Received: 29 August 2020 | Accepted: 27 December 2020

DOI: 10.1002/prot.26042

RESEARCH ARTICLE

PROTEINS WILEY

Human SARS CoV-2 spike protein mutations

Lalitha Guruprasad

School of Chemistry, University of Hyderabad, Hyderabad, India

Correspondence
Lalitha Guruprasad, School of Chemistry, University of Hyderabad, Hyderabad 500046, Telangana, India.
Email: lalitha.guruprasad@uohyd.ac.in

Funding information
School of Chemistry, University of Hyderabad

Abstract

The human spike protein sequences from Asia, Africa, Europe, North America, South America, and Oceania were analyzed by comparing with the reference severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) protein sequence from Wuhan-Hu-1, China. Out of 10333 spike protein sequences analyzed, 8155 proteins comprised one or more mutations. A total of 9654 mutations were observed that correspond to 400 distinct mutation sites. The receptor binding domain (RBD) which is involved in the interactions with human angiotensin-converting enzyme-2 (ACE-2) receptor and causes infection leading to the COVID-19 disease comprised 44 mutations that included residues within 3.2 Å interacting distance from the ACE-2 receptor. The mutations observed in the spike proteins are discussed in the context of their distribution according to the geographical locations, mutation sites, mutation types, distribution of the number of mutations at the mutation sites and mutations at the glycosylation sites. The density of mutations in different regions of the spike protein sequence and location of the mutations in protein three-dimensional structure corresponding to the RBD are discussed. The mutations identified in the present work are important considerations for antibody, vaccine, and drug development.

KEYWORDS
mutations, receptor binding domain, SARS-CoV-2, sequence and structural mapping, spike proteins

1 | INTRODUCTION

The epicenter of the ongoing COVID-19 pandemic caused by the human severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) was first identified in the city of Wuhan-Hubei-1 province, China during mid-December 2019.[1] Since then, the disease has spread rapidly and has affected millions of people in the populations worldwide leading to more than 778102 deaths until date (https://www.worldometers.info/coronavirus/). The SARS-CoV-2 belongs to the family of Coronaviridae, subfamily of Orthocoronavirinae and genera of β-CoV (https://www.ncbi.nlm.nih.gov/taxonomy/694009). The spread of the disease is attributed to contact via respiratory droplets either due to coughing or sneezing or through surface contact. As on date, no vaccines or specific drugs are available to treat the disease, however, there is enormous ongoing efforts worldwide in this direction.

The SARS-CoV-2 is a spherical shaped virion with a positive-stranded RNA viral genome of size 30 kb that is translated into structural and non-structural proteins. The spike glycoprotein is a homotrimer present on the surface of the coronavirus that plays a vital role in recognition of human host cell surface receptor angiotensin-converting enzyme-2 (ACE-2).[2] This recognition is required for fusion of viral and host cellular membranes for transfer of the viral nucleocapsid into the host cells. SARS-CoV-2 is reported to have originated in bats[3] and subsequently transmitted to humans via pangolins as intermediate host species.[4-6] In order to be able to jump species and infect a new mammalian host, the viral genome undergoes several mutations in the spike proteins.

The spike protein comprises an N-terminal S1 subunit and a C-terminal membrane proximal S2 subunit. The S1 subunit consists $S1^A$, $S1^B$, $S1^C$ and $S1^D$ domains. The $S1^A$ domain, referred as N-terminal domain (NTD), recognizes carbohydrate, such as, sialic acid

required for attachment of the virus to host cell surface. The $S1^B$ domain, referred as receptor-binding domain (RBD) of the SARS-CoV-2 spike protein interacts with the human ACE-2 receptor.[2,7] The structural elements within the S2 subunit comprises three long α-helices, multiple α-helical segments, extended twisted β-sheets, membrane spanning α-helix, and an intracellular cysteine rich segment. The PRRA sequence motif located between the S1 and S2 subunits in SARS-CoV-2 presents a furin-cleavage site.[8] In the S2 subunit, a second proteolytic cleavage site S2′, upstream of the fusion peptide is present. Both these cleavage sites participate in the viral entry into host cells.

In a study on the infectivity and reactivity to a panel of neutralizing antibodies and sera from convalescent patients,[9] mutations and glycosylation site modifications have been reported in human SARS-CoV-2 spike proteins. Few mutations have been reported in the spike glycoprotein.[10] The D614G mutation is reported to be relatively more common[11-14] and is known to increase the efficiency of causing infection.[2] Mutation sites for spike proteins from some of the SARS-CoV-2 Indian isolates have been mapped on to protein three-dimensional structure.[15-17]

In light of the large number of SARS-CoV-2 spike protein sequences currently available in the NCBI virus database, I intended to carry out an exhaustive analysis, in order to understand the current scenario of mutations in the spike proteins. This study informs us of all the mutations present in the human SARS-CoV-2 spike proteins relative to Wuhan-Hu-1 reference sequence from China, according to their geographical locations, positions of the mutation sites, distribution of the number of mutations at the mutation sites, the different mutation types observed so far, mutations at glycosylation sites, occurrence of multiple mutations in a single spike protein and mutations within the RBD close to the host-cell ACE-2 receptor interactions. This study has implications from the perspective of vaccine, antibody, and drug design.

2 | METHODS

The SARS-CoV-2 spike protein sequences were obtained in the FASTA format from the NCBI virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/). The multiple sequence alignment[18] of the proteins was achieved using NGPhylogeny server[19] (http://www.NGPhylogeny.fr). The human SARS-CoV-2 spike protein sequence from Wuhan-Hu-1, China (NCBI accession code: YP_009724390.1)[1] was used as reference sequence to examine the mutations. The identification of mutations in the spike proteins and further analyses was carried out using the software suite of programs developed by ABREAST (https://www.abreast.in). The mutations were analyzed according to their presence within different regions of the spike protein sequence. The locations of mutations in RBD ($S1^B$ domain) of the spike protein were mapped on to the three-dimensional crystal structure available in the Protein Data Bank (PDB)[20] (PDB code: 6LZG).[7] The molecular visualization was carried out using PyMol.[21]

3 | RESULTS AND DISCUSSIONS

3.1 | Worldwide distribution of mutations in spike protein

The NCBI virus database contained 10333 human SARS-CoV-2 spike proteins that were analyzed. These represented spike proteins from Africa (103), Asia (996), Europe (370), North America (8268), South America (29), and Oceania (567). The length of protein sequences ranged between 1250 and 1273 amino acid residues. One or more amino acid mutations were observed in 8155 proteins. A total of 9654 mutations were observed corresponding to 400 distinct mutation sites. The distribution of total number of mutations in spike proteins analyzed from different geographical locations is shown in Table 1. The large number of spike protein sequences available has provided an opportunity to evaluate the wide-spectrum of mutations observed across the continents of the world. The mutations analyzed are a result of human-to-human transmission of the virus since January 2020.

3.2 | Distribution of the number of mutations at mutation sites in the spike protein sequence

Nearly one-third of the spike protein sequence is associated with mutations. The list of the mutation sites along with total number of mutations observed at individual mutation sites is shown in Table S1. The top 10 mutation sites according to the total number of occurrences were; D614(7859), L5(109), L54(105), P1263(61), P681(51), S477(47), T859(30), S221(28), V483(28), A845(24).

3.3 | Mutation density in different regions of the spike protein sequence

The distribution of the mutations within different regions of the spike protein is shown in Table 2. Mutations are distributed in almost all regions of the protein. The $S1^D$ domain (594-674) that comprises the D614G mutation is the most predominant and is observed in 7859 of the 7915 mutations in this region of the spike protein. Our analysis is in agreement with the high frequency of D614G mutation in the spike

TABLE 1 Geographical distribution of human SARS-CoV-2 spike proteins and their associated number of mutations

Continent	Number of spike proteins	Number of mutations
Africa	103	121
Asia	996	1169
Europe	370	360
North America	8268	7453
South America	29	26
Oceania	567	525

TABLE 2 Distribution of mutations in the different regions of human SARS-CoV-2 spike proteins

Regions	Total number of mutations	Number of distinct mutation types
$S1^A$ domain (1-302)	759	196
$S1^A$-$S1^B$ linker (303-332)	31	11
$S1^B$ domain (333-527)	204	52
$S1^B$ - $S1^C$ linker (528-533)	1	1
$S1^C$ domain (534-589)	75	15
$S1^C$ - $S1^D$ linker (590-593)	0	0
$S1^D$ domain (594-674)	7915	34
Protease cleavage site (675-692)	126	35
S1-S2 subunits linker (693-710)	13	7
Central β-strand (711-737)	13	7
Downward helix (738-782)	24	18
S2' cleavage site (783-815)	27	13
Fusion peptide (816-828)	6	3
Connecting region (829-911)	111	26
Heptad repeat region (912-983)	73	18
Central helix (984-1034)	8	6
β-hairpin (1035-1068)	7	4
β-sheet domain (1069-1133)	79	27
Heptad repeat region (1134-1213)	65	22
Transmembrane region (1214-1236)	28	7
Cytoplasmic region (1237-1273)	89	15

proteins or the more common occurrence as reported previously.[11-14] The mutation density evaluated as a function of the number of mutations observed over the sequence length corresponding to different regions in the spike protein is shown in Figure 1. The protease cleavage site (between residues 675 and 692) in the spike protein is associated with the maximum mutation density. The mutations at this site in the spike protein may be of advantage for the virus to undergo proteolytic cleavage by a large number of host enzymes in its evolution. Further, the NTD ($S1^A$ domain) is another region where mutations have accumulated relatively more in number compared with rest of the spike protein.

3.4 | Human SARS-CoV-2 spike protein mutation sites and mutation types

More than one mutation type can be found at the same position in the spike protein sequence. For instance, at position 88 the amino acid residue D is observed to be mutated either to N, E, Y, or A. At position 675, the amino acid residue Q is mutated either to R, H, K or is deleted among the spike proteins. The geographical location-wise distribution of the mutation sites and mutation types is shown in Table 3. Accordingly, the total number of mutation sites observed were; North America (300), South America (4), Europe (42), Africa (16), Asia (166), and Oceania (51) and the total number of mutation types observed were; North America (350), South America (4), Europe (43), Africa (16), Asia (181), and Oceania (51). It is clear from our study that the human SARS-CoV-2 spike protein undergoes mutations at multiple sites and there can be more than one mutation type associated with a mutation site. The D614G is the only mutation, that has so far been commonly observed among the spike proteins from all the continents. Table 3 serves as a reference to consult the presence or absence of a particular mutation within and between the various continents.

3.5 | Mapping mutations in receptor-binding domain of SARS-CoV-2 spike protein

The spike protein plays a vital role for the attachment to host cell-surface specific receptors and subsequently catalyzes the virus - host cell membrane fusion required for causing infection. The RBD in spike protein interacts with host ACE-2 receptor to cause the novel coronavirus infection leading to COVID-19 disease. The three-dimensional crystal structure of human SARS-CoV-2 RBD (between residues 333 and 527) complexed with the ACE-2 receptor (PDB code: 6LZG) was used to map the mutation sites as shown in Figure 2. The RBD of human SARS COV-2 spike proteins from different continents is associated with 44 distinct mutation sites. The mutations are located at positions; 337, **344**, 345, 348, 354, 357, 367, 368, 379, 382, 384, 393, 395, 403, 407, 408, 411, 413, 441, 453, 457, 458, 468, 471, 476, **477**, 479, **483**, 484, 485, 486, 491, 493, 494, 498, 500, **501**, 506, 507, 508, 518, 519, 520, 522. The numbers marked in bold indicate instances of more than 10 occurrences of the mutations observed at the particular position. The numbers underlined indicate instances of more than five occurrences of the mutations observed at the particular position. The distribution of the number of mutations in RBD observed at the 44 different positions is shown in Figure 3. The mutation sites and mutation types of the human SARS-CoV-2 spike protein in the RBD according to the geographical location-wise distribution is included in Table 3. Accordingly, the total number of distinct mutation sites in RBD observed were; North America (27), South America (0), Europe (7), Africa (1), Asia (15), and Oceania (9) and the total number of distinct mutation types observed were; North America (28), South America (0), Europe (7), Africa (1), Asia (16), and Oceania (9). The mutations occurring in relatively large numbers in RBD

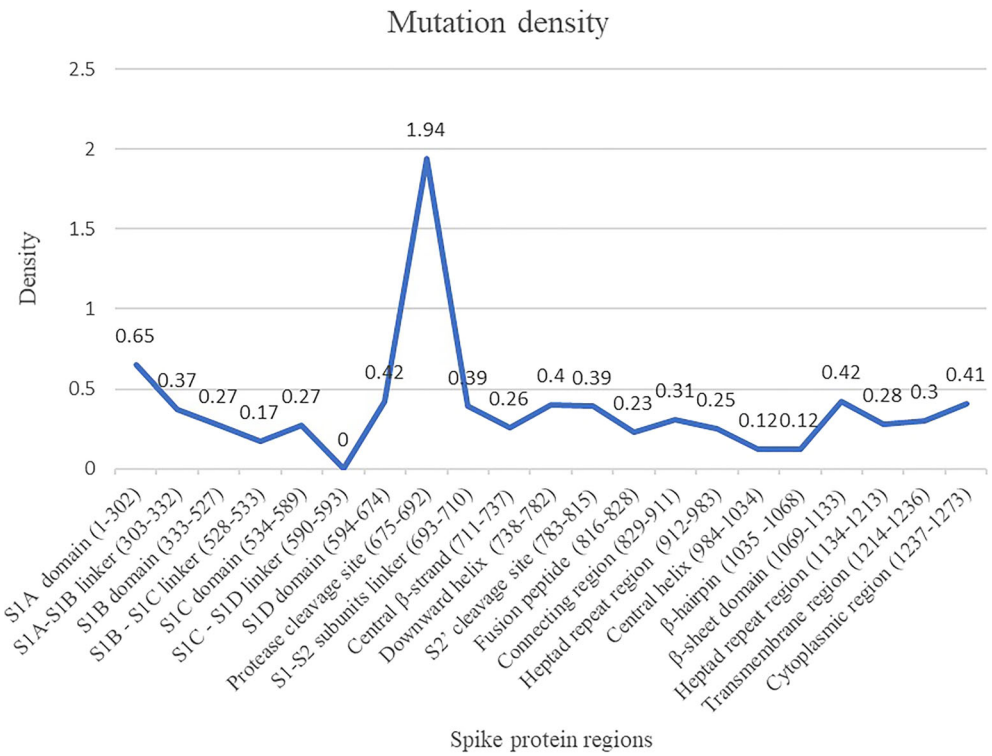

FIGURE 1 Mutation density in human SARS-CoV-2 spike protein regions

were examined. A maximum of 47 mutations were observed at position 477 in the RBD of the spike protein. The S477N mutation was present in 45 spike proteins from Oceania, one from Asia (NCBI ID: QLR12405.1) and one from North America (NCBI ID: QMU91291.1). The V483A mutation was observed in 27 spike proteins representative of North America and one spike protein from Oceania (QLG76529.1) contained a mutation at the same position, but with mutation type V483F. The A344S mutation was observed 18 times and all the associated spike proteins were representative of North America.

Some of the residues in the spike protein RBD that are involved in the interactions with the ACE-2 receptor are mutated as shown in Figure 4. The residues in spike protein RBD that are ≤3.2 Å interatomic distance from the ACE-2 receptor as observed in the crystal structure of the human SARS-CoV-2 spike protein RBD complexed with ACE-2 receptor (PDB code: 6LZG) are; K417, Y449, Y453, A475, N487, T500, N501, and G502. The Y453 residue in spike protein RBD is close to His34 in ACE-2 receptor and T500 and N501 residues in spike protein RBD are close to Tyr41 in ACE-2.[22] Few residues that are close to residues ≤3.2 Å from ACE-2 receptor are also associated with the mutations. These residues are G476 (next to A475), F486 (next to N487) and G502 (next to N501).

Mutations were observed for the residues Y453, T500, N501. The Y453F mutation is present in five spike proteins from Europe (NCBI accession codes: QJS39603.1, QJS39579.1, QJS39543.1, QJS39591.1, QJS39555.1). The T500I mutation is present in the spike protein from Oceania (QLG75833.1). The N501Y mutation is present in 14 spike proteins from Oceania (QLG76793.1, QLG76805.1, QLG76817.1, QLG76085.1, QLG76181.1, QLG76685.1, QLG76697.1, QLG76709.1, QLG76277.1, QLG76097.1, QLG76745.1, QLG76397.1, QLG76469.1, QLG75761.1) and the N501T mutation is present in the spike proteins from Europe (QJS39507.1). The G476S mutation is present in eight spike proteins from North America (NCBI accession codes: QIS30425.1, QIS30625.1, QKS90479.1, QIQ49882.1, QKS90179.1, QIQ50152.1, QJC20487.1, QJD48075.1) and the one F486L mutation in the spike protein from Europe (NCBI accession codes: QJS39567.1).

Therefore, residues at positions; 453, 476, 486, 500, 501 that are associated with the RBD mutations and that are close to the ACE-2 receptor would affect the shape and charge of the protein near the protein-receptor interaction interface. However, despite these mutations in the RBD of human SARS-CoV-2 spike protein, the novel coronavirus causes the infection leading to COVID-19 disease. The regions of protein-protein interactions between virus and host are being targeted as sites for drug design and the spike protein and epitopes are the ideal targets for vaccine design. Therefore, changes in the shape of the protein surface due to the mutations, especially near the protein-receptor binding regions would be important considerations for antibody, vaccine and drug development.

3.6 | Spike protein sequences with multiple mutations

The 8155 spike proteins comprise anywhere from 1 to 16 mutations. The spike proteins containing eight or more mutations in the same protein are discussed below. The spike protein with 16 mutations (QMS95041.1) corresponds to the SARS-CoV-2 isolated on 1 May

TABLE 3 Mutation sites and mutation types observed in human SARS-CoV-2 spike proteins according to geographical locations

North America	F2L, L5F, L5I, V6F, L7V, P9L, S12C, Q14H, C15F, N17K, L18F, T20I, T22N, T22A, T22I, Q23K, P25S, A27S, A27V, T29I, F32L, R34C, H49Y, S50L, T51I, Q52L, Q52H, L54F, L54W, F55I, P57L, H69Y, S71F, G72V, T73I, G75V, T76I, F79L, D80N, D80Y, N87Y, D88N, D88E, D88Y, D88A, V90F, T95A, T95I, E96D, K97T, S98F, R102I, I105L, D111N, K113R, L118F, V130A, E132D, C136R, D138H, L141-, L141F, G142V, G142-, V143F, V143-, Y144-, Y144V, Y145H, H146Y, K147E, N148S, S151I, M153T, M153V, M153I, E154V, F157L, R158S, L176F, M177I, D178N, G181V, L189F, R190K, I203M, I210-, R214L, D215Y, D215G, L216F, Q218L, F220L, S221L, A222V, A222P, D228H, L229F, Q239R, T240I, L242F, A243S, A243V, H245R, H245Y, R246K, D253Y, D253G, S254F, S256L, W258L, G261V, G261R, A262S, Y265C, V267L, R273S, E281Q, A288S, L293V, D294E, P295S, E298G, T307I, V308L, E309Q, Q314K, Q314L, Q314H, T315I, Q321L, T323I, P330S, A344S, T345S, A348T, A348S, N354K, R357K, V367F, V382L, P384L, V395I, R403K, V407I, A411S, G413R, L441I, R457K, K458Q, G476S, S477N, P479L, V483A, E484Q, Q493L, S494P, Y508H, H519Q, A520S, A522V, K529E, G545S, T547I, L552F, T553I, E554D, K558N, A570V, T572I, D574Y, E583D, I584V, S596I, I598V, N603H, Q613H, D614G, V615F, T618A, P621S, V622F, V622I, V622A, A623S, H625Y, A626V, P631S, W633R, G639V, S640F, A647S, A647V, E654Z, E654K, H655Y, N658Y, A668S, A672V, Q675R, Q675H, T676S, T676I, Q677H, Q677R, T678I, P681L, P681H, R682W, A684S, A684T, V687L, A688V, A688S, S689I, S691F, S698L, N703D, S704L, V705F, A706V, I714M, T716I, I720V, T724A, M731I, T732A, T732I, G744V, D745G, N751D, L754F, R765S, R765H, T768I, G769A, A771S, T778I, Q779H, E780Q, A783S, D808V, D808G, P809S, I818V, L822F, D830H, D830Y, Q836P, Q836L, G838D, A845S, A845D, A845V, A845D, A845S, A846V, R847I, K854R, N856S, T859I, D867N, A879V, A879S, A893E, A893V, E918V, L922F, A924S, A924V, S929I, D936H, D936Y, L938F, S939F, T941I, G946V, A958S, N969S, L981F, T1006I, V1008T, T1009I, A1016S, A1020V, F1052L, P1053T, L1063F, V1065L, A1070V, Q1071H, E1072V, K1073N, A1078V, A1078S, G1085R, K1086N, R1091L, H1101Y, V1104L, P1112L, D1118Y, T1120I, V1122L, S1123P, G1124V, G1124C, V1129A, I1130M, T1136I, D1139H, L1141F, D1146H, S1147L, D1153Y, P1162Q, P1162S, P1162Q, P1162S, D1163Y, G1167V, D1168H, V1176F, N1187Y, K1191N, N1192T, E1195Q, L1203F, K1205N, E1207A, 1219 V, G1219S, I1221T, V1228L, M1229I, V1230L, T1231I, T1231A, C1235F, M1237I, M1237V, M1237T, T1238I, K1245N, C1247F, G1251R, D1259H, S1261F, P1263L, V1264L
South America	N74K, I197V, D614G, V1176F
Europe	L5F, T22I, H49Y, Q115R, M153I, L176I, L176F, F186S, N188D, I197V, V213L, T240I, S254F, G261D, V367F, C379F, V382E, T393P, Y453F, F486L, N501T, T553N, K558R, T572I, L611F, D614G, T676I, S686G, M740I, G769V, Y789D, F797C, D839Y, A845S, A1020V, H1101Y, V1122L, P1162L, K1191N, M1229I, D1260N, D1260H, P1263L
Africa	L5F, S12F, T29I, H49Y, V70F, Y144-, L242F, A288T, Q314R, R408I, A570S, D614G, S640A, A653V, Q677H, P812L
Asia	F2L, L5F, L8V, S12F, S13I, Q14H, T22I, P25L, Y28H, Y28N, T29A, G35V, Y38C, H49Y, S50L, L54F, A67V, A67S, I68-, I68R, H69-, V70-, S71-, G72-, T73-, N74-, N74K, G75V, G75-, T76I, T76-, R78M, F86S, T95I, E96G, K97Q, S98F, V127F, D138Y, D138H, F140L, L141-, G142-, V143-, Y144-, H146Y, H146R, N148Y, S151G, W152L, M153I, S155I, E156D, S162I, Q173H, M177I, G181A, N185K, R190S, N211Y, V213L, S221W, Y248H, S255F, W258L, G261R, G261S, A262S, V267L, G268D, Q271R, A292V, L293M, D294I, P295H, L296F, S297W, C301F, P337R, V367F, L368P, V382L, R408I, A411D, E471Q, S477N, E484Q, P491L, Q506H, P507H, P507S, Y508N, L518I, H519Q, A520S, A570V, T572I, D574Y, E583D, G594S, Q607L, Q613H, D614G, V622F, A653V, E654Q, H655Y, Y660F, A672D, Q675-, Q675H, Q675R, T676-, Q677H, Q677-, T678-, N679-, S680-, P681-, R682Q, R682W, R682-, A684V, A684-, R685-, S686-, V687-, A688V, A688-, Q690H, A701V, A706S, M731I, L754F, R765L, A771S, V772I, Q774R, E780D, A783S, K786N, T791I, K795Q, G798A, P809S, T827I, A829T, I834V, A879S, S884F, A892V, M900I, A930V, D936Y, L938F, S939F, S939Y, S974P, Q1002E, L1063F, T1077I, H1083Q, D1084Y, H1088R, F1089V, V1104L, F1109L, D1139Y, S1147L, D1153Y, G1167S, K1181R, R1185H, N1187K, K1191N, E1195Q, Q1201K, V1230E, C1243F, G1246A, D1259Y
Oceania	L5F, T22I, T29I, H49Y, S50L, T76I, S98F, I128F, D138H, M153I, L176F, D178N, E180K, I210-, S221L, S247R, D253G, W258L, A262T, G283V, I468T, E471Z, S477N, V483F, G485R, Q498Z, T500I, N501Y, H519Q, P561L, E583D, D614G, P621S, A626V, Q675K, Q675H, S704L, M731I, T791I, D808B, P812S, D839N, A846V, I931V, D936Y, K1073N, 1079S, G1124V, D1163G, C1254F, D1260N

2020 from Iran, Asia. The mutations; H146R, R190S, G268D, P337R, L368P, A411D, Q607L, D614G, A672D, Q774R, G798A, S974P, S1147L, R1185H, V1230E, G1246A are located at different positions that lie scattered in the spike protein. The protein (QKN61229.1) isolated on 18 March 2020 from Taiwan has two regions of inframe deletions; 68 to 76 corresponding to the NTD domain and 675 to 679 corresponding to the protease cleavage site. The protein (QKS67443.1) isolated in March 2020 from Hong Kong has 11 mutations between residues 679 and 688 also within the protease cleavage site and a V367F mutation in RBD. The protein (QJD23249.1) isolated 20th March 2020 from Malaysia has eight mutations; between residues 292 and 297 in NTD and two mutations; P491L and H519Q in RBD.

3.7 | Spike proteins not associated with mutations

We observed 2178 spike proteins that do not show any mutations relative to the human SARS-CoV-2 spike protein RefSeq from Wuhan-Hu-1, China. Interestingly, some of the human SARS-CoV-2 genomes isolated during June 2020 have not undergone any mutation yet. These include 42 spike protein sequences from North America (NCBI accession codes: QMT50797, QMT51409, QMT51505, QMT51865, QMT52129, QMT52237, QMT52249, QMT52393, QMT52561, QMT52741, QMT52765, QMT53017, QMT53041, QMT53053, QMT53065, QMT53089, QMT53101, QMT53149, QMT53173, QMT53197, QMT53221, QMT53233, QMT53245, QMT55880, QMT57260, QMT57332, QMT57572, QMT57584, QMT57608,

QMT57644, QMT57656, QMT57692, QMT94108, QMT94756, QMT94780, QMT95200, QMT95308, QMT95356, QMT95368, QMT95452, QMT95488, QMT95560), five from Asia (QLL26046, QLI49781, QLF98260, QKY60061, QKV26077) and two from Africa (QKR84420, QKS66940).

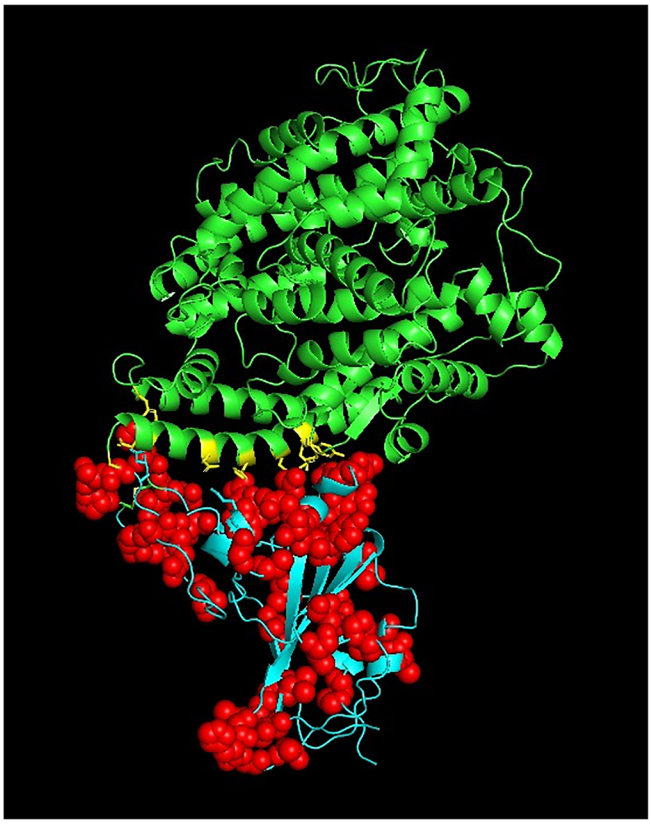

FIGURE 2 The 44 mutations (red spheres) mapped on to the crystal structure of the spike protein RBD (cyan) complexed with ACE-2 receptor (green) (PDB code: 6LZG). PDB, Protein Data Bank; RBD, receptor binding domain [Color figure can be viewed at wileyonlinelibrary.com]

3.8 | Glycosylation site mutations

The NxT/S represent glycosylation site sequence motifs. The deletions at glycosylation sites; N331 and N343 in the spike proteins are known to have caused lesser infections revealing the importance of glycosylation for viral infectivity.[9] The examination of three-dimensional electron microscopy structure of human SARS-CoV-2 spike protein (PDB code: 6VSB) on the graphics showed a number of glycosylation sites in the human SARS-CoV-2 spike protein. I observed instances of the SARS-CoV-2 spike proteins where either the N or the S/T in the NxT, NxS sequence motifs were mutated. The spike proteins from North America; (NCBI accession code: QKU31901.1) is associated with N17K mutation and (QKV40463.1) is associated with T1136I mutation that is glycosylated at N1134. The spike proteins from Asia; (QLG99547.1) is associated with S151I mutation and (QLA46612.1) with S151G mutation and these spike proteins are glycosylated at N149.

4 | CONCLUSIONS

The human SARS-CoV-2 spike proteins comprised 400 distinct mutation sites with reference to the first human SARS-CoV-2 sequence from Wuhan-Hu-1, China. The mutations are present in 8155 proteins among 10333 human SARS-CoV-2 spike protein sequences analyzed in the present work. The total number of mutations observed were 9654 for all the spike proteins. The mutation sites are distributed over whole length of the protein sequence with the maximum mutation density being near the protease cleavage site between residues 675 and 692. The RBD is associated with 44 mutation sites and within the RBD, the mutations; S477N, V483A, A344S, N501Y were more frequent. The D614G mutation is predominant and is the only common mutation in the spike protein observed so far among all the continents. Some of the SARS-CoV-2 spike proteins are associated with the mutations at glycosylation sites. Nearly, 21% SARS-CoV-2 spike proteins have not undergone any mutations yet with respect to the first human SARS-CoV-2 Wuhan-Hu-1 reference sequence that became available during

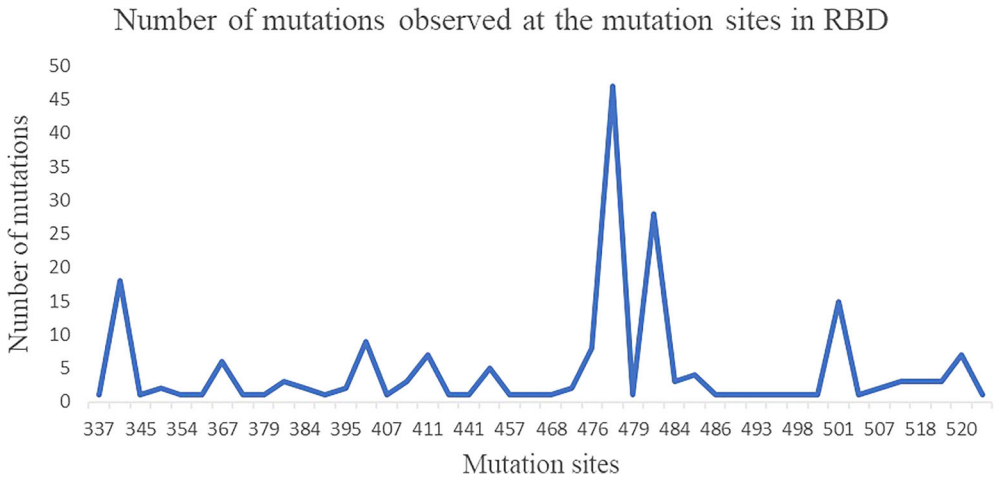

FIGURE 3 Mutations in human SARS-CoV-2 spike protein RBD. RBD, receptor binding domain [Color figure can be viewed at wileyonlinelibrary.com]

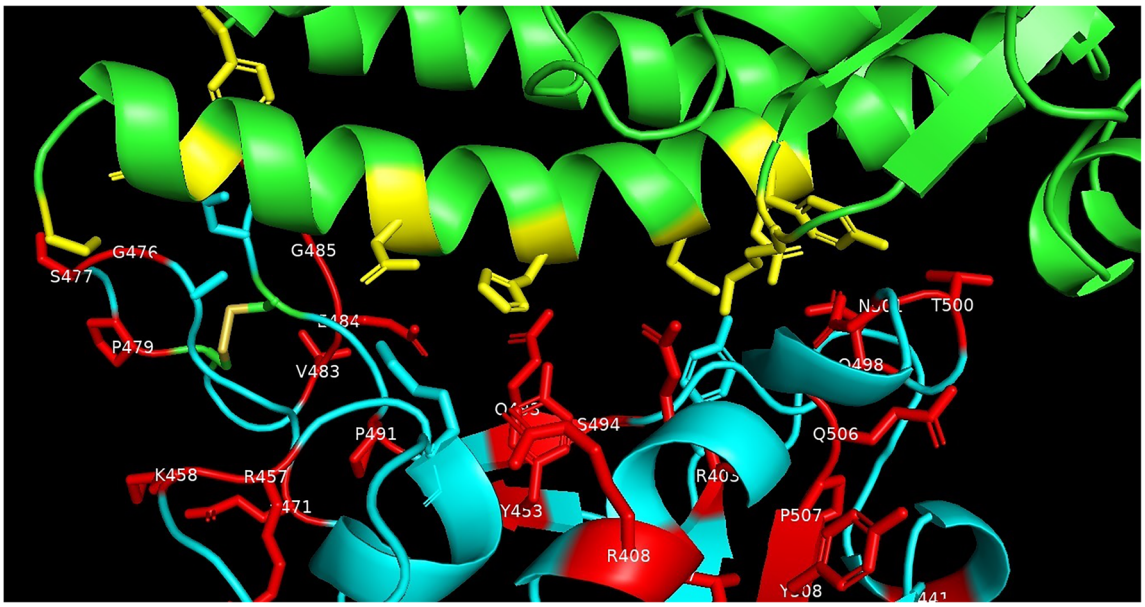

FIGURE 4 Interactions of spike protein residues (cyan) with ACE-2 (green) side-chain residues (yellow) that are within 3.2 Å in crystal structure of human SARS-CoV-2 spike protein RBD complexed with ACE-2 receptor (PDB code: 6LZG). The spike protein mutated residues are shown in (red). PDB, Protein Data Bank; RBD, receptor binding domain [Color figure can be viewed at wileyonlinelibrary.com]

December 2019. Within the RBD, mutations were observed for Y453, G476, F486, T500, N501 that are close to the ACE-2 receptor. The mutations present at the interface between the spike protein and ACE-2 receptor could potentially affect vaccine performance and drugs designed at the interface of protein-protein interactions. Therefore, the mutations identified in the present work would be important considerations for antibody, vaccine, and drug development.

ACKNOWLEDGEMENTS

LGP thanks School of Chemistry, University of Hyderabad for research facilities and ABREAST (https://www.abreast.in) for making available the computer programs used in this work for the identification and analyses of the mutations.

CONFLICT OF INTEREST

The author declares no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26042.

DATA AVAILABILITY STATEMENT

All data is available in the manuscript and Table S1.

ORCID

Lalitha Guruprasad https://orcid.org/0000-0003-1878-6446

REFERENCES

1. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269.
2. Zhang H, Penninger JM, Li Y, Zhong N, Slutsky AS. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Med*. 2020;46(4):586-590.
3. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270-273.
4. Han GZ. Pangolins harbor SARS-CoV-2-related coronaviruses. *Trends Microbiol*. 2020;28(7):515-517.
5. Lam TT, Jia N, Zhang YW, et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*. 2020;583:286-289.
6. Guruprasad L. Human coronavirus spike protein-host receptor recognition. *Prog Biophys Mol Biol*. 2020. https://doi.org/10.1016/j.pbiomolbio.2020.10.006.
7. Wang Q, Zhang Y, Wu L, et al. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell*. 2020;181:894-904.
8. Ou X, Liu Y, Lei X, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun*. 2020;11:1620.
9. Li Q, Wu J, Nie J, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*. 2020;182(5):1284-1294.
10. Kaushal N, Gupta Y, Goyal M, Khaiboullina SF, Baranwal M, Verma SC. Mutational frequencies of SARS-CoV-2 genome during the beginning months of the outbreak in USA. *Pathogens*. 2020;9(7):565.
11. Korber B, Fischer WM, Gnanakaran S, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020;82(4):812–827.e19.
12. Phelan J, Deelder W, Ward D, Campino S, Hibberd ML, Clark TG. Controlling the SARS-CoV-2 outbreak, insights from large scale whole genome sequences generated across the world. *BioRxiv*. 2020. https://doi.org/10.1101/2020.04.28.066977.
13. Mercatelli D, Giorgi FM. Geographic and genomic distribution of SARS-CoV-2 mutations. *Front Microbiol*. 2020;11:1800. https://doi.org/10.3389/fmicb.2020.018002020.
14. Guruprasad K. Amino acid mutations in the protein sequences of human SARS CoV-2 Indian isolates compared to Wuhan-Hu-1

reference isolate from China. *ChemRxiv*. 2020. https://doi.org/10.26434/chemrxiv.12300860.v1.
15. Guruprasad K. Mapping mutations in proteins of SARS CoV-2 Indian isolates on to the three-dimensional structures. *ChemRxiv*. 2020. https://doi.org/10.26434/chemrxiv.12683771.v1.
16. Yadav PD, Potdar VA, Choudhary ML, et al. Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J Med Res*. 2020;151(2):200-209.
17. Saha P, Banerjee AK, Tripathi PP, Srivastava AK, Ray U. A virus that has gone viral: amino acid mutation in S protein of Indian isolate of coronavirus COVID-19 might impact receptor binding, and thus, infectivity. *Biosci Rep*. 2020;40(5):BSR20201312.
18. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772-780.
19. Lemoine F, Correia D, Lefort V, et al. NGPhylogeny. Fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Res*. 2019;47(W1):W260–W265.
20. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28(1):235-242.
21. DeLano WL. Pymol: an open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*. 2002;40:44-53.
22. Guruprasad L. Evolutionary relationships and sequence-structure determinants in human SARS coronavirus-2 spike proteins for host receptor recognition. *Proteins*. 2020;88:1387-1393.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Guruprasad L. Human SARS CoV-2 spike protein mutations. *Proteins*. 2021;89:569–576. https://doi.org/10.1002/prot.26042