A simple method for predicting emerging SARS-CoV-2 variants using outgroups infecting non-human hosts

Kazutaka Katoh and Daron M. Standley

Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita 565-0871, Japan

April 3, 2021

Abstract

The ability to predict emerging variants of SARS-CoV-2 would be of enormous value, as it would enable proactive design of vaccines in advance of such emergence. Based on molecular evolutionary analysis of S protein, we found a significant correspondence in the location of amino acid substitutions between SARS-CoV-2 variants recently emerging and their relatives that infected bat and pangolin before the pandemic. This observation suggests that a limited number of sites in this protein are repeatedly substituted in independent lineages of this group of viruses. It follows, therefore, that the sites of future emerging mutations in SARS-CoV-2 can be predicted by analyzing their relatives (outgroups) that have infected non-human hosts. We discuss a possible evolutionary mechanism behind these substitutions and provide a list of frequently substituted sites that potentially include future emerging variants in SARS-CoV-2.

Introduction

In December 2020, three SARS-CoV-2 variants emerged with increased infectivity from England, South Africa and Brazil. The fact that certain mutations in the spike (S) protein had occurred independently prompted us to reexamine our September 2020 study of the evolution

of this protein [1]. In our original study, we characterized the *Importance* of each residue position in the S protein by comparing its diversity in SARS-CoV-2 with that in relatives (outgroups) that infected bats or pangolins by using a simple equation:

$$Importance = diversity(\text{SARS-CoV-2 + outgroup}) - diversity(\text{SARS-CoV-2}),$$

where $diversity(x)$ is defined as the number of different amino acids observed at the site in question in virus group x. This equation, which was meant to be descriptive rather than predictive, identified twenty positions of high *Importance*. We were thus surprised to find that, of these twenty positions, four were characteristic of the above emerging variants: Histidine 69, Valine 70, Glutamine 484 and Asparagine 501. These sites coincide with four out of the five residues (69, 70, 417, 484, 501) that are observed multiple times in the three emerging lineages or the lineage transmitted between human and mink [5]. We reanalyzed the underlying sequence data and found that the *Importance* values of these sites were determined primarily by *diversity*(outgroup), rather than *diversity*(SARS-CoV-2). In hindsight, this is somewhat expected, as the latter term was close to unity at the time when we performed the analysis (i.e., before the emergence of new variants).

Theory

A natural question, then, is why a limited set of sites with high diversity in outgroups have also recently mutated in SARS-CoV-2. One possible explanation is that these sites are rapidly evolving under low functional constraints (i.e. neutral evolution) and thus frequently substituted in multiple lineages. This explanation is contradicted by the fact that the sites in question are estimated to be under positive selection (nonsynonymous substitutions more frequent than synonymous substitutions) using Bayes Empirical Bayes analysis [6] applied to closely related outgroups (see the 4th column in https://mafft.cbrc.jp/alignment/pub/sarscov2/fulllist.tsv), although the estimation is sensitive to sequence selection. A more likely explanation, then, is that the sites are involved in either infection of host cells, evasion of host immunity, or both. Indeed, Glutamine 484 and Asparagine 501 are structurally close to the interface with the host cell receptor ACE2, which, in turn, is targeted by neutralizing antibodies. Histidine 69 and

Valine 70, on the other hand, are far from the ACE2 binding site but proximal to a recently-reported epitope for infection-enhancing antibodies [2, 3]. An overlapping region has been reported to bind sialic acids [4]. Modification of these processes could thus enable the virus to escape from the host's immune system, albeit temporarily, as the change will inevitably be counteracted by a shift in the antibody repertoire of the host, resulting in an effective "arms race". In this scenario, the sites with higher diversity imply direct or indirect host-pathogen interactions and are thus in a constant state of flux.

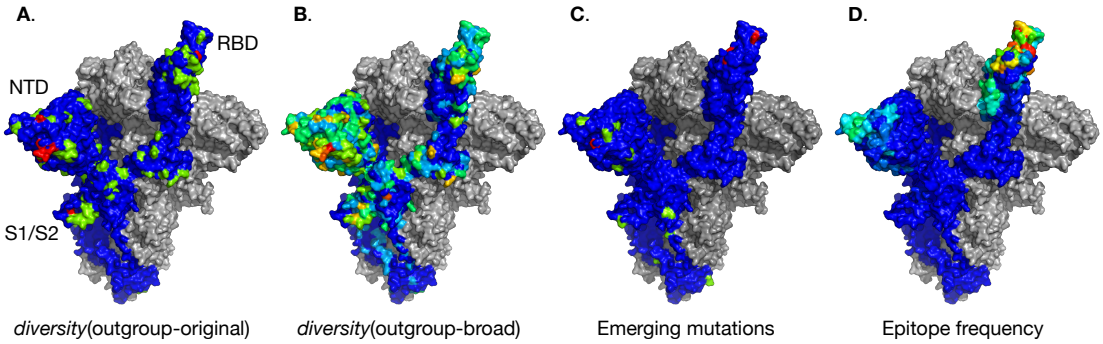

Figure 1: Diversity of the the S protein. For clarity, a single S protein is shown in the context of a spike trimer. **A**, **B**. The *diversity*(outgroup) was computed using the original and broad definitions of outgroup and is shown as gradient between low diversity (blue) and high diversity (red). **C**. Emerging mutations are colored based on their frequency of appearance: 1 (green); 2 (red). **D**. Epitope frequency, the number of antibodies that contact each residue (<6Å), was counted based on currently available Protein DataBank (PDB) entries of S protein-antibody complexes listed in https://mafft.cbrc.jp/alignment/pub/sarscov2/epitopefrequency.txt. 0 (blue); 15 (red). This value is not expected to represent all spike-targeting antibodies.

Results and Discussion

According to the latter interpretation, it is possible that positions of mutations in future emerging variants can be predicted simply by identifying sites with high diversity in outgroups, where such an arms race has been played out longer than between SARS-CoV-2 and humans. Because of their potential importance in the design of vaccines against future emerging variants, we list residue positions with the highest *diversity*(outgroup) in Table 1, where we have considered

two definitions of outgroups: one that is identical to that used in our original analysis in which 6 sequences were used and a broader definition (11 sequences) to increase the amount of data used in the calculation. Both datasets are available at https://mafft.cbrc.jp/alignment/pub/sarscov2/. When viewed as a heatmap on the spike molecular surface, it is apparent that the residue positions with high diversity are not evenly distributed, but form clusters in the N terminal domain (NTD), receptor binding domain (RBD) and S1/S2 cleavage site (Fig. 1). We note that the correspondence between the positions of emerging mutations and those with high *diversity*(outgroup) is significant by Fisher's Exact Test regardless whether the original outgroup (Table 2A) or the broad outgroup (Table 2B) is used. Both the five positions (69, 70, 417, 484 and 501) observed in multiple emerging variants and the union of all variant positions were considered in calculating p. Mutations in the five positions are expected to continue to spread in humans as they are likely to affect interactions with host factors. The proposed simple method is suitable to predict such sites because they appear to be under positive selection in independent lineages. Consistently, the p values for the positions that are sampled independently in multiple lineages are relatively low in Table 2A and B.

To anticipate new variants of SARS-CoV-2 as early as possible, a straightforward strategy would be to intensively collect a large amount of sequence data from human-infecting lineages. Our observation above leads to a complementary strategy: prepare against new variants in advance by decoding the long history of host-pathogen interactions recorded in the outgroup sequences infecting non-human hosts. Unfortunately, currently efforts have focused almost exclusively on the former strategy and available outgroup sequences are limited. If richer sequence data of close relatives of SARS-CoV-2 infecting bat, pangolin and other possible hosts become available, we can, in principle, gain an advantage in the arms race with this virus.

Conflcts of interest

None declared.

References

[1] Saputri, D. S. et al. Flexible, Functional, and Familiar: Characteristics of SARS-CoV-2 Spike Protein Evolution. *Front Microbiol* 11: 2112 doi:10.3389/fmicb.2020.02112 (2020)

[2] Li, D. et al., The functions of SARS-CoV-2 neutralizing and infection-enhancing antibodies in vitro and in mice and nonhuman primates. *bioRxiv* 2020.2012.2031.424729, doi:10.1101/2020.12.31.424729 (2021)

[3] Liu, Y. et al., An infectivity-enhancing site on the SARS-CoV-2 spike protein is targeted by COVID-19 patient antibodies. *bioRxiv* 2020.2012.2018.423358, doi:10.1101/2020.12.18.423358 (2020)

[4] Baker, A. N. et al., The SARS-COV-2 Spike Protein Binds Sialic Acids and Enables Rapid Detection in a Lateral Flow Point of Care Diagnostic Device. *ACS Cent. Sci.* 6:2046-2052, https://doi.org/10.1021/acscentsci.0c00855 (2020)

[5] Lassaunière, R. et al. Working paper on SARS-CoV-2 spike mutations arising in Danish mink, their 2 spread to humans and neutralization data. `https://files.ssi.dk/Mink-cluster-5-short-report_AFO2` (2021)

[6] Yang, Z. et al. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Mol Biol Evol* 22:1107-1118, doi:10.1093/molbev/msi097 (2005)

Residue	AA	Orig	Broad	Epitope	ACE2	SialicAcid	Cleavage	Emerge
7	L	2	5	0				
12	S	1	5	0				
23	Q	2 *	4	0		1		
27	A	2	5	0				
33	T	1	5	2				
69	H	3 *	5	2		1		2
70	V	3 *	5	2		1		2
71	S	3 *	5	1				
72	G	2 *	3	1				
73	T	3 *	5	1				
74	N	3 *	6	1				
75	G	1	5	1				
76	T	3 *	7	2				
137	N	2	5	0				
147	K	2	5	4				
213	V	1	5	2				
218	Q	2 *	5	2				
224	E	1	5	0				
253	D	3 *	5	2				
255	S	2 *	5	1				
256	S	1 *	2	0				
272	P	2	5	0				
417	K	2 *	2	8	1			2
439	N	2 *	4	0				
440	N	2	5	0				
441	L	2 *	3	3				
444	K	2 *	3	0				
445	V	2 *	3	3				
449	Y	3	3	2	1			
450	N	2 *	5	11				
501	N	2 *	4	8	1			2
529	K	2 *	2	0				
532	N	2	5	0				
554	E	2 *	4	0				
556	N	2	5	0				
640	S	2	6	0				
677	Q	2 *	4	0				
679	N	2 *	5	0				
680	S	2 *	4	0				
684	A	3	1	0			1	
688	A	2	5	0				
689	S	2 *	3	0				

Table 1: High diversity residues. The most diverse residue positions are listed, along with several annotations. Orig, *diversity*(outgroup-original); *, (nonsynonymous substitutions) / (synonymous substitutions) > 1 in outgroup-original; Broad, *diversity*(outgroup-broad); Epitope, epitope frequency. See the caption of Figure 1. ACE2, residue is within 6Å of ACE2 in PDB entry 7DF4; SialicAcid, reported sialic acid binding residue [4]; Cleavage, known protease cleavage site; Emerge, frequency of mutations in emerging variants or human/mink transmitted lineage. See https://mafft.cbrc.jp/alignment/pub/sarscov2/fulllist.tsv for a full list.

A. *diversity*(outgroup-original)	1	2	3					
No. sites	1179	85	9					
No. mutated sites in emerging variants	15	4	2	$p = 0.00099$				
No. mutated sites in multiple emerging variants	0	3	2	$p = 1.5 \times 10^{-7}$				

B. *diversity*(outgroup-broad)	1	2	3	4	5	6	7	
No. sites	822	242	126	57	23	2	1	
No. mutated sites in emerging variants	8	3	6	2	2	0	0	$p = 0.0057$
No. mutated sites in multiple emerging variants	0	1	1	1	2	0	0	$p = 0.00046$

Table 2: Correspondence between *diversity*(outgroup) and emerging mutations.