Current Topics in Biochemical Research
Vol. 21, 2020

Original Article

The extraordinarily conserved spike glycoprotein domains of SARS-CoV-2 and the potential for COVID-19 vaccine protection against future SARS outbreaks

Babu V. Bassa* and Rao M. Uppu

Department of Environmental Toxicology, College of Sciences and Engineering, 108 Fisher Hall, James L. Hunt Street, Southern University and A&M College, Baton Rouge, LA 70813, USA.

ABSTRACT

The pathogen that causes COVID-19 has been designated as severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2 or SARS-2). It is widely speculated that SARS-CoV-2 has originated from the bat strains of coronaviruses. The spike glycoprotein (SGP) is the main component of the coronavirus envelope which plays important role in the entry of the virus into mammalian cells. In our earlier reports we have eluded to the presence of large peptide domains in the SGP that are highly conserved in the *Sarbecovirus* group. Many of the severe acute respiratory syndrome (SARS)–causing coronaviruses belong to this group. We now show that three large peptide domains of SGP are present in un-substituted forms in close to 100% of the *Sarbecovirus* strains. Most of the recently introduced mRNA vaccines employ the SGP as the target protein. Although the receptor binding domain of the SGP has evolved out and deviated significantly from that of the SARS and SARS-like viruses, the mRNA vaccines still have the potential to offer protection against the present as well as future SARS pandemics due to the presence of these three extraordinarily conserved polypeptide domains in the SGP molecule.

KEYWORDS: coronavirus, SARS-CoV-2, COVID-19, spike protein.

*Corresponding author: bassa_babu@subr.edu

INTRODUCTION

The recent outbreak of severe acute respiratory syndrome occurred in the Wuhan city, China in December 2019. The pathogen responsible for the syndrome was later identified as a coronavirus strain and it was designated as severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2 or SARS-2). The World Health Organization (WHO) declared the outbreak as a pandemic and named the disease as Coronavirus Disease-19 (COVID-19) [1]. Coronaviruses that cause severe acute respiratory syndrome in humans and animals are grouped under the subgenus *Sarbecovirus* [2]. Apart from SARS-CoV-2, included in this group are SARS-CoV, the pathogen from 2003 SARS outbreak, and several bat, civet, and pangolin strains of coronaviruses. Whereas the source of SARS-CoV-2 is still not known definitively, a bat origin is widely speculated. Spike glycoprotein (SGP) is an important structural protein of the coronavirus envelope and the first protein to contact the cellular plasma membrane in the process of entry of the virus into the host cells. It is for this reason that most vaccines are designed to target the SGP [3]. In our earlier investigations we tentatively identified a 111 amino acid (111-aa) and an 11 amino acid (11-aa) SGP domain to have been highly conserved in the *Sarbecovirus* sub-genus [4-8]. Our later surveys have revealed one more highly conserved peptide domain in the SGP that is 82 amino acids long (82-aa). The receptor binding domain (RBD), the proteolytic

cleavage site (PCS), the heptad-1-repeating-domain-1 (HR1), and the heptad-2-repeating-domain (HR-2) are the main functional domains of the SGP molecule. In the present study we evaluated the conservation of 11-aa, 82-aa, and 111-aa in the *Sarbecovirus* group, in order to obtain quantitative information. Conservation was determined as the number of times each motif appeared in un-substituted forms (identical permutations) in the analyzed number of SGP sequences. We found that these three peptide domains were conserved in close to 100% of the GenBank annotated *Sarbecovirus* strains. The mRNA vaccines introduced recently are designed to express and present the full-length SGP to the immune system of the human host. Owing to their size the 111-aa and 82-aa are likely to contribute significantly to the constellation of the spike protein antigenic sites, and therefore these vaccines are likely to offer protection against the present and future SARS outbreaks, even though the RBD of the SARS-CoV-2 is not similarly conserved.

MATERIALS AND METHODS

All the SGP sequences analyzed in this study were obtained from GenBank. The Basic Local Alignment Search Tool (BLAST) [9], and *Compare* [4] were used in the analysis of the sequences. As described previously, *Compare* identifies and presents the profile of common amino permutations occurring between the queried pair of sequences.

Determination of conservation

For the BLAST analysis the whole motif of 11-aa, 82-aa, or 111-aa was queried. The sequences resulting from the query were subjected to testing by *Compare* and taking guidance from the BLAST scores the number of times that the motifs appeared in the analyzed sequences was determined and respective percentages were calculated. Owing to the presence of over-whelming number of human SARS-CoV-2 genomes in GenBank deposits, at the present, locating the bat and other animal strains of *Sarbecovirus* proved to be a difficult task. For this reason, the animal (bat, civet, pangolin, and mink) sequences were accessed separately and analyzed by *Compare*. Please note that sequence homology was not used to determine the degree of conservation. Instead, the number of times each motif occurred in an un-substituted form in each case was expressed as a percentage of the number of samples analyzed (Table 1). This method was adopted in view of the unprecedented degree of conservation of these motifs among the coronaviruses of the *Sarbecovirus* group. The values presented are therefore applicable only to the comparisons made in this report.

RESULTS AND DISCUSSION

In the SGP as shown in Figure 1, the 111-aa and the 82-aa form parts of the fusion fork. The fusion fork is formed by HR1 and HR2 following the binding of the virus to the cellular plasma membrane through the ACE2 receptor. The fusion fork, which is a hydrophobic tool-like structure, in turn facilitates the membrane fusion and entry of SARS-Co-V-2 into mammalian cells. Whereas a portion of the 111-aa extends into HR-1, whole of HR2 is embedded in the 82-aa (Figure 1).

The exact significance of 11-aa of the SGP is not clear, but as reported by us earlier, the motif is densely hydrophobic in nature and the well-recognized D614G mutation is embedded in this motif. The specific locations of these three peptide motifs relative to other functional domains of the SGP molecule are diagrammatically depicted in Figure 1. The present context is the significance of 11-aa, 111-aa and 82-aa in the possible extension of COVID-19 vaccine protection against probable future SARS pandemics. Therefore, following our initial identification, analysis of how these three fragments are retained by the *Sarbecovirus* strains is very relevant. In this study all the three motifs were separately queried through BLAST and the BLAST scores corresponding to un-substituted forms (identical amino acid permutations with the motif) were identified, and used in the counting of the number of times each motif appeared as an identical permutation. Percentages as presented in Table 1 are calculated using these numbers. In speculating the possible extension of protection from COVID-19 vaccines to the future SARS pandemics, the points to remember are (1) the RBD of SARS-CoV-2 evolved out to distinguish itself from the RBD of all known SARS and

Table 1. Conservation of peptide motifs of SGP in the *Sarbecovirus* sub-genus.

Sarbecovirus sub-genus		Peptide motif (% Conservation)[a]		
		11-aa	82-aa	111-aa
SARS-1[b]	Human	100 (71)	83 (67)	98 (67)
	Animal[c]	100 (58)	68 (44)*	70 (44)
SARS-2	Human	100 (1906)	100 (4909)**	100 (4901)[e]
	Animal[d]	100 (36)	97 (36)	94 (36)

Number of samples analyzed for each data point is indicated in parentheses.

[a]Conservation is computed as percentage of the times a motif occurs in an un-substituted form in the number of samples analyzed. [b]Includes Exo N1 and Witc. M strains. [c]Bat and Civet. [d]Pangolin, Bat-RaTG13 and Mink. [e]Please note that the S982A mutation of the British strain B.1.1.7 is located in the 111-aa although it is not detectable in the BLAST search.

*All strains available in the GenBank records were included. Margin of error 6.4 at confidence level 95%.

**Margin of error 1.4 at 95% confidence level.

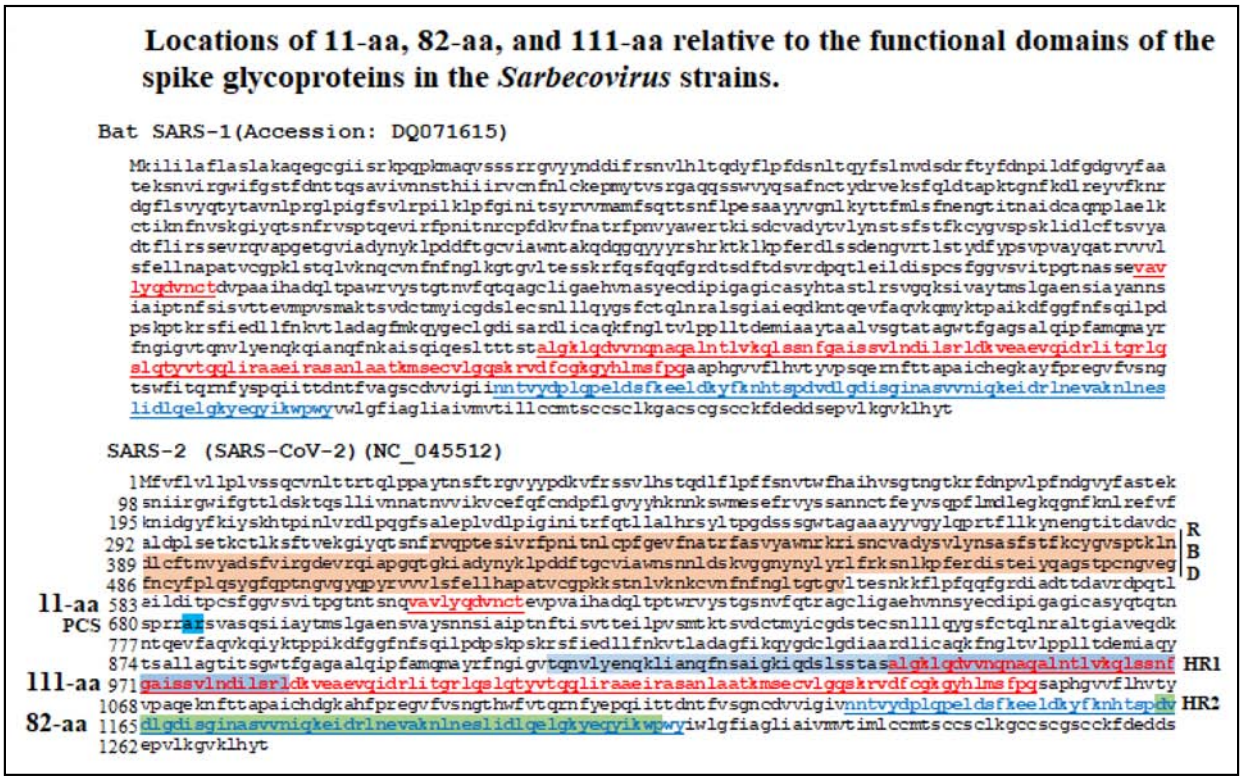

Figure 1. The locations of 11-aa (red alphabet underlined), 82-aa (magenta alphabet, underlined), and 111-aa (red alphabet underlined) are depicted relative to the functional domains, namely, the receptor binding domain (RBD) (orange shade), the proteolytic cleavage site (PCS) (blue shade), the heptad-repeating-1-domain (HR1) (light blue shade), the heptad-repeating-1-domain (HR2) (light green shade) locations in the spike glycoproteins of the *Sarbecovirus* group. The spike glycoprotein sequences of representative bat SARS-1 and human SARS-CoV-2 are shown here. Please note that information on the locations of RBD, PCS, HR1, and HR2 was obtained from published literature [10].

SARS-like animal strains of coronavirus (2) the dramatic conservation of the three polypeptide motifs points out to the critical role played by these polypeptides in the survival of the virus (3) the 111-aa and 82-aa form parts of the fusion fork that plays a critical role in the entry of the virus into mammalian cells (4) although the significance of 11-aa is not yet clear, like the other two motifs, 11-aa is also conserved extraordinarily, in the *Sarbecovirus* group and as reported by us earlier the 11-aa forms a local concentration of hydrophobic amino acids in the larger milieu of the SGP sequence. It is also important to note that the D614G mutation which appeared early in the course of the pandemic and completely displaced the original Wuhan strain, globally is embedded in11-aa.

CONCLUSION

It is widely speculated that the pathogens SARS-CoV and SARS-CoV-2 have originated in bats. The spike glycoprotein is a structural component of the coronavirus envelope that plays an important role in the initial binding and subsequent entry of the virus into mammalian cells. For the same reason the spike glycoprotein is the target protein of almost all COVID-19 vaccines. In this study we have identified two large polypeptide domains and one shorter polypeptide domain that are commonly present in the spike glycoproteins of human SARS and animal SARS-like strains of coronaviruses. Owing to their sizes together these peptides likely contribute to a significant number of the spike protein antigenic sites. Therefore, the recently introduced COVID-19 vaccines have the potential to offer protection against any future SARS outbreaks originating in animals especially bats.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

REFERENCES

1. Timeline of WHO's response to COVID-19. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline
2. National Center for Biotechnology Information (NCBI). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1988 – [cited 2020 January, 2021]. Available from: https://www.ncbi.nlm.nih.gov/
3. Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D, Spector, S. A., Rouphael, N., Creech, C. B., McGettigan, J., Khetan, S., Segall, N., Solis, J., Brosz, A., Fierro, C., Schwartz, H., Neuzil, K., Corey, L., Gilbert, P., Janes, H., Follmann, D., Marovich, M., Mascola, J., Polakowski, L., Ledgerwood, J., Graham, B. S., Bennett, H., Pajon, R., Knightly, C., Leav, B., Deng, W., Zhou, H., Han, S., Ivarsson, M., Miller, J. and Zaks, T. 2020, New Engl. Med., doi:10.1056/NEJMoa2035389.
4. Bassa, B. V. and Brown, O. R. 2020, Front. Biosci. (Landmark Edn.), 25, 1894-1900. doi:10.2741/4883.
5. Bassa, B. and Olen, B. 2020, Preprints 2020070488; doi:0.20944/preprints202007.0488.v1.
6. Bassa, B. V. and Uppu, R. M. 2020, Science eLetters. https://science.sciencemag.org/content/368/6493/829/tab-e-letters
7. Bassa, B. V. and Uppu, R. M. 2020, Science eLetters. https://science.sciencemag.org/content/368/6492/681/tab-e-letters
8. Bassa, B. V. and Uppu, R. M. 2020, Science eLetters. https://science.sciencemag.org/content/368/6491/561/tab-e-letters
9. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
10. Xia, S., Liu, M., Wang, C., Xu, W., Lan, Q., Feng, S., Qi, F., Bao, L., Du, L., Liu, S., Qin, C., Sun, F., Shi, Z., Zhu, Y., Jiang, S. and Lu, L. 2020, Cell Res., 30, 343-355.