Characteristics of SARS-CoV-2 Spike Protein Evolution. Front. Microbiol. 11:2112. doi: 10.3389/fmicb.2020.02112

INTRODUCTION

Over 200 viruses are known to infect humans (Woolhouse et al., 2012). Among recent human virus outbreaks, three (SARS-CoV-1, MERS-CoV, and SARS-CoV-2) have arisen from beta coronaviruses. The close interaction between pathogen and host can be a driving force for molecular evolution. This is nowhere more apparent than on the surfaces of the viruses themselves. The characteristic crown-shaped spikes, for which coronaviruses are named, enable binding to and entering host cells, and also provide camouflage from the host immune system. The ectodomain – the most outer part of the spike (S) protein – consists of two functional subunits, the receptor-binding subunit (S1) and the membrane fusion subunit (S2) (**Figure 1A**). The S1 subunits are highly variable across genera, while the S2 subunits are much more conserved. These differences reflect their distinct functions: Whereas the S1 regions engage with receptors on the surfaces of host cells, the primary function of S2 is to mediate fusion with host cell membranes. The S1 subunit is located within the N-terminus of the S protein and can be further divided into an N-terminal domain (NTD) and a C-terminal domain, which, in itself, can be divided into a receptor-binding domain (RBD) located at the apex of the protein when viewed from the side and two additional domains

connecting it to the NTD (Wang et al., 2020). In SARS-CoV-2, the RBD contains a receptor-binding motif (437–508) that contains host receptor-binding residues. The structural domains of the S protein wind around each other such that the three RBDs and NTDs constitute a nearly continuous surface at the apex of the trimeric protein (**Figure 1B**).

The targets of the S1 NTD and RBD can differ greatly among beta coronaviruses. For example, the NTD can recognize sugar derivatives in human coronavirus (HCoV)-HKU1 and HCoV-OC43, which facilitate attachment to host cells; in mouse hepatitis coronavirus (MHV), the NTD binds to the host protein carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1). Meanwhile, the RBD binds hACE2 in SARS-CoV-1 and SARS-CoV-2, but binds aminopeptidase N (APN) and dipeptidyl peptidase 4 (DPP4) in HCoV-229E and MERS-CoV, respectively (Wang et al., 2020). This large variability in binding partners suggests that NTD and RBD are sites of intense evolutionary pressure.

In order to better understand this evolutionary pressure, we estimated the evolutionary importance of residue positions in SARS-CoV-2 by comparing the amino acid diversity of each position to that of equivalent positions in closely related viruses that infect non-human hosts. We found that evolutionary importance was high in the NTD and RBD. Moreover, within these domains, residues with high evolutionary importance could be characterized by three features: they are more *flexible* (when simulated by molecular dynamics) than surrounding residues, they occur in or around known *functionally important* host – protein binding sites, and their sequences are much more self-like or *familiar* to the host immune system than other residues.

Estimating Evolutionary Importance

It is possible to infer evolutionarily important residues in the S protein by observing sites that are conserved within a given branch of the phylogenetic tree but vary among different branches. To construct a phylogenetic tree, 20 SARS-CoV-2 S protein sequences, 6 close outgroups that infect bat and pangolin, and several sequences from other lineages of beta coronavirus (SARS-CoV, MERS-CoV, and HCoV-HKU1) were collected. Amino acid sequences were aligned by MAFFT (Katoh and Standley, 2013), and a neighbor-joining (Saitou and Nei, 1987) tree was estimated to roughly visualize the phylogenetic relationship (**Figure 1C**). We subsequently estimated the sequence diversity at each position using 9,827 SARS-CoV-2 sequences (https://www.gisaid.org; after filtering out those with many ambiguous bases and/or fragmentary sequences). These sequences were compared only with the close outgroups that infect bat and pangolin. The diversity for the combination of human + outgroup and for the human group alone was compared. We defined "evolutionary importance" as the difference:

$$\text{diversity(human + outgroup)} - \text{diversity(human)}$$

assuming that this difference reflects the change in evolutionary pressure when this virus is transmitted to humans. This resulted in three levels of importance: low (0), medium (1), and high (2) as indicated in the heatmap projected onto the molecular surface of the S protein. While low- and medium-importance positions were distributed widely across the S protein surface, most of the positions with high importance were confined to two domains: the NTD and the RBD (**Figure 1D**).

For comparison, local evolutionary rate in the human-infecting lineage was estimated by (100 AA) sliding window analysis. The evolutionary rate in this lineage is proportional to the evolutionary distance between the present-day sequences infecting humans and the common ancestor of human-infecting and bat-infecting lineages. The average distance, D, between these two points was estimated, for each window, by the relative rate test (Sarich, 1969):

$$D = \frac{1}{2}\{d(h,p) + d(h,b) - d(b,p)\}$$

where, h, p, b denote human, pangolin, and bat, respectively; pairwise distance was computed using the Poisson correction, $d(,) = -\ln(1 - x/y)$, where x is the number of differences between sequences, and y is the number of sites. A high evolutionary rate in this lineage was clearly observed near hACE2-binding sites (**Figure 1E**). This observation is consistent with the site-specific diversity observed in the evolutionary importance; such sites have apparently changed radically upon transfer to humans and have been highly conserved thereafter. However, regarding the NTD region, the evolutionary rate was not estimated to be as high as the RBD in the human lineage. This local evolutionary rate analysis has three limitations: (1) it uses average rates of multiple adjacent residues, (2) it does not consider conservation within the human-infecting lineage, and (3) it cannot distinguish changes in a specific lineage from background changes in the same region. By defining evolutionary importance as we have above, we clearly observe sites that are specifically conserved in the lineage infecting humans.

Evolutionary Importance of Flexible Regions

It has been established that in SARS-CoV-1 and SARS-CoV-2, the RBD undergoes a large conformational change from the "closed" state to the "open" state upon engagement with hACE2 (Walls et al., 2020; Wrapp et al., 2020). In order to visualize flexible regions in the SARS-CoV-2 S protein, we carried out molecular dynamic simulations of the S protein in the open conformation followed by the root-mean square fluctuation (RMSF) analysis (**Figure 2A**). Not surprisingly, the most flexible parts of the protein were in loop regions. We observed that the beta sheet cores of both the S1 NTD and RBD domains were stable, as was most (but not all) of the S2 subunits. There were two exceptions with a higher RMSF: residue alanine 684 is part of the furin cleavage site (RRAR), which has been shown to be essential for infection of human lung cells (Hoffmann et al., 2020) and residues 830–840, which constitute a fusion peptide. Overall, we observed a nearly linear correlation between evolutionary importance and mean RMSF of these regions (Spearman correlation 0.30, $p < 2.2 \times 10^{-16}$) (**Figure 2B**). It is possible that flexibility in the NTD and RBD loops provide an induced-fit binding mechanism, wherein loop regions rearrange

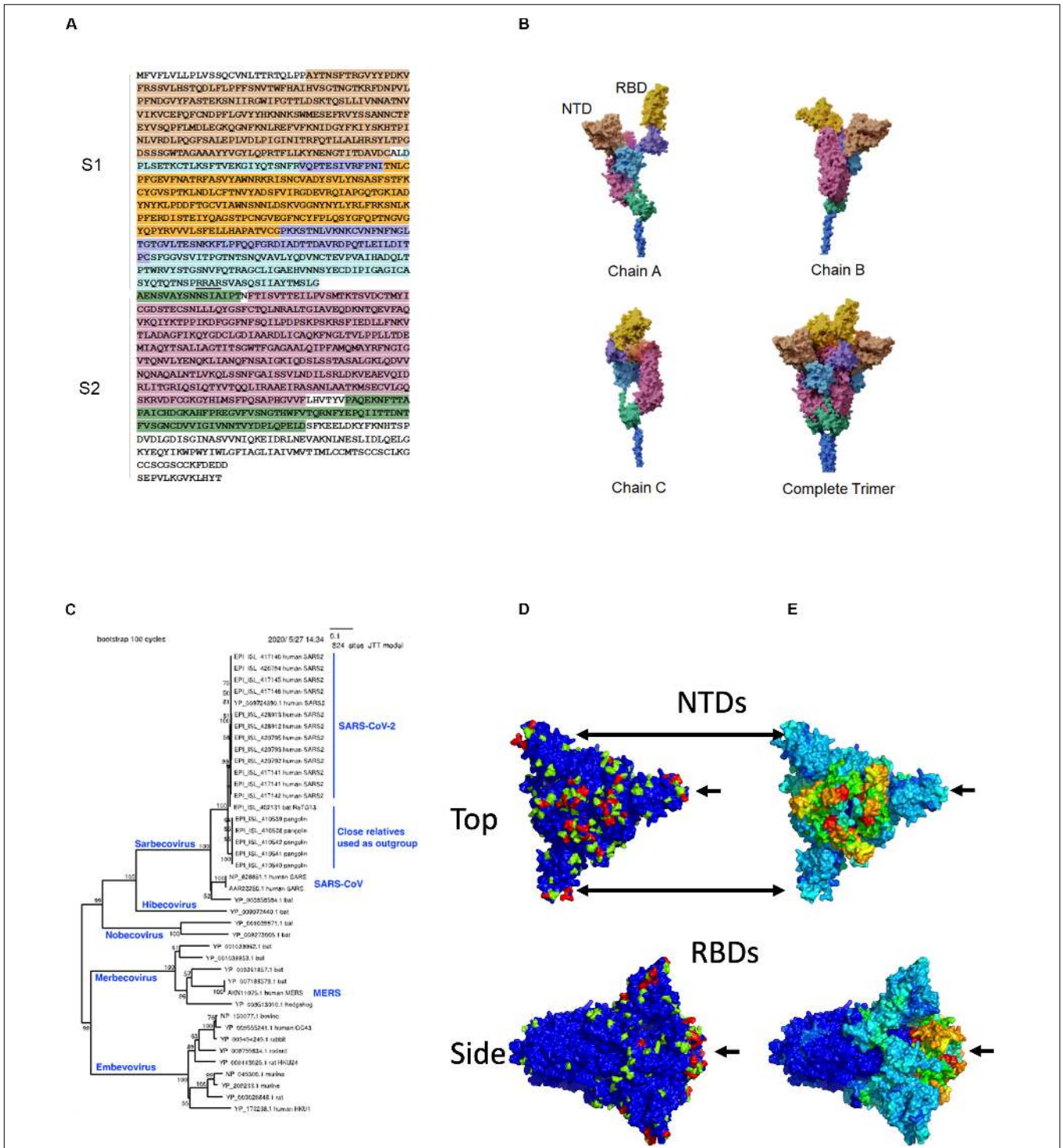

FIGURE 1 | Sequence, structure, and evolution of SARS-CoV-2 spike protein. **(A)** The sequence of the ectodomain is obtained from Uniprot with accession number P0DTC2 and is shaded according to six structural domains that correspond to those of **(B)**. The trimeric ectodomain can be divided into S1 and S2 subunits. In this figure, S1 is defined as residues 27–700 and S2 as 701–1,146. The S1 subunit can be further divided into the N-terminal domain (NTD, residues 27–291) and C-terminal domain (CTD, residues 294–700). The S1 CTD can further be divided into smaller structural domains, including the receptor-binding domain (RBD), in magenta, which is composed of residues 333–526 in this figure. The furin "RRAR" cleavage site at the end of S1 is underlined. **(B)** The three chains are interwoven to form the complete trimer. Here, the overall model was based on PDB entry 6vsb with NTD domains replaced by those of entry 5×4s. **(C)** Phylogenetic tree of closely related spike proteins from viruses that infect humans along with a representative outgroup. **(D)** Molecular surface of spike with evolutionary importance represented as a heatmap at three levels: low (blue), medium (green), and high (red). **(E)** Molecular surface of spike with evolutionary rate represented as a heatmap on a scale from lowest (blue) to highest (red).

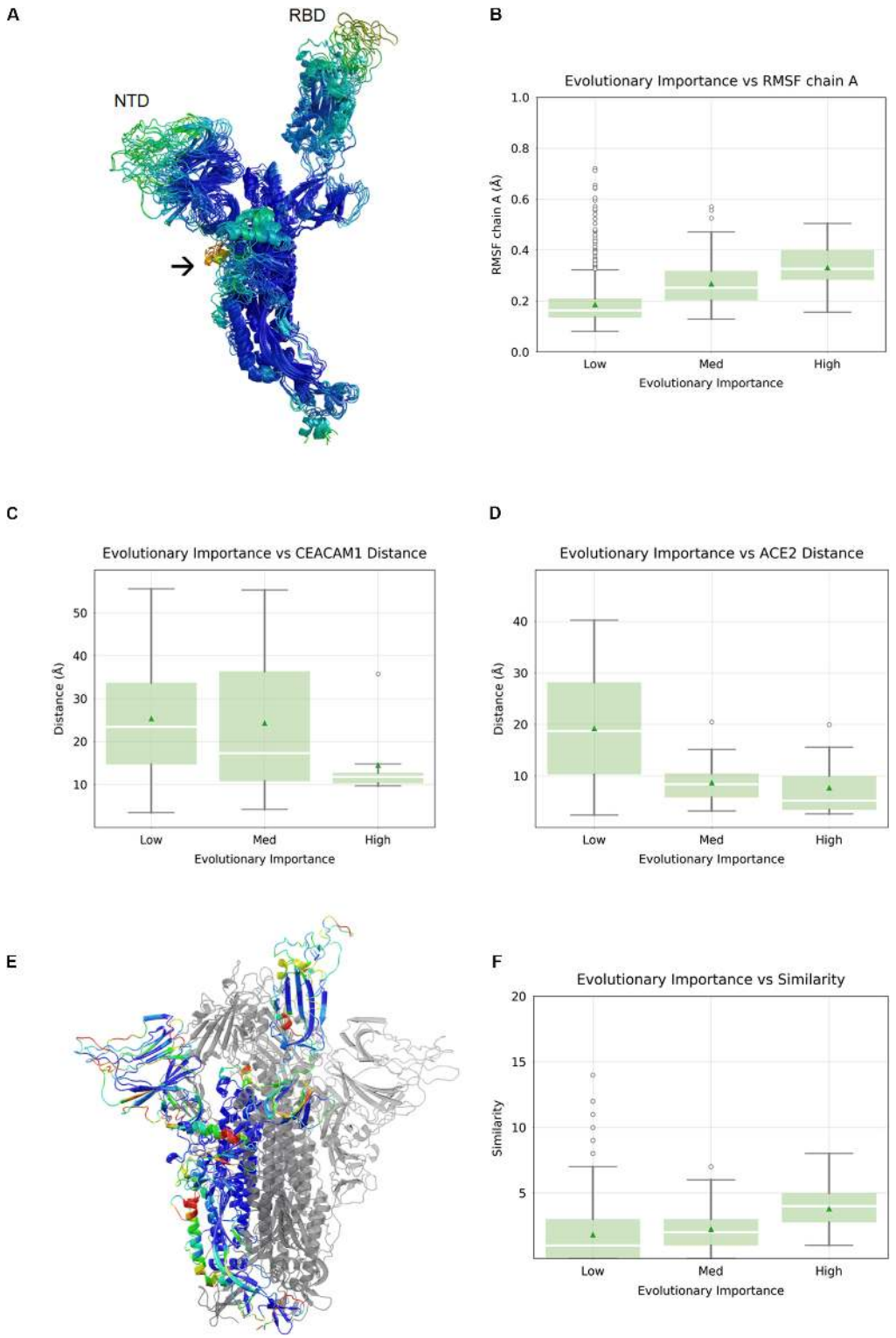

FIGURE 2 | Correlation between evolutionary importance and other properties. **(A)** The ectodomain of the spike protein colored according to the root-mean square fluctuation (RMSF), with the scale going from blue (low RMSF), via green and yellow to red (high RMSF). The NTD and RBD are labeled, and an arrow indicates the location of the fusion peptide. **(B)** Boxplot of RMSF as a function of evolutionary importance indicating that evolutionary importance is higher in flexible regions. **(C)** Boxplot of distance from carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1) interface, as inferred from superposition on MHV, for residues in the three groups of evolutionary importance. **(D)** Boxplot of distance from ACE2-binding site for residues in the three groups of evolutionary importance. **(E)** Similarity is represented as a heatmap (from low to high: blue, cyan, green, yellow, orange, red) on a single chain of the spike protein with the other two chains shown in gray. **(F)** Boxplots showing similarity to human proteins highest for evolutionary important residues.

in order to properly bind to their host receptors. To explore this idea further, we analyze the relationship between evolutionary importance and distance from host receptor-binding sites below.

Evolutionary Importance and Proximity to Functional Binding Sites

The SARS-CoV-2 RBD mediates host cell entry by binding to hACE2. While the target of the SARS-CoV-2 NTD is still unknown, the high evolutionary importance in the NTD suggests a potential binding partner. Even without knowing the target of the NTD, we can assume that the location of the binding site is roughly conserved, and the distance of each residue in the NTD from this location using the NTDs of other viruses as proxies was measured. We can, of course, perform a similar and more precise analysis in the RBD using the known RBD–ACE2 complex crystal structure (Lan et al., 2020). When we compared the evolutionary importance in SARS-CoV-2 S1-NTD with the distance to the MHV NTD–CEACAM1 interface (Peng et al., 2011), we observed a negative correlation with distance (Spearman correlation -0.19, $p = 1.8 \times 10^{-3}$) (**Figure 2C**). The fact that evolutionary importance is higher in residues located near the equivalent site suggests that SARS-CoV-2 S1-NTD may have retained host binding and that the location of the binding site is roughly conserved. We compared the evolutionary importance with the distance to the hACE2-binding site (Yan et al., 2020) and observed that the evolutionary importance was higher in residues located near the ACE2 interface, consistent with its functional importance (Spearman correlation -0.38, $p < 8.8 \times 10^{-8}$) (**Figure 2D**). Taken together, we can say that evolutionary important residues occur often in flexible loops in or near known or putative virus – host binding interfaces.

Evolutionary Importance of Host-Like Sequences

Since the outer parts of the virus are most exposed to the host immune system, we aimed to look for their similarity with human cell surface proteins, as such similarity may indicate immune evasion. We carried out local alignment of all five-residue sequence fragments with a representative set of 507 human cell surface proteins as annotated by the Cell Surface Protein Atlas (Bausch-Fluck et al., 2015). The local sequence similarity was computed for each SARS-CoV-2 residue using rigorous matching criteria for each fragment. This analysis revealed several hotspots of similarity, including the NTD and RBD (**Figure 2E**). We quantified the relationship between similarity to human cell surface proteins and evolutionary importance and found that the similarity was highest for residues with the greatest importance (Spearman correlation 0.13, $p < 7.7 \times 10^{-6}$) (**Figure 2F**).

DISCUSSION

We estimated evolutionary importance based on generally diverse residue positions that are conserved within the SARS-CoV-2. We observed that such residues were primarily restricted to two domains, the NTD and RBD, both of which have host receptor-binding functions in a number of closely related viruses. Interestingly, these "important" residues were more flexible than less important residues, suggesting that the flexibility is a characteristic of rapid molecular evolution. Moreover, the residues tended to cluster near or within known or predicted host receptor-binding sites. This is not surprising, since the Evolutionary Trace method, on which our simple definition of evolutionary importance was based, has widely been used for predicting protein – protein interactions (Wodak and Mendez, 2004).

The fact that the NTD includes many evolutionary important residues strongly hints at a role in host receptor binding. Moreover, the correlation of evolutionary importance with distance from the known CEACAM1-binding site implies that the location of the binding site might be conserved. A recent report that anti-NTD antibodies can be neutralizing (Chi et al., 2020) supports this notion. We observed that evolutionary important residues appeared to be biased toward "human-like" sequence motifs more than other residues suggesting that they may have more potential to evade the immune system through mimicking the host protein. Although the sequence data on SARS-CoV-2 is still limited, the patterns may provide clues about the identity of targeted human cell surface receptors.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

DS carried out structural analysis of domains in related viruses. SL constructed a full-length model of SARS-CoV-2 S protein. FE performed molecular dynamics simulations of S protein. JR developed software for structural alignment. ZX carried out S protein docking. HI performed sequence analysis of SARS-CoV-2 S protein. AD carried out immunogenic (epitope) prediction on S protein. ST performed statistics calculations. KK did phylogenetic analysis. DS conceived of the project and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by a Grant-in-Aid for Scientific Research by the Japan Society for the Promotion of Science (JSPS) and by the Platform Project for Supporting Drug Discovery and Life Science Research [Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)] from AMED under Grant No. 17am0101108j0001. Molecular Dynamics calculations were performed using machine time

on the TSUBAME supercomputer granted by the Tokyo Institute of Technology.

ACKNOWLEDGMENTS

We would like to thank Prof. Tatsuo Shioda, Osaka University, for his helpful suggestions and careful reading of the manuscript. We would also like to thank all members of the Systems Immunology Lab for helpful discussions regarding the preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.02112/full#supplementary-material

REFERENCES

Bausch-Fluck, D., Hofmann, A., Bock, T., Frei, A. P., Cerciello, F., Jacobs, A., et al. (2015). A mass spectrometric-derived cell surface protein atlas. *PLoS One* 10:e0121314. doi: 10.1371/journal.pone.0121314

Chi, X., Yan, R., Zhang, J., Zhang, G., Zhang, Y., Hao, M., et al. (2020). A potent neutralizing human antibody reveals the N-terminal domain of the Spike protein of SARS-CoV-2 as a site of vulnerability. *BioRxiv [Preprint]* doi: 10.1101/2020.05.08.083964

Hoffmann, M., Kleine-Weber, H., and Pohlmann, S. (2020). A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol. Cell.* 78, 779-784 e775. doi: 10.1016/j.molcel.2020.04.022

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., et al. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 581, 215–220. doi: 10.1038/s41586-020-2180-5

Peng, G., Sun, D., Rajashankar, K. R., Qian, Z., Holmes, K. V., and Li, F. (2011). Crystal structure of mouse coronavirus receptor-binding domain complexed with its murine receptor. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10696–10701. doi: 10.1073/pnas.1104306108

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Sarich, V. M. (1969). Pinniped origins and the rate of evolution of carnivore albumins. *Systematic Zoology* 18, 286–295. doi: 10.2307/2412325

Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., Mcguire, A. T., and Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 181, 281-292 e286. doi: 10.1016/j.cell.2020.02.058

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., et al. (2020). Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* 181, 894-904 e899. doi: 10.1016/j.cell.2020.03.045

Wodak, S. J., and Mendez, R. (2004). Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.* 14, 242–249. doi: 10.1016/j.sbi.2004.02.003

Woolhouse, M., Scott, F., Hudson, Z., Howey, R., and Chase-Topping, M. (2012). Human viruses: discovery and emergence. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 2864–2871. doi: 10.1098/rstb.2011.0354

Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C. L., Abiona, O., et al. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367, 1260–1263. doi: 10.1126/science.abb2507

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* 367, 1444–1448. doi: 10.1126/science.abb2762

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.