Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

In Silico Investigation and Correlation of Hydrophobic Stretches in Spike Proteins of SARS-CoV-2, SARS-CoV and MERS-CoV

Uma Shekhawat[1,2], Anindita Roy Chowdhury (Chakravarty)[1,*]

[1] School of Engineering and Sciences, GD Goenka University, Gurugram Haryana, 122103, India
[2] Department of Physics, Pt. Jawaharlal Nehru Govt. College, Faridabad, Haryana, 121002, India

ARTICLE INFO

Article history:
Received 4 December 2023
Received in revised form 18 April 2024
Accepted 24 May 2024
Available online 25 July 2024

Keywords:

Consecutive hydrophobic; MERS-CoV; SARS-CoV; SARS-CoV-2; Spike protein; Secondary structure

ABSTRACT

Hydrophobic force is a key factor for the three-dimensional structure and stability that a protein will adopt. It is important to understand how hydrophobic interactions affect the protein folding process for a protein to become functional. In SARS-CoV-2, SARS-CoV, and MERS-CoV, spike protein serves as both the primary structural protein and a multifunctional protein. The authors aim to investigate the secondary structure of spike protein sequences of three primary coronaviruses corresponding to the stretches of consecutive hydrophobic amino acid residues by developing and implementing computer programs. Besides the primary sequences, similar investigations were carried out for other aligned sequences from different coronaviruses and source organisms. Understanding the potential impacts of hydrophobic amino acid residues in spike protein would be helpful to gain insight into its stability and thus, aid in the study of viral pathogenicity as well as their implications effects on immunogenicity and treatment.

1. Introduction

The protein's native structure is governed by the balance between various competing forces; the primary force exerting the most significant influence on the three-dimensional structure, stability, and folding of a protein is the hydrophobic force [1-3]. To keep the protein stable and biologically active, hydrophobic interactions are essential [4-6]. The primary structure of a protein refers to the sequential arrangement of amino acids within the polypeptide chain, while protein secondary structures encompass repetitive and regular conformations of the polypeptide chain [7]. The interaction between neighbouring or long-distance amino acid residues is represented by protein secondary structures [7,8] which fundamentally include alpha helix, beta sheet and coil [7,9,10].

A new coronavirus known as the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the cause behind COVID-19 disease. The Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) which were responsible for the preceding epidemics in 2002 and 2012, respectively., are known to cause fatal pneumonia

* *Corresponding author.*
E-mail address: aninditaroy.chowdhury@gdgu.org

https://doi.org/10.37934/araset.49.1.1125

[11-18]. According to research by Hatmal *et al.*, SARS-CoV-2, SARS-CoV and MERS-CoV coronaviruses are genetically highly similar, although there are significant changes between them at the level of protein production [19]. The structural proteins of these coronavirus strains are majorly responsible for the infection. Amongst structural proteins, spike protein is the most crucial for both viral pathogenesis and the infection process [12]. The spike protein is a primary target for the development of antibodies as it is present in all coronaviruses [8,20]. The spike protein is a Type-I transmembrane (TM) glycoprotein that serves as a homotrimer on the viral surface and is crucial for viral attachment and entrance into host cells as well as binding to ACE2 receptors [11,19,21].

Quantitative studies of hydrophobic mapping of protein sequences and secondary structure configuration would offer major insights into the structural elements crucial to structural stability for several significant challenges in molecular biology, including protein-protein interactions, protein folding, tertiary properties, and protein-nucleic acid interactions [22,23]. The protein structure of the virus is crucial to its functions, and a variation in shape of the structure might result in non-functional proteins by affecting its functionalities, virulence, infectiousness, and transmissibility [8].

A missing or wrong amino acid can cause proteins to misfold, resulting in proteins with different properties that may lead to diseases. Therefore, an understanding of amino acid residues and protein structure is vital in protein research [9,24,25]. Cognizance of stability and structural integrity of the spike protein can be gained by investigating quantitative distribution of hydrophobic residues which can be carried out using an experimental scale of Fauchere & Pliska [4,22,26-29]. In the present investigation, the authors aim to examine the presence of consecutive aromatic or aliphatic hydrophobic amino acid residues and their correlation with secondary structure elements of the spike protein sequences of SARS-CoV-2, SARS-CoV, and MERS-CoV along with their aligned protein sequences. The authors followed a computational approach in order to carry out this investigation.

2. Methodology

2.1 Selection of Protein Sequences

The spike protein sequence of three primary coronavirus sequences (SARS-CoV-2 UniProt ID: P0DTC2, SARS-CoV UniProt ID: P59594, and MERS-CoV UniProt ID: A0A7D5J875) and their aligned protein sequences as obtained with BLAST tool was considered for this investigation. Through the utilization of the BLASTP sequence alignment tool 42, 48, and 153 aligned protein sequences have been observed for the primary spike protein sequences of SARS-CoV-2, SARS-CoV, and MERS-CoV, respectively. Based on E-value < 10^{-5} and specific query coverage range, the aligned sequences have been selected for further analysis [22,30-32].

Furthermore, multiple sequence alignment was carried out using ClustalW [33,34], and the highest alignment scored protein sequences were identified. In Figure 1, the Venn diagram illustrates the grouping of the different aligned protein sequences in reference to the spike protein sequence of the three primary coronaviruses. In addition to the three primary spike protein sequences, there were twelve more protein sequences, which were aligned to at least either of the three. So, in total there were fifteen protein sequences selected for further analysis.

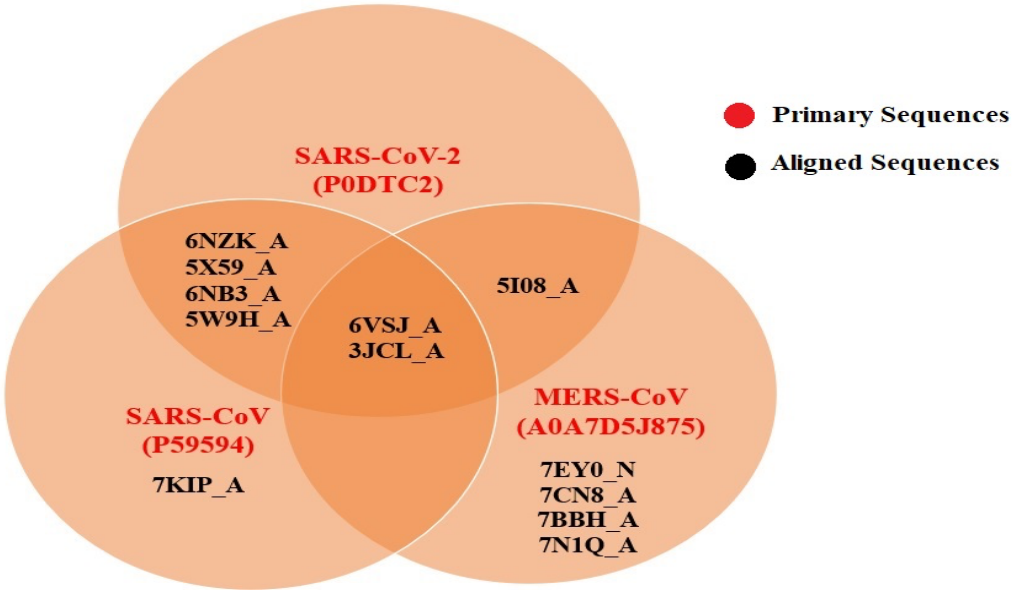


Fig. 1. Primary and aligned protein sequences related to SARS-CoV-2, SARS-CoV, and MERS-CoV [30]

2.2 Determination of Hydrophobic Stretches

The authors wanted to investigate the presence of stretches of consecutive hydrophobic amino acids (HS) in the spike protein sequence of primary sequences and their aligned sequences. In order to do so, they considered the presence of minimum of three consecutive hydrophobic (aromatic/aliphatic) amino acids as one hydrophobic stretch; consecutive pairwise presence of hydrophobic amino acids may be too frequent and hence, was not considered as a hydrophobic stretch.

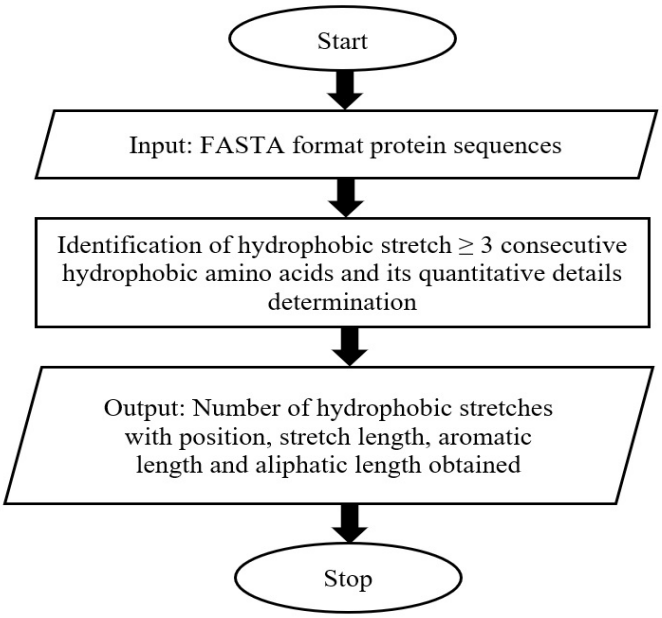


Fig. 2. Flowchart of the program to identify and determine essential details of hydrophobic stretches

A C++ program 'hspseq.cpp' had been devised to determine the hydrophobic stretches (HS) with respective details for each of the fifteen protein sequences considered in this study. The downloaded fasta sequences of SARS-CoV-2, SARS-CoV, and MERS-CoV spike proteins and their aligned sequences were used as input to the C++ program. On executing the program, output files were generated showing the hydrophobic stretches with the position, stretch length, aromatic and aliphatic length for each protein sequence. The uniqueness of the program lies in identification of three or more consecutive aromatic and aliphatic hydrophobic amino acid residues in each protein sequence. The flowchart of the program to identify the hydrophobic stretches in each protein sequence has been described below (Figure 2).

2.3 Determination of Secondary Structure

A BhageerathH+ software that requires amino acid sequence information in FASTA (*.fasta) format as an input has been used to get the secondary structure. BhageerathH+ is consisting of three major steps in its algorithm which include structure creation for conformation sampling, scoring the structures to identify the best conformation as well as optimization of side chains. While carrying out the above processes it can also deliver information about protein annotation [35-37]. On executing BhageerathH+ software on spike protein and their aligned sequences, an output file containing information about the secondary structure elements of each of the protein sequences was generated. This was used as an input file for another C++ program 'seconhs.cpp'. This program was developed to determine hydrophobic stretches of each of the primary spike protein sequence of coronavirus as well as their aligned protein sequences along with their respective secondary structure; the flowchart of the program has been revealed in Figure 3.

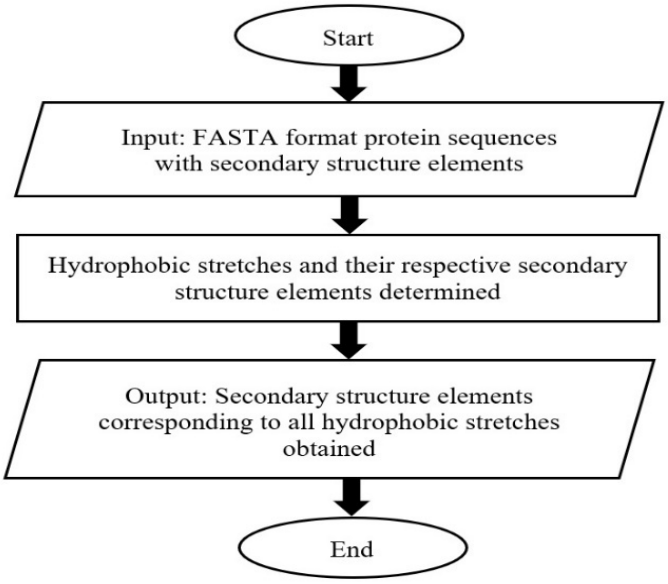


Fig. 3. Flowchart of the program to identify and determine secondary structure elements corresponding to all hydrophobic stretches

The output of the program exhibited specific details about secondary structure elements, i.e., beta sheet (E), alpha helix (H), and coils (C) of each of the residues present in hydrophobic stretches for all the protein sequences of this study. To normalize the results, percentage of hydrophobic stretches on different secondary structure elements was also determined.

3. Results

The provided methodology was employed to examine and compare the hydrophobic stretches and secondary structure of protein sequences from the spike proteins of three coronavirus strains, SARS-CoV-2, SARS-CoV, and MERS-CoV, and their corresponding aligned sequences. which involved the implementation of different computational tools and execution of specifically designed computer programs.

3.1 Multiple Sequence Alignment

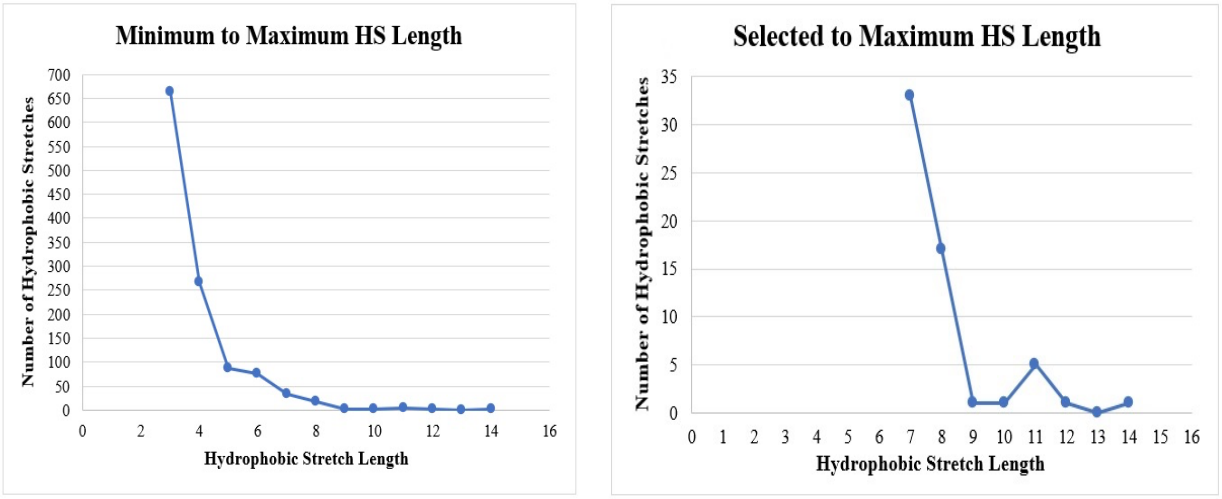
The Bioinformatics tool, ClustalW [33] was employed to determine the sequence similarity and identity among the best-aligned spike protein sequences of SARS-CoV-2, SARS-CoV, and MERS-CoV (SI Table 1). The twelve aligned protein sequences were examined. It was observed that some of the sequences had the highest alignment score with one another while others had the lowest (SI Table 1). The protein sequence 3JCL_A exhibited optimal alignment with the protein sequence 6VSJ_A, achieving a 98% alignment score. This indicates a high degree of similarity in the amino acid sequences between these two IDs, and both PDB IDs are associated with the Murine Hepatitis Virus (MHV-A59). Among the twelve aligned protein sequences, two protein sequences with ID 3JCL_A and 6VSJ_A were aligned with all three spike protein sequences of coronavirus strains. Based on the alignment score, it can be mentioned that the protein sequences which were found in human coronavirus (HCoV-NL63, HCoV-HKU1, and HCoV-OC43) with IDs 7KIP_A, 5I08_A, and 6NZK_A [38] can be considered to be unique as they had a very low alignment score with all other protein sequences in this group.

PDB ID 5X59_A appeared to be the closest relative of PDB IDs 6NB3_A and 5W9H_A with 99% and 98% alignment scores and all were associated to Middle East Respiratory Syndrome (MERS-CoV). The alignment score between protein sequence 6NB3_A and 5W9H_A was 97%. Likewise, the protein sequence of PDB IDs 7EY0_A, 7CN8_A, 7BBH_A, and 7N1Q_A had the alignment score varying in the range from 85% to 95% between each other, and the two PDB IDs, 7EY0_A and 7N1Q, belong to the SARS-CoV-2 and other two PDB IDs, 7CN8_A and 7BBH_A were related to the Pangolin Coronavirus. Interestingly, all four were found to be aligned with MERS CoV. Phylogenetic analysis done by Zheng and Song [39], as well as Verma and Subbarao [20], also revealed the similarity among these protein sequences. The evolutionary relationship of these viruses was also indicated by other researchers which supports this multiple sequence alignment results [19,40,41].

3.2 Determination of Hydrophobic Stretches

On identifying the different hydrophobic stretches (HS) of fifteen protein sequences related to the spike protein of coronavirus, it was observed that the maximum hydrophobic stretch length was fourteen residues long and the minimum length as considered was three (Methodology Section – determination of hydrophobic stretches).

In Figure 4, the frequency plot highlighted the distribution of hydrophobic stretches in all the sequences. As per probability (Figure 4(a)) it is obvious that the presence of consecutive three/four/five/six hydrophobic amino acids would be more frequent compared to the higher length hydrophobic stretches, so the authors applied a second filter here and for further investigation considered hydrophobic stretches whose length was minimum seven residues long (Figure 4(b)). From Figure 4(b), it can be mentioned that the highest number of hydrophobic stretches were seven residues long, followed closely by hydrophobic stretches which were eight residues long.



(a) (b)

Fig. 4. Distribution of Hydrophobic Stretches (HS) (a) Minimum to maximum HS length (b) Selected to maximum HS length

3.3 Determination of Secondary Structure

The secondary structure elements corresponding to the different hydrophobic stretches for all the protein sequences were identified by executing the designed program 'seconhs.cpp'. The distribution of hydrophobic stretches on different types of secondary structure elements were illustrated in Table 1 and Table 2. The maximum number of hydrophobic stretches was identified for the protein sequence 6NZK_A (HCoV-OC43) whereas the minimum number of stretches occurred in two protein sequences, 3JCL_A (MHV-A59) and 5X59_A (MERS-CoV). Notably, all these three sequences were best aligned with SARS-CoV-2 and SARS-CoV coronavirus.

In case of determination of the secondary structure elements (Table 1), the percentage of the beta sheet (E), alpha helix (H), and coil (C) corresponding to the hydrophobic stretches varied from 25.33% to 42.31%, 7.04% to 24.36%, and 18.18% to 34.72% in all the protein sequences, respectively. Also, there were some stretches identified where on a single hydrophobic stretch, different secondary structure elements occurred. The percentage of hydrophobic stretches with combination of both beta sheet and coil varied in the range of 1.28% to 30.68% whereas that of alpha helix and coil was 1.28% to 10.96%, which was comparatively lesser than the range variation for beta sheet and coil combination.

It was observed that a greater number of hydrophobic stretches lay on only one type of secondary structure element, i.e., beta sheet/alpha helix/coil, followed by hydrophobic stretches on combination of beta sheet and coil. The distribution of hydrophobic stretches on combination of alpha helix and coil was comparatively much lower and was quite similar to the percentage of hydrophobic stretches on beta sheet and alpha helix combination. For the miscellaneous secondary structure element on hydrophobic stretches, the percentage was very less, approximately 1% (Table 1). It was interesting to note that there were some protein sequences in which not even a single hydrophobic stretch corresponded to beta sheet and helix combination or the miscellaneous secondary structure combination.

The Protein sequence of 3JCL_A which was common with primary spike protein sequence of SARS-CoV-2, SARS-CoV, and MERS-CoV, had the minimum percentage of alpha helix and comparatively higher percentage of miscellaneous secondary structures among all the protein sequences. Also, it had nearly similar number of hydrophobic stretches as MERS-CoV and nearly similar percentage of beta-sheet elements with SARS-CoV and SARS-CoV-2.

In comparison to spike protein sequences of SARS-CoV-2, SARS-CoV, MERS-CoV, and their aligned sequences, it was observed that the maximum percentage of one type of secondary structure element i.e., beta sheet and alpha helix, and the minimum percentage of combination of beta sheet and coil as well as combination of alpha helix and coil lie on the hydrophobic stretches that belongs to spike protein sequence of SARS-CoV-2. Interestingly, the percentage of the beta sheet was more than alpha helix in the spike protein sequence of SARS-CoV-2. *In silico* results here indicate that increased number of beta sheets were present on hydrophobic stretches of spike protein sequence of SARS-CoV-2. Experimental findings of *D'Arco et al.*, also endorsed the fact that SARS-CoV-2 was found to have significantly large proportion of intermolecular beta sheets [12]. Also, in three primary spike protein sequences of SARS-CoV-2, SARS-CoV, and MERS-CoV, the percentage of secondary structure elements was much higher than the twelve aligned spike protein sequences.

In this work, authors have identified the multiple stretches of consecutive hydrophobic aromatic/aliphatic amino acid residues and determined their respective secondary structure conformation. Hydrophobic stretches predominantly laid on beta sheet even for the primary spike protein sequence of SARS-CoV-2, SARS-CoV, and MERS-CoV. On close inspection, it further revealed that the presence of beta sheet and alpha helix are highest on SARS-CoV-2. Coils can be more easily mutated than beta sheets and alpha helices. Beta sheets exhibit a greater susceptibility to mutations compared to alpha helices., but without any change in their secondary structure conformation [42].

Table 1
Hydrophobic stretches (HS) on secondary structure of primary and aligned protein sequences related to SARS-CoV-2, SARS-CoV, and MERS-CoV

Sr. No.	ID	Percentage of Hydrophobic Stretches lying on							Total Number of HS
		Beta Sheet(E)	Alpha Helix (H)	Coil (C)	Beta Sheet + Alpha Helix	Beta Sheet + Coil	Alpha Helix + Coil	Miscellaneous (Beta Sheet + Alpha Helix + Coil)	
1	UniProt ID P0DTC2	42.31	24.36	30.77	-	1.28	1.28	-	78
2	UniProt ID P59594	40.79	21.05	28.95	-	6.58	2.63	-	76
3	UniProt ID A0A7D5J875	41.67	19.44	34.72	-	2.78	1.39	-	72
4	PDB ID 5I08_A	35.37	10.97	26.83	1.22	15.85	8.54	1.22	82
5	PDB ID 3JCL_A	40.85	7.04	19.72	-	22.53	8.45	1.41	71
6	PDB ID 6VSJ_A	35.14	10.81	20.27	-	24.32	8.11	1.35	74
7	PDB ID 6NZK_A	36.05	16.28	18.6	-	20.93	6.98	1.16	86
8	PDB ID 5X59_A	30.98	9.86	28.17	-	19.72	9.86	1.41	71
9	PDB ID 6NB3_A	25.33	13.33	24	1.33	30.68	4	1.33	75
10	PDB ID 5W9H_A	30.14	10.96	26.02	-	20.55	10.96	1.37	73
11	PDB ID 7KIP_A	35.8	18.52	19.75	-	18.52	6.17	1.24	81
12	PDB ID 7EY0_A	27.63	13.16	23.68	1.32	26.31	6.58	1.32	76

13	PDB ID 7CN8_A	29.49	14.1	24.36	-	25.64	6.41	-	78
14	PDB ID 7BBH_A	29.27	12.19	26.83	1.22	21.95	8.54	-	82
15	PDB ID 7N1Q_A	27.27	18.18	18.18	-	25.97	7.8	1.3	77

Based on the details inferred regarding secondary structure elements (Table 1), the hydrophobic stretches of length seven and above (Result and Discussion section 3.2) for the three primary protein sequences were considered to correlate the hydrophobic amino acid stretches with secondary structure elements, shown in Table 2 and Figure 5 to 7. All the hydrophobic stretches were considered highly hydrophobic as the maximum number of hydrophobic amino acid residues, like leucine, phenylalanine, alanine, and valine were present there [43]. The similar details of hydrophobic stretches of remaining twelve protein sequences showed that beta sheet dominated in majority of the protein sequences, however, alpha helix dominated in the hydrophobic stretches of few protein sequences like 6VSJ_A, 7KIP_A, 7BBH_A, and 7N1Q_A (Table 2).

Table 2
Hydrophobic stretches (HS) corresponding to secondary structure elements of spike protein sequences of SARS-CoV-2, SARS-CoV, MERS-CoV, and their aligned sequences

Sr. No.	HS	Position	Stretch Length	Aromatic Length	Aliphatic Length	Secondary Structure of Respective residue
UniProt ID P0DTC2						
1	MFVFLVLLPLV	01-11	11	2	9	CEEEEECCCHH
2	WFHAIHV	64-70	7	4	3	EEEEECC
3	VPVAIHA	620- 626	7	1	6	CCEEEEC
4	WPWYIWL	1212-1218	7	4	3	HHHHHHH
5	LIAIVMV	1224-1230	7	0	7	HHHHHHH
UniProt ID P59594						
1	MFIFLLFL	01-08	8	3	5	CCCCCCCH
2	WPWYVWL	1194-1200	7	4	3	HHHHHHH
3	LIAIVMV	1206-1212	7	0	7	HHHHHHH
UniProt ID: A0A7D5J875						
1	VFLLMFLL	5-12	8	2	6	HHHCHHCC
2	IYPAFML	143- 149	7	2	5	HCCEEEE
3	AWAAFYVY	309- 316	8	4	4	CCCCEEEE
4	FAAIPFA	967- 973	7	2	5	HCCCCHH
5	WPWYIWL	1295-1301	7	4	3	HHHHHHH
6	LVALALCVFFILCC	1307-1320	14	2	12	HHHHHHHHHHHHHH
PDB ID 5I08_A						
1	WLYFHFY	184- 190	7	6	1	EEEEEEE
2	VFYAYYA	195- 201	7	4	3	EEEEEEE
3	HYYVMPL	221- 227	7	3	4	EEEEEEE
4	VAAMFPPW	951- 958	8	2	6	HHHHCCCC
PDB ID 3JCL_A						
1	AFYFHFY	187- 193	7	6	1	EEEEEEE
2	YYVLPFIC	225- 232	8	3	5	EEEEEEEE
3	AAAMFPPW	919- 926	8	2	6	HHHHCCCC
PDB ID 6VSJ_A						
1	VALVFMVVYI	7-16	10	2	8	HHHHHHHHHH
2	AFYFHFY	208- 214	7	6	1	EEEEEEE
3	YYVLPFIC	246- 253	8	3	5	EEEEEEEE
4	AAAMFPPW	940- 947	8	2	6	HHHHCCCC
5	HHHHHHHH	1268-1275	8	8	0	CCCCCCCC

PDB ID 6NZK_A						
1	YLYFHFY	221- 227	7	6	1	EEEEEEE
2	HYYVMPL	258- 264	7	3	4	EEEEEEE
PDB ID 5X59_A						
1	IYPAFML	126- 132	7	2	5	HCCEEEE
2	AWAAFYVY	292- 299	8	4	4	CCCCEEEE
3	FAAIPFA	950- 956	7	2	5	HCCCCHH
PDB ID 6NB3_A						
1	IYPAFML	157- 163	7	2	5	CCCCEEE
2	AWAAFYVY	323- 330	8	4	4	CCCEEEEE
3	FAAIPFA	981- 987	7	2	5	EEECCHH
4	HHHHHHHH	1352-1359	8	8	0	CCCCCCCC
PDB ID 5W9H_A						
1	VFLLMFLL	5-12	8	2	6	HHHHHHCC
2	IYPAFML	143- 149	7	2	5	HCCEEEE
3	AWAAFYVY	309- 316	8	4	4	CCCCEEEE
4	FAAIPFA	967- 973	7	2	5	HCCCCHH
PDB ID 7KIP_A						
1	LFLILLVLPLA	3-13	11	1	10	CEEHHCCCCHH
2	FYVPAAY	160- 166	7	3	4	EECCCCC
3	YVALPIYY	473- 480	8	3	5	EEECCCEE
4	WPWWVWLII	1295-1303	9	4	5	HHHHHHHHH
5	VVFVVLL	1305-1311	7	1	6	HHHHHHH
6	LLVFCCL	1313-1319	7	1	6	HHHHHHH
PDB ID 7EY0_A						
1	MFVFLVLLPLV	1-11	11	2	9	CCEECCCCHHH
2	WFHAIHV	64-70	7	4	3	EEECCCC
3	VPVAIHA	617- 623	7	1	6	CCEEEEC
4	HHHHHHHH	1247-1254	8	8	0	CCCCCCCC
PDB ID 7CN8_A						
1	MFVFLFVLPLV	1-11	11	3	8	CEEEEEECCCC
2	VPMAIHA	618- 624	7	1	6	CCCEEEC
3	HHHHHHHH	1247-1254	8	8	0	CCCCCCCC
PDB ID 7BBH_A						
1	MLFFFFLHFALV	1-12	12	6	6	CCHHHHHHHHHH
2	VPVAIHA	616- 622	7	1	6	CCEEEEC
PDB ID 7N1Q_A						
1	MFVFLVLLPLV	1-11	11	2	9	CCEECCCHHHH
2	WFHAIHV	64-70	7	4	3	EEECEEC
3	VPVAIHA	617- 623	7	1	6	CCCEEEC
4	WPWYIWL	1209-1215	7	4	3	HHHHHHH
5	LIAIVMV	1221-1227	7	0	7	HHHHHHH

Based on the results given in Table 1, it was observed that the hydrophobic stretches with beta sheet (E) were dominant for all the three primary spike protein sequences of SARS-CoV-2, SARS-CoV, and MERS-CoV. The authors were interested to closely investigate the hydrophobic stretches of length greater than equal to the chosen length of the hydrophobic stretches i.e., seven amino acid length (Table 2). On ChimeraX software [44-46], the three- dimensional structure of these three primary spike protein sequences was investigated and the hydrophobic stretches were highlighted in blue (Figures 5 to 7).

In spike protein sequence of SARS-CoV-2, HS 4, and HS 5 whereas in SARS-CoV, HS 2 and HS 3, and in MERS-CoV, HS 5 and HS 6 corresponded to the alpha helix element of the secondary structure. Interestingly the HS 5 of SARS-CoV-2 and HS 3 of SARS-CoV spike protein sequence were identical both in terms of amino acid content and secondary structure elements; however, positional

occurrence of the respective HS was different in the two protein sequences. A similar result was identified for HS 4 of SARS-CoV-2 with respect to HS 5 of MERS-CoV. Comparing the HS 4 of SARS-CoV-2 with HS 2 of SARS-CoV, it was noticed that in the otherwise identical stretch, the fifth residue of this hydrophobic stretch has been replaced. This finding indicates that with evolution of the spike protein of coronavirus, replacement of this aliphatic hydrophobic amino acid has possibly occurred. It can be said that a lower scale valued hydrophobic aliphatic amino acid, valine has been replaced by a higher-scale valued isoleucine [29]. Metsuki *et al.*, and He *et al.*, suggested that a virus can resist neutralizing antibodies by changing just one amino acid [47,48]. This possibly can be a reason for this aliphatic hydrophobic replacement as identified in this study.

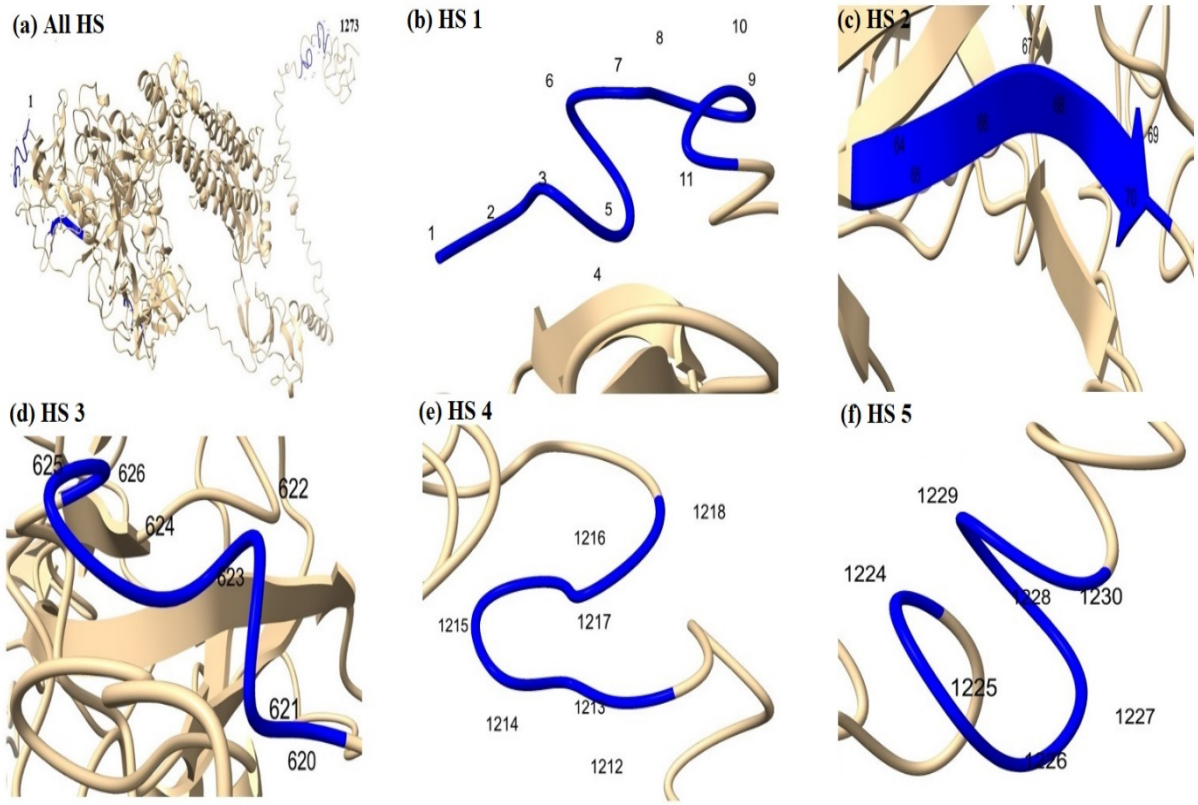


Fig. 5. Hydrophobic stretches (HS) of spike protein sequence of SARS-CoV-2 (a) All hydrophobic stretches (HS); (b)-(f) Specific hydrophobic stretches (HS 1 to HS 5)

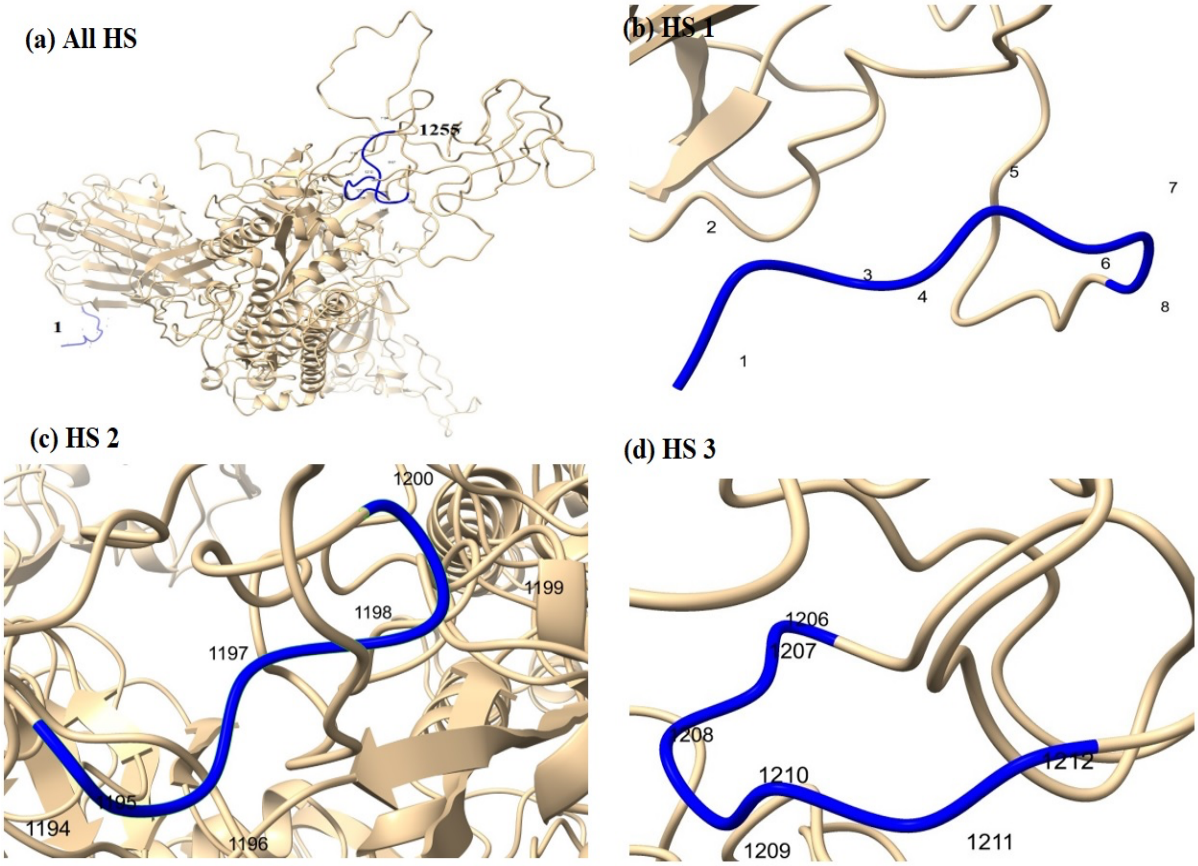


Fig. 6. Hydrophobic Stretches (HS) of spike protein sequence of SARS-CoV (a) All hydrophobic stretches (HS); (b)-(d) Specific hydrophobic stretches (HS 1 to HS 3)

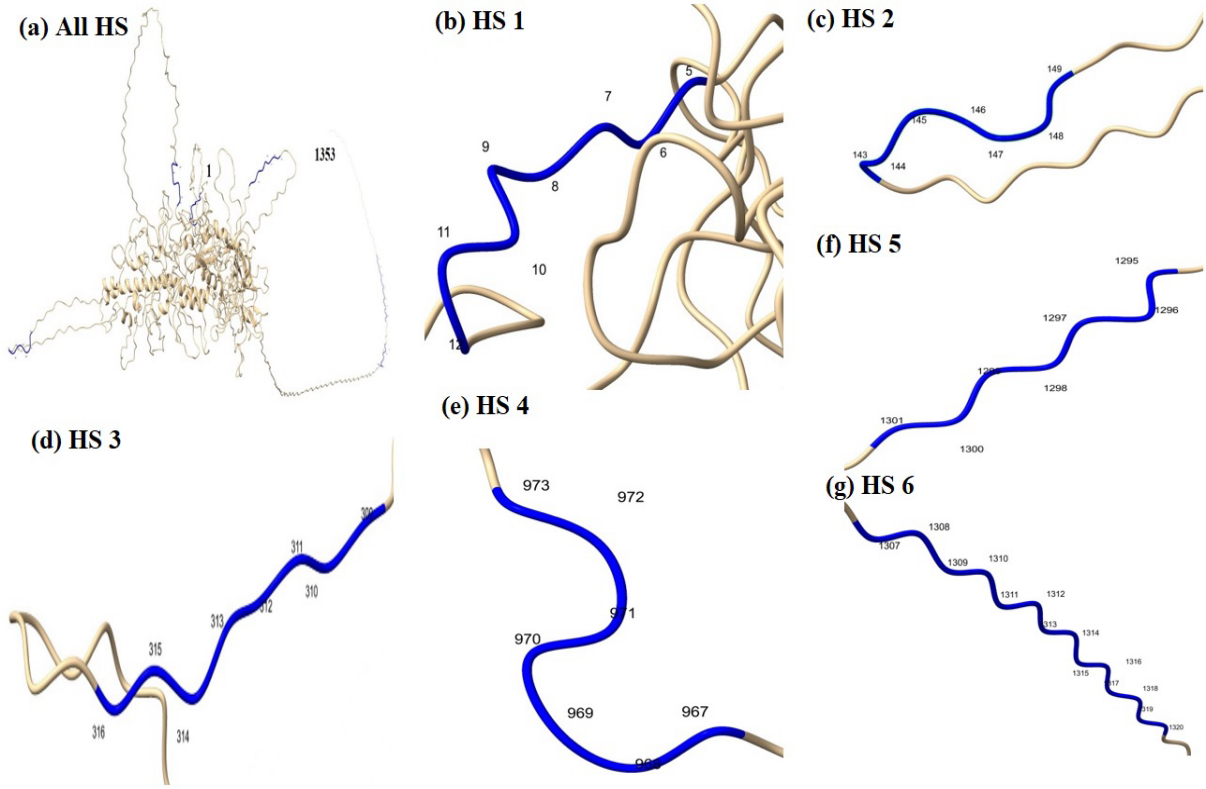


Fig. 7. Hydrophobic Stretches (HS) of spike protein of MERS-CoV; (a) All hydrophobic stretches (HS); (b) – (g) Specific hydrophobic stretches (HS 1 to HS 6)

In the hydrophobic stretches compared, it was recognized that the total length of hydrophobic stretch was approximately in same range of seven to eight amino acid length. Moreover, aliphatic length was higher compared to aromatic length for all hydrophobic stretches; this observation is endorsed by the fact that probability of occurrence of seven aliphatic amino acid residues was more than that of four aromatic amino acid residues [29]. Results of this investigation get endorsed by the findings of Xia who stated the conserved nature of hydrophobic residues, in SARS-CoV-2 in general, and also mentioned that it is preferable to develop antiviral medications or vaccines that only target highly conserved areas [49], thus, highlighting the significance of this result. Among the three primary coronavirus sequences, the hydrophobic stretch with maximum length had been identified in MERS-CoV.

In general, hydrophobic residues have large R-group and their specific geometry is also critical for a stable protein structure. If such hydrophobic residues appear consecutively in long stretches, it might lead towards an unstable structure [5]. Findings here indicate that even for spike protein sequences, large hydrophobic stretches are very rare [25] whereas frequency of hydrophobic stretches with lower number of amino acid residues is more. The fact that too many consecutive amino acid residues may create instability but at the same time their presence in small stretches would enhance the stability is well illustrated in this work.

In this paper, authors focussed on the secondary structure elements of the hydrophobic stretches and interestingly found the propensity of beta sheet highest on hydrophobic stretches of SARS-CoV-2 and closely followed by MERS-CoV and SARS-CoV. Using Infrared Vibrational spectroscopy technique D'Arco *et al.*, also showed the propensity of beta sheet following an increasing trend from MERS-CoV to SARS-CoV-2 through SARS-CoV [12]. Through *in silico* investigation of the spike protein sequence of SARS-CoV-2, SARS-CoV and MERS-CoV, the authors identified that amino acids in positions 1212 – 1217 lie on helical region; experimental work done by Lie *et al.*, using NMR technique endorsed that the same stretch lies on helical region [11]. Estimation of the potential effects of consecutive hydrophobic amino acid residues in spike protein and their correlation with secondary structure elements may provide significant lead towards computational life sciences and drug designing.

4. Conclusions

In this paper, stretches of consecutive hydrophobic residues for spike protein sequences of coronavirus strains SARS-CoV-2, SARS-CoV, MERS-CoV, and their aligned sequences have been identified, along with their corresponding secondary structure elements. The propensity of hydrophobic stretches on a particular type of secondary structure was determined and conserved hydrophobic stretches were also identified. It can be inferred that the distribution of amino acid residues and secondary structure within hydrophobic stretches is distinctive in SARS-CoV-2 when compared to MERS-CoV and SARS-CoV. The spike protein sequence of SARS-CoV-2 presents a higher percentage of hydrophobic stretches on beta sheet, thus, indicating a more stable protein structure. Quantitative investigation of consecutive hydrophobic residues will give an insight into the protein structure at molecular level and its stability. This might facilitate further investigation regarding pathogenicity, evolution, and transmission of these viruses, thus helping in drug development.

Acknowledgment
This research was not funded by any grant. The authors acknowledge the availability of protein sequence data from authentic websites. The authors thank Anu Khanna for technical support.

References

[1] Van Dijk, Erik, Arlo Hoogeveen, and Sanne Abeln. "The hydrophobic temperature dependence of amino acids directly calculated from protein structures." *PLoS computational biology* 11, no. 5 (2015): e1004277. https://doi.org/10.1371/journal.pcbi.1004277

[2] Sarkar, Aurijit, and Glen E. Kellogg. "Hydrophobicity-shake flasks, protein folding and drug discovery." *Current topics in medicinal chemistry* 10, no. 1 (2010): 67-83. https://doi.org/10.2174/156802610790232233

[3] Simm, Stefan, Jens Einloft, Oliver Mirus, and Enrico Schleiff. "50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification." *Biological research* 49 (2016): 1-19. https://doi.org/10.1186/s40659-016-0092-5

[4] Aydin, Halil, Dina Al-Khooly, and Jeffrey E. Lee. "Influence of hydrophobic and electrostatic residues on SARS-coronavirus S2 protein stability: Insights into mechanisms of general viral fusion and inhibitor design." *Protein Science* 23, no. 5 (2014): 603-617. https://doi.org/10.1002/pro.2442

[5] Nelson, David L., Albert L. Lehninger, and Michael M. Cox. *Lehninger principles of biochemistry*. Macmillan, 2008.

[6] Wang, Hai-Jing, Xue-Kui Xi, Alfred Kleinhammes, and Yue Wu. "Temperature-induced hydrophobic-hydrophilic transition observed by water adsorption." *Science* 322, no. 5898 (2008): 80-83. https://doi.org/10.1126/science.1162412

[7] Zhang, Buzhong, Jinyan Li, and Qiang Lü. "Prediction of 8-state protein secondary structures by a novel deep learning architecture." *BMC bioinformatics* 19 (2018): 1-13. https://doi.org/10.1186/s12859-018-2280-5

[8] Nguyen, Thanh Thi, Pubudu N. Pathirana, Thin Nguyen, Quoc Viet Hung Nguyen, Asim Bhatti, Dinh C. Nguyen, Dung Tien Nguyen, Ngoc Duy Nguyen, Douglas Creighton, and Mohamed Abdelrazek. "Genomic mutations and changes in protein secondary structure and solvent accessibility of SARS-CoV-2 (COVID-19 virus)." *Scientific Reports* 11, no. 1 (2021): 3487. https://doi.org/10.1038/s41598-021-83105-3

[9] Chaudhary N, Saini S. "A Progress on Protein Structure Prediction using Various Soft Computing Techniques." Computer Science and Information Technology Trends, Academy and Industry Research Collaboration Center (AIRCC), (2022): 113–30. https://doi.org/10.5121/csit.2022.121410

[10] Kurgan, Lukasz A., Wojciech Stach, and Jishou Ruan. "Novel scales based on hydrophobicity indices for secondary protein structure." *Journal of theoretical biology* 248, no. 2 (2007): 354-366. https://doi.org/10.1016/j.jtbi.2007.05.017

[11] Li, Qingxin, Qiwei Huang, and Congbao Kang. "Secondary structures of the transmembrane domain of SARS-CoV-2 spike protein in detergent micelles." *International Journal of Molecular Sciences* 23, no. 3 (2022): 1040. https://doi.org/10.3390/ijms23031040

[12] D'Arco, Annalisa, Marta Di Fabrizio, Tiziana Mancini, Rosanna Mosetti, Salvatore Macis, Giovanna Tranfo, Giancarlo Della Ventura, Augusto Marcelli, Massimo Petrarca, and Stefano Lupi. "Secondary Structures of MERS-CoV, SARS-CoV, and SARS-CoV-2 Spike Proteins Revealed by Infrared Vibrational Spectroscopy." *International Journal of Molecular Sciences* 24, no. 11 (2023): 9550. https://doi.org/10.3390/ijms24119550

[13] Chang, Tai-Jay, De-Ming Yang, Mong-Lien Wang, Kung-How Liang, Ping-Hsing Tsai, Shih-hwa Chiou, Ta-Hsien Lin, and Chin-Tien Wang. "Genomic analysis and comparative multiple sequences of SARS-CoV2." *Journal of the Chinese Medical Association* 83, no. 6 (2020): 537-543. https://doi.org/10.1097/JCMA.0000000000000335

[14] Rahim, Nasimah, Siti Zalita Talib, Nur Ainun Mokhtar, Nurulbahiyah Ahmad Khairudin, and Ragheed Hussam Yousif. "In-Silico Search Analysis of Potential Inhibitors for 3-Chymotrypsin-Like Protease Of Sars-Cov-2 (Covid-19)." *Journal of Research in Nanoscience and Nanotechnology* 4, no. 1 (2021): 49-56. https://doi.org/10.37934/jrnn.4.1.4956

[15] Fardhyanti, Dewi Selvia, Maharani Kusumaningrum, Nadya Alfa Cahaya Imani, Junaidah Jai, Nurul Asyikin Md Zaki, Ririn Andriyani, and Melinia Rahmahani Putri. "The Temperature Effect of Madeira Vine (Anredera Cordifolia) Leaf Oil Extraction and Its Characterization as An Additive in Health Supplement Product." *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences* 97, no. 2 (2022): 1-7. https://doi.org/10.37934/arfmts.97.2.17

[16] Abd Rahman, Muhammad Faqhrurrazi, Nor Zelawati Asmuin, Ishkrizat Taib, Juwaidi Nazar, Azizan Ismail, and Riyadhthusollehan Khairulfuaad. "Investigate flow characteristics of metered-dose inhaler (MDI) disposable inhaler spacer (AeroCup) for COVID-19 patient by using computational fluid dynamic (CFD)." *CFD Letters* 12, no. 12 (2020): 63-74. https://doi.org/10.37934/cfdl.12.12.6374

[17] Sethu, Vasanthi, Peck Loo Kiew, Swee Pin Yeap, and Lian See Tan. "Time to Get Serious about Sustainable Water Management." *Progress in Energy and Environment* (2020): 13-15.

[18] Nor, Siti Rohani Mohd, Adina Najwa Kamarudin, and Nurul Aini Jaafar. "Comparison on the Student's Performances during Physical and Online Learning in Financial Mathematics Course." *International Journal of Advanced Research in Future Ready Learning and Education* 28, no. 1 (2022): 1-8.

[19] Hatmal, Ma'mon M., Walhan Alshaer, Mohammad Al Al-Hatamleh, Malik Hatmal, Othman Smadi, Mutasem O. Taha, Ayman J. Oweida, Jennifer C. Boer, Rohimah Mohamud, and Magdalena Plebanski. "Comprehensive

structural and molecular comparison of spike proteins of SARS-CoV-2, SARS-CoV and MERS-CoV, and their interactions with ACE2." *Cells* 9, no. 12 (2020): 2638. https://doi.org/10.3390/cells9122638

[20] Verma, Jyoti, and Naidu Subbarao. "A comparative study of human betacoronavirus spike proteins: structure, function and therapeutics." *Archives of Virology* 166, no. 3 (2021): 697-714. https://doi.org/10.1007/s00705-021-04961-y

[21] Jain, Mukul, Nil Patil, Darshil Gor, Mohit Kumar Sharma, Neha Goel, and Prashant Kaushik. "Proteomic Approach for Comparative Analysis of the Spike Protein of SARS-CoV-2 Omicron (B. 1.1. 529) Variant and Other Pango Lineages." *Proteomes* 10, no. 4 (2022): 34. https://doi.org/10.3390/proteomes10040034

[22] Shekhawat, Uma, and Anindita Roy Chowdhury. "Computational and comparative investigation of hydrophobic profile of spike protein of SARS-CoV-2 and SARS-CoV." *Journal of Biological Physics* 48, no. 4 (2022): 399-414. https://doi.org/10.1007/s10867-022-09615-x

[23] Pelton, John T., and Larry R. McLean. "Spectroscopic methods for analysis of protein secondary structure." *Analytical biochemistry* 277, no. 2 (2000): 167-176. https://doi.org/10.1006/abio.1999.4320

[24] Argyrou, Agathi. "The Misfolding of Proteins." *GeNeDis 2018: Genetics and Neurodegeneration* (2020): 249-254. https://doi.org/10.1007/978-3-030-32633-3_33

[25] Zhu, Chaogeng, Guiyun He, Qinqin Yin, Lin Zeng, Xiangli Ye, Yongzhong Shi, and Wei Xu. "Molecular biology of the SARs-CoV-2 spike protein: A review of current knowledge." *Journal of medical virology* 93, no. 10 (2021): 5729-5741. https://doi.org/10.1002/jmv.27132

[26] Chowdhury, Anindita Roy, H. G. Nagendra, and Alpana Seal. "Correlation among hydrophobic aromatic and aliphatic residues in the six enzyme classes." *International Journal of Computational Biology and Drug Design* 13, no. 2 (2020): 209-223. https://doi.org/10.1504/IJCBDD.2020.10029441

[27] Chowdhury, A. R., A. Seal, and H. G. Nagendra. "Computational analysis of hydrophobicity across six enzyme classes revealing relative contribution of aliphatic and aromatic residues." *Biotechnol. Bioinf. Bioeng* 1 (2011): 83-91.

[28] Chowdhury AR, Kothari A. "Hydrophobic Mapping of Chlorobium Tepidum, The Energy Generating Bacteria." Journal of Harmonized Research in Applied Science no. 7 (2019): 98–106. https://doi.org/10.30876/JOHR.7.3.2019.98-106

[29] Fauchere, Jean-Luc, and Vladimir Pliska. "Hydrophobic Parameters Pi of Amino-Acid Side Chains from The Partitioning Of N-Acetyl-Amino Amides." (1983).

[30] NCBI, NIH. "Protein BLAST: search protein databases using a protein query." (2019).

[31] Shekhawat, Uma, and Anindita Roy Chowdhury. "Computational and comparative investigation of hydrophobic profile of spike protein of SARS-CoV-2 and SARS-CoV." *Journal of Biological Physics* 48, no. 4 (2022): 399-414. https://doi.org/10.1007/s10867-022-09615-x

[32] UniProt. https://www.uniprot.org

[33] Multiple Sequence Alignment - CLUSTALW. https://www.genome.jp/tools-bin/clustalw

[34] Hung, Jui-Hung, and Zhiping Weng. "Sequence alignment and homology search with BLAST and ClustalW." *Cold Spring Harbor Protocols* 2016, no. 11 (2016): pdb-prot093088. https://doi.org/10.1101/pdb.prot093088

[35] BhageerathH+ | SCFBio 2014. http://scfbio-iitd.res.in/bhageerathH+

[36] Kaushik R, Singh A, Dasgupta D, Pathak A, Shekhar S, Jayaram B. "BhageerathH+: A hybrid methodology based software suite for protein tertiary structure prediction." *CASP12 Proceedings*, (2016): 25-26

[37] Jayaram, Bhyravabhotla, Priyanka Dhingra, Avinash Mishra, Rahul Kaushik, Goutam Mukherjee, Ankita Singh, and Shashank Shekhar. "Bhageerath-H: a homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins." *BMC bioinformatics* 15 (2014): 1-12. https://doi.org/10.1186/1471-2105-15-S16-S7

[38] Data, RCSB Protein. "RCSB PDB: Homepage." (2021).

[39] Zheng, Ming, and Lun Song. "Novel antibody epitopes dominate the antigenicity of spike glycoprotein in SARS-CoV-2 compared to SARS-CoV." *Cellular & molecular immunology* 17, no. 5 (2020): 536-538. https://doi.org/10.1038/s41423-020-0385-z

[40] Lu, Roujian, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang *et al.,* "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding." *The lancet* 395, no. 10224 (2020): 565-574. https://doi.org/10.1016/S0140-6736(20)30251-8

[41] Robson, Barry. "Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus." *Computers in biology and medicine* 119 (2020): 103670. https://doi.org/10.1016/j.compbiomed.2020.103670

[42] Abrusán, György, and Joseph A. Marsh. "Alpha helices are more robust to mutations than beta strands." *PLoS computational biology* 12, no. 12 (2016): e1005242. https://doi.org/10.1371/journal.pcbi.1005242

[43] Higgs, Trent, Bela Stantic, Md Tamjidul Hoque, and Abdul Sattar. "Hydrophobic-hydrophilic forces and their effects on protein structural similarity." In *Suppl. Conf. Proc*, (2008): 1-12.

[44] UCSF ChimeraX Home Page. https://www.rbvi.ucsf.edu/chimerax
[45] Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Elaine C. Meng, Gregory S. Couch, Tristan I. Croll, John H. Morris, and Thomas E. Ferrin. "UCSF ChimeraX: Structure visualization for researchers, educators, and developers." *Protein science* 30, no. 1 (2021): 70-82. https://doi.org/10.1002/pro.3943
[46] Goddard, Thomas D., Conrad C. Huang, Elaine C. Meng, Eric F. Pettersen, Gregory S. Couch, John H. Morris, and Thomas E. Ferrin. "UCSF ChimeraX: Meeting modern challenges in visualization and analysis." *Protein science* 27, no. 1 (2018): 14-25. https://doi.org/10.1002/pro.3235
[47] Mitsuki, Yu-ya, Kazuo Ohnishi, Hirotaka Takagi, Masamichi Oshima, Takuya Yamamoto, Fuminori Mizukoshi, Kazutaka Terahara *et al.,* "A single amino acid substitution in the S1 and S2 Spike protein domains determines the neutralization escape phenotype of SARS-CoV." *Microbes and infection* 10, no. 8 (2008): 908-915. https://doi.org/10.1016/j.micinf.2008.05.009
[48] He, Yuxian, Jingjing Li, and Shibo Jiang. "A single amino acid substitution (R441A) in the receptor-binding domain of SARS coronavirus spike protein disrupts the antigenic structure and binding activity." *Biochemical and biophysical research communications* 344, no. 1 (2006): 106-113. https://doi.org/10.1016/j.bbrc.2006.03.139
[49] Xia, Xuhua. "Domains and functions of spike protein in Sars-Cov-2 in the context of vaccine design." *Viruses* 13, no. 1 (2021): 109. https://doi.org/10.3390/v13010109