Comparative analysis based on the spike glycoproteins of SARS-CoV2 isolated from COVID 19 patients of different countries.

Preeti Mangar¹, Smriti Pradhan², Subecha Rai³, Khusboo Lepcha⁴, Vivek Kumar Ranjan², Aditi Rai⁴*
¹Department of Botany, University of North Bengal, Raja Rammohanpur, Darjeeling, West Bengal-734013, India.
²Department of Biotechnology, University of North Bengal, Raja Rammohanpur, Darjeeling, West Bengal-734013, India.
³Institute of Bioresources and Sustainable Development, Sikkim Centre, 5th Mile, Metro Point, Tadong, Gangtok, Sikkim-737102, India.
⁴Department of Microbiology, University of North Bengal, Raja Rammohanpur, Darjeeling, West Bengal-734013, India.

Abstract

SARS-CoV2 popularly known as (COVID-19) has presently received worldwide attention. It has been considered a pandemic by the World Health Organisation. Owing to its high transmittance factor the virus has brought about many deaths and spread to all the major countries of the world. Scientists and Researchers worldwide are giving their full efforts to develop a vaccine. In our present study, we have included the comparative analysis of the different spike glycoprotein sequences of the patients suffering from COVID-19 from different countries where this pandemic has occurred. Spike glycoproteins are the structural proteins that bring about the binding of the SARS-CoV-2 viral molecule to the ACE2 receptor of the host following which infection occurs. Through this data, we have shown the different point mutations in the spike glycoproteins that occurred over time in different countries as the disease progressed.

Keywords: SARS-CoV2, spike glycoprotein, mutation

*Corresponding Author- aditimicro009@nbu.ac.in

1. Introduction

Diseases since time immemorial has always cost mankind health and wealth, not surprisingly the quest for survival of mankind has been an everlasting battle. Epidemics and pandemics are not new to the history of mankind and records of diseases like Plague and Spanish Flu devastating lives have been an integral part of the epidemiological study of the human race. In December 2019, an incident occurred in Wuhan, Southern China where a series of pneumonia cases, was reported. It wasnt long before the cases were classified as viral pneumonia and the virus was speculated to belong to β coronavirus. Primarily it was named as 2019- novel coronavirus (2019-nCoV) by World Health Organization (WHO), which later named the disease as coronavirus disease 2019 (COVID-19). Also identified with the name SARS-CoV2,
the epidemic COVID-19 progressed by leaps and bounds via human-to-human transmission, that made it basically impossible to contain it within a certain area thus leading to a pandemic that crossed borders spreading into the International community. The SARS-CoV2 being a β-coronavirus (Zhu et., 2019) like SARS-CoV and MERS-CoV is responsible for causing severe and potentially fatal respiratory tract infections (Yin et al., 2018) Sharing 96.2% identity to a bat CoV RaTG13, and 79.5% identity to SARS-CoV, it may be presumed that the SARS-CoV2 might be transmitted from bats to humans. However, recent studies comparing the receptors on host surfaces, suggest the possibility of alternative intermediate hosts (Liu et al., 2020).

**Spike glycoprotein of SARS-CoV 2**

The complete genome analysis of a strain of SARS-CoV2, obtained from Wuhan revealed that this enveloped virus has a positive stranded RNA genome with a size about 29.9 kb (Wu et al., 2020) The study of genomes of CoVs has revealed a variable number of (6–11) open reading frames (ORFs) (Song et al., 2019). The first ORF (ORF1a/b) encodes 16 non-structural proteins (NSP), and translates two polyproteins, pp1a and pp1ab, and the remaining ORFs encode accessory and structural proteins. However, four essential structural proteins are encoded by the viruses including spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins( Wang et al., 2020; Du et al., 2016). Among them, an envelope-anchored spike protein specifically recognizes its host receptor and it serves as a target for development of antibodies, entry inhibitors and vaccines (Du et al., 2016; He et al., 2006) The S protein is trimeric and each monomer is about 180 kDa, and contains two subunits, S1 and S2. The minimal RBD (Receptor Binding Domain) region is the fragment covering the residues 318-510 in S1 subunit (Xiao et al., 2003, Wong et al., 2004). The receptor-binding motif (RBM) is the RBD containing a loop region (residues 424-494), which makes complete contact with the receptor ACE2 (Angiotensin Converting Enzyme), the RBM region being tyrosine-rich, It was observed that six out of 14 residues of RBM were tyrosines that were in direct contact with ACE2 (Zhu et al., 2013). The S protein first binds to a host receptor through the receptor-binding domain (RBD) in the S1 subunit and then fuses the viral and host membranes through the S2 subunit (Liu et al., 2004). In the structure, N-terminal domain (NTD) and C-terminal domain (C-domain) portions of S1 fold as two independent domains, Depending on the virus, either NTD or C-domain (occasionally both) binds to a host receptor and functions as a receptor-binding domain (RBD) (Breslin et al., 2003; Lin et al., 2008). Recently, (Zhou et al., 2020) reported that SARS-CoV-2 uses ACE2 as the receptor which is similar to the S1 C-domain of SARS-CoV in the RBD that recognizes host angiotensin-converting enzyme 2 (ACE2) as its receptor (Babcock et al., 2004; Li et al., 2003). In one of the studies by (Wan et al., 2020) it was observed that among the ACE2-contacting residues in the RBD, 9 are fully conserved and 4 are partially conserved among 2019-nCoV and SARS-CoV from human, civet, and bat. ACE2 is a zinc-dependent peptidase that functions in the renin-angiotensin pathway and regulates blood pressure (Donoghue et al., 2000; Yagil and Yagil, 2003). However, the physiological function of ACE2 is not related to its role as the SARS-CoV receptor (Li et al., 2005b). ACE2 contains an N-terminal peptidase domain and a C-terminal collectrin domain. The
enzymatic active site of ACE2 is buried in a claw-like structure with two lobes of the peptidase domain (Towler et al., 2004). The binding interactions between SARS-CoV, RBD and ACE2 largely determine the host range and cross-species infections of SARS-CoV (Lu et al., 2015). Using computer modeling, Xu et al. (2020), found that the spike proteins of SARS-CoV-2 and SARS-CoV share 76.5% identity in amino acid sequences having almost identical 3-D structures in the receptor-binding domain that maintains Van der Waals forces. It has been reported that residue 394 (glutamine) in the SARS-CoV-2 receptor-binding domain (RBD), which corresponds to residue 479 in SARS-CoV, can be recognized by the critical lysine 31 on the human ACE2 receptor (Wu et al., 2012). Through recent cryo-EM structure studies which further deciphered the S protein of SARS-CoV2 and ACE2 interaction at Angstrom resolution level it was revealed that the overall ACE2-binding mode of S protein of SARS-CoV2 is almost identical to the mode of S protein of SARS-CoV (Yan et al., 2020; Lan et al., 2020). Yan et al., 2020 performed the bioinformatics analysis of the S protein sequences of coronaviruses and found an evolutionary mutation of K403R in S protein of SARS-CoV2 compared with that in S protein of SARS-CoV, forming an adjacent RGD (R:Arginine, G:Glycine, D:Aspartic Acid) motif at the interaction surface. The RGD motif is the cell attachment site of a large amount of adhesive extracellular matrix and cell surface proteins and recognized by integrins. The evolutionary obtainment of the RGD (R:Arginine,G:Glycine,D:Aspartic Acid) motif in SARS-CoV2 may play an important role in promoting rapid human to human transmission. Yan et al., 2020 also suggested that the RGD motif on the S glycoprotein may bind to the integrin on the surface of host cells which can be a new potential mechanism for SARS-CoV2 infection, resulting in higher affinity with the host cells in comparison with SARS-CoV.

Spike proteins has been found to be immunogenic and induce high IFN-γ-specific T-cell response (Janice et al., 2012). Mutations in the spike protein could change the tropism of a virus, including new hosts or increasing viral pathogenesis (Shang et al., 2020). SARS-CoV possess some residues in RBD that allows the interspecies infection, known as Y442, L472, N479, D480, and T487 (Lu et al., 2015). However, in SARS-CoV2, slight modification of some residues could improve the interaction with the human cellular receptor: L455, F486, Q493, and N501. In SARS-CoV, two main residues (479 and 487) have been associated to the recognition of the human ACE2 receptor (Lu et al., 2015). These residues suffered a punctual mutation from civet to human, K479N and S487T (Li, 2013). In the SARS-CoV2, the residues corresponding to N479 correspond to Q493 and T487 to N501. Moreover, a model shows the presence of the two capping loops in the binding domain which produces a stabilization effect over the interaction with the cellular receptor (Ortega et al., 2020). Thus the amino acid substitutions and the longer capping loops could explain the increase in binding affinities in SARS-CoV2 compared to SARS-CoV. Since mutations play a major role, our focus of the present study was to understand the mutations in the spike glycoproteins from different countries as it could provide us an idea about the constant shift in the structure of the spike glycoproteins and probably enabling it to be transmitted to different regions. However, is the mutation dependant on the race or ethnicity of a person or the gene pool is an entire new story altogether.
2. Methodology

A total of 22 coding sequences for spike glycoproteins were retrieved from NCBI database (http://www.ncbi.nlm.nih.gov). Multiple Sequence alignment of the CDS region was performed using CLUSTAL Omega (https://www.ebi.ac.uk/Tools/msa/clustalomega) (Madeira et al., 2019). The sequences were selected on the basis of their origin to 14 different countries affected by the COVID19 pandemic namely USA, India, China, Australia, Finland, South Korea, Brazil, Italy, Japan, Vietnam, Pakistan, Sweden, Taiwan and Spain (Accession numbers have been mentioned in Table 1). The phylogenetic analysis of SARS-CoV2 spike proteins of the different countries was done using MEGAX software (MEGA-X Version 7.0) (Kumar et al., 2018). The phylogenetic analysis was accomplished through multiple comparisons using the neighbor-joining algorithm in the MEGA-X. Multiple comparisons were done by ClustalW multiple sequence alignment and the neighbor-joining phylogenies were estimated by p-distance method.

3. Result and Discussion

A total of 22 different amino acid sequences of the spike glycoproteins from different countries were analysed by Multiple Sequence alignment (Table 1, Fig 2). The spike glycoprotein sequence of India showed mutations in the S1 and S2 domains. The mutation Ala930Val in the spike protein of the Indian sequence (Accession Number Q1A985839) has been observed to be in the S2 domain. It is well understood that these point mutation enhances the surface area for interaction with the ACE2 receptor while conserving the physico-chemical property of the side chain. Additionally, increasing the chance of vander-walls interactions and contributing to the protein core stability another mutation of Arg408Ile is noted to in the RBD region of the spike protein of another Indian sequence (Accession number MT012098). It has been seen that the RBD regions are mostly tyrosine rich to ensure proper contact with the ACE2 receptor (Zhu et al., 2013). However in contradiction reducing the binding affinity, the same spike glycoprotein of India shows a deletion at 145 amino acid position whereas the rest of the spike sequences from the other countries have a tyrosine residue. The spike glycoprotein of China (Accession numbers Q1A20044 and Q1004367) and Finland (Accession number QHU79173) showed mutations in Tyr28Asn, Asn74Lys and His49Tyr respectively. Our study is the first report that indicates a point mutation of His49Tyr in the sequence of Finland (QH QHU79173) sharing similarity with SARS-CoV Sβ(s-domain) of Urbani sequence(Accession number AAP1344)(data not shown) isolated in the late phase of the 2002-2003 SARS-CoV epidemic and dissimilarity with the remaining SARS-CoV2 spike sequences indicating reversion of mutation at specific sites. Likewise the spike glycoproteins of Australia (Accession Number QHR84449) showed mutation at Ser247Arg. It is quite interesting to note that the four different sequences of United States from Cruise A(Accession numbers QII57278, QI96493, QIK50427) and Washington (Accession number QH1187830) show mutation at two different positions where one mutation is at Phe157Leu and the other mutation is at Gly181Val ,Asp614Gly and His655Tyr
respectively. According to our study the spike glycoproteins of USA showed the maximum variations
between the submitted sequences from United States itself. Though a drastic difference in the RBD region
is not observed it is remarkable how small point mutations in other regions has shifted towards providing
a favourable environment through hydrophobic interactions and hydrogen bonds as it binds to the host
receptor. The spike glycoprotein of Sweden (Accession number Q1C53204) showed an amino acid
deletion at the 910 position, however comparatively in all the other sequences of the spike glycoprotein
of the other countries glycine was present. Interestingly, another spike glycoprotein sequence of Sweden
(Accession number Q1C53204) showed a mutation in the Phe797Cys. The two Korean spike glycoprotein
sequences (Accession number QH200379 and Accession number MT039890) showed mutation in the
Ser221Trp however, it is interesting to note that the first spike glycoprotein sequence QH200379 was
taken from a patient who had travelled to Wuhan from Korea and it is likely that he/she may have been
infected in Wuhan, China. Probably these respective mutations in the SARS-CoV2 spike proteins of the
different countries contribute to the occurrence of branches forming different clades in the phylogenetic
tree (Fig 1). In the phylogenetic tree (Fig 1) SARS CoV-2 spike glycoprotein sequence from Sweden
(QIC53204), China (YP009724390), Spain (QIQ08810) and Taiwan( MT06617) form a separate clade
from the other sequences. Most of the SARS-CoV2 spike glycoprotein sequences of USA (QII57278,
QII87830, QII96493, QIK50427), China(QIA20044), India(QIA98583), Finland(QHU79173,
Australia(QHR84449) and South Korea(QHZ00379, MT039890) formed distinct cluster within the
second clade. The remaining sequences of Brazil (MT126808), Italy(MT066156 ), Japan (LC528232)
and China(NC_045512 )corresponded to the second clade of the phylogenetic tree. However, the
sequences of India (MT012098) and Vietnam (MT192772) formed a separate cluster within the
phylogenetic tree. It is very important to analyse the spike glycoprotein sequences within different
locations so as to monitor the spread and mobilization of the SARS-CoV2 to different countries across
the globe.

Continuous analysis of the spike glycoprotein sequences of SARS-CoV2 obtained from different regions
interacting with the ACE2 receptor is important. The various conserved domains as well as point
mutations were noted across amino acid sequences pertaining to the spike glycoprotein of 14 different
countries. Thus, one or a few seemingly trivial mutations at the receptor-binding surface of a virus may
lead to dramatic epidemic outcomes by facilitating cross-species infections and human- to-human
transmission of the virus. The changes in the amino acid residue of the Receptor binding domain may
determine the host’s fate and role in a viral epidemic by presenting species barriers for viral infections.
The findings can fill in an important missing link and lead to development of vaccines and therapeutics
associated with the COVID-19 pandemic. Also, the present work gives a better insight to understand the
positions of amino acids which may be susceptible to mutations and can drastically aid SARS-CoV2 to
evolve in the near future to another potential pathogenic strain.

References-


Shijia Y, Haixia S, Xianzhang B, Guohui W. 2020. An evolutionary RGD motif in the spike protein of SARS-CoV-2 may serve as a potential high risk factor for virus infection?


Table 1: It represents the Accession Number, Country and the Date of submission of the sequences in the NCBI database.
Fig 1. Phylogenetic analysis of SARS cov2 spike proteins obtained from different countries. The evolutionary history was inferred using the Neighbor-Joining method. The optimal tree with the sum of branch length = 0.00942794 is shown. The evolutionary distances were computed using the p-distance method and are in the

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>Accession No.</th>
<th>Place</th>
<th>Date of submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>QHR84449</td>
<td>Australia</td>
<td>30.01.2020</td>
</tr>
<tr>
<td>2</td>
<td>QHU79173</td>
<td>Finland</td>
<td>17.03.2020</td>
</tr>
<tr>
<td>3</td>
<td>QH200379</td>
<td>South Korea</td>
<td>11.02.2020</td>
</tr>
<tr>
<td>4</td>
<td>QIA20044</td>
<td>Yunnan(China)</td>
<td>09.02.2020</td>
</tr>
<tr>
<td>5</td>
<td>QIA98583</td>
<td>India</td>
<td>11.02.2020</td>
</tr>
<tr>
<td>6</td>
<td>QIC53204</td>
<td>Sweden</td>
<td>20.02.2020</td>
</tr>
<tr>
<td>7</td>
<td>Q1157278</td>
<td>Cruise A (U.S.A.)</td>
<td>07.03.2020</td>
</tr>
<tr>
<td>8</td>
<td>Q1187830</td>
<td>WA/ (U.S.A.)</td>
<td>06.03.2020</td>
</tr>
<tr>
<td>9</td>
<td>Q1J96493</td>
<td>Cruise A (U.S.A.)</td>
<td>12.03.2020</td>
</tr>
<tr>
<td>10</td>
<td>QIK50427</td>
<td>Cruise A (U.S.A.)</td>
<td>16.03.2020</td>
</tr>
<tr>
<td>11</td>
<td>Q1004367</td>
<td>Yunnan (China)</td>
<td>20.03.2020</td>
</tr>
<tr>
<td>12</td>
<td>Q1Q08810</td>
<td>Valencia (Spain)</td>
<td>23.03.2020</td>
</tr>
<tr>
<td>13</td>
<td>MT012098</td>
<td>India</td>
<td>01.02.2020</td>
</tr>
<tr>
<td>14</td>
<td>YP009724390</td>
<td>Wuhan (China)</td>
<td>17.01.2020</td>
</tr>
<tr>
<td>15</td>
<td>MT240479</td>
<td>Pakistan</td>
<td>25.03.2020</td>
</tr>
<tr>
<td>16</td>
<td>MT126808</td>
<td>Brazil</td>
<td>02.03.2020</td>
</tr>
<tr>
<td>17</td>
<td>MT192772</td>
<td>Vietnam</td>
<td>17.3.2020</td>
</tr>
<tr>
<td>18</td>
<td>LC528232</td>
<td>Japan</td>
<td>29.02.2020</td>
</tr>
<tr>
<td>19</td>
<td>MT039890</td>
<td>South Korea</td>
<td>11.02.2020</td>
</tr>
<tr>
<td>20</td>
<td>MT066175</td>
<td>Taiwan</td>
<td>14.2.2020</td>
</tr>
<tr>
<td>21</td>
<td>NC045512</td>
<td>Wuhan (China)</td>
<td>3.3.2020</td>
</tr>
<tr>
<td>22</td>
<td>MT066156</td>
<td>Italy</td>
<td>09.03.2020</td>
</tr>
</tbody>
</table>
units of the number of amino acid differences per site. The analysis involved 22 amino acid sequences. Evolutionary analyses were conducted in MEGAX software.

Fig 2. Multiple sequence alignment of 22 SARS-CoV2 glycoprotein encoding sequences using CLUSTAL Omega (https://www.ebi.ac.uk/Tools/msa/clustalw2). The “*” means the identical and fully conserved amino acid residue. “:” means conservation within strong group of amino acid residue. “.” means conservation within weaker group of amino acid residues. The red bordered boxes indicate the point mutations.