doi:10.20944/preprints202008.0621.v1

World-wide Sequence Variant and Non-synonymous Amino Acid Substitution Signature in SARS-COV-2 Structural Proteins

Jayanta Kumar Das[a], Swarup Roy[b,*],

[a] *Department of Pediatrics, Johns Hopkins University School of Medicine, Maryland, USA*
[b] *Network Reconstruction & Analysis (NetRA) Lab, Department of Computer Applications, Sikkim University, Gangtok, India*

Abstract

Like other viruses, SARS-COV-2 too mutating and thus creating divergent variants across the world. Protein sequence variation occurs due to non-synonymous single-nucleotide polymorphism (SNP) that alter the amino acid. Amino acid substitutions on homooligomer interfaces may change the structure of the protein and hence alter the regular or known functional activities of a viral protein. Studies reveal that even a single point mutation in virus protein can significantly change their biology, leads to peculiar pathogenic properties. Therefore, an in-depth investigation of the amino acid substitution in the genomic signature of a protein is highly essential for the rapidly evolving virus-like SARS-COV-2. Investigation of world-wide and country-specific substitution features may be crucial and highly essential to decipher pathogenicity. These might be also helpful to precise structure prediction and identification of possible therapeutic targets for effective drug design.

We perform extensive analysis towards highlighting and characterizing the amino acid substitution signature occurs in the four structural proteins (Spike-S, Nucleocapsid-N, Membrane-M, Envelope-E) of SARS-COV-2. We use a total of 9587 viral sequences reported from 49 different countries across

[*] Corresponding Author
Email addresses: jdas4@jhmi.edu (Jayanta Kumar Das), sroy01@cus.ac.in (Swarup Roy)

the globe. In this study, we try to study the amino acid substitution patterns and its impact on change in biochemical properties, thereby possible changes in protein structures. We perform the following analysis: a) isolating and grouping variants we considered, for different protein sequences; b) identifying amino acid substitution type that are frequently and rarely occurring and reporting their location within the sequence; c) change in chemical properties due to amino acid substitution; and f) highlight country-specific divergent variation and substitution signature.

In terms of mutational changes, E and M proteins are relatively stable than N and S proteins. A significant quantity of variations is observed in spike (S) proteins. Our study further reveals an interesting fact that the substitution location is random in N protein, whereas the substitution sites in M protein is less varying and almost stable. Substitutions specific to active sub-domains in S and N proteins reveals that sub-domains like Heptapeptide Repeat (HR2), Fusion peptides (FP), and Transmembrane (TM), which are involved in cellular membrane fusion and entry of the virus into the host cells, are significantly mutated. Majority of the substitutions leads to change in biochemical properties (side chain and hydropathy) of amino acid. A good number of exclusive variants are found specific to a particular country. We strongly believe that the current findings will be helpful for protein structure analysis of viral structural proteins and antiviral drug discovery.

Keywords: Mutation, Amino acid substitution, Structural proteins, Biochemical properties, Functional sub-domains.

1. Introduction

Recent world pandemic due to SARS-COV-2 creates a havoc and lethal for the entire mankind. With the passage of time, SARS-COV-2 infecting different parts of the globe with varying intensity. Though the rate of mutation in SARS-COV-2 observed to be relatively low, even then a good number of strains are reported across the globe. Genome of any viruses are differed by several mutations in their genome. Mutations in the nucleotide sequence can prevent binding PCR primers to target sequences [1]. Different biological properties such as pathogenicity, tissue tropism or host range can happen diversely in closely related strains due to variation [2, 3]. Couple of studies reveal that mutation can trigger on enzyme motion in dihydrofolate reductase [4], effect in three-dimensional structure, stability and redox potential [5], hydrophobic effect [6], functional diversity, even for the structurally similar proteins [7, 8] and pathogenic mutation [9].

The genetic diversity of SARS-COV-2 strains may impact on molecular functionality of the protein [10]. Mutation and sequence variant analysis is crucial for understanding the disease pathogenesis and structural changes in SARS-COV-2 proteins, which ultimately helps in designing more stable small molecules that binds viral proteins.

Mutation is the key factor that trigger a virus switch hosts including human [11]. SARS-COV-2 genome shows different degree of similarity with other coronaviruses. SARS-COV-2 genome is highly similar with SARS-related coronaviruses [1] derived from Pangolin [12] or Bat [13, 14]. A minor dissimilarity [15] might leads to the variation in functionality of SARS-COV-2 protein with other class of coronaviruses. Scientists observing variants in novel coronavirus strains due to mutation, insertion or deletion reported from different geographical regions [16, 17, 18].

At least 26 protein-coding genes are available in each SARS-COV-2 genome that encodes mainly three classes of proteins, which are nonstructural (Nsp1-Nsp16), structural (Spike glycoprotein-S, Envelope-E, Membrane-M and Nucleocapsid - N), and several accessory protein chains [19, 20]. The two-third of the complete genome is at the $5'$ site that encoding the nonstructural proteins, and one-third are at the $3'$ side that encoding structural and accessory proteins [19]. SARS-COV-2 proteins play a diverse functional role in which structural proteins are highly essential. This is mainly because of two reasons: a) proteins from this class initiate host jumping mechanism, and b) high mutational rate (particularly in Spike and Nucleocapsid protein) compared to other categories. For example, the envelope protein promotes viral assembly and release [21]. The Spike protein responsible for the occurrence of spikes on the viral surface that binds to host receptors [21] and it is mainly responsible for receptor recognition, cell attachment, and fusion during viral infection [22, 23, 24]. The Nucleocapsid protein capable of self-association through a C-terminal [25, 26] and it activates the expression of cyclooxygenase-2 [27]. The roles of M protein include promoting membrane fusion, regulating viral replication, packing genomic RNA into viral particles, interaction with other proteins [28, 29]. Both the spike and nucleocapsid proteins are in the main target for antibody detection [30]. Importantly, many of these sequence vari-

[1] https://www.ecohealthalliance.org/2020/01/

ability features of structural proteins are yet to be investigated thoroughly.

Studies on three severe class of coronaviruses have shown that hydrophobic interaction in receptor binding motif for SARS-COV and SARS-COV-2 could allow the spike protein for zoonotic transmission [31]. Further, insertion of an extra charged amino acid residue in SARS-COV-2 receptor-binding domain, creates a larger hydrophobic surface that underlies the higher binding affinity of SARS-COV-2 to ACE2 compared to SARS-COV [31] and that might allow more flexibility for zoonotic transmission. In comparison to Bat and Pangolin coronavirus, SARS-CoV-2 sequence seems to possess specific modifications and characteristics than other SARS-COV viruses [32]. For instance, in case of homologous coronavirus proteins, positively charged amino acid (Arg69) replaces negatively and neutrally charged Glu or Gln residues. Moreover, a deletion specific to SARS-COV-2 proteins flanks this position. In a different studies have highlighted the impact of the non-synonymous substitutions, that change biochemical properties of amino acids, are highly crucial for protein stability, binding with receptors, substrate specificity, affinity of amino acid change [33, 34, 35, 36, 37, 38]. Therefore, amino acid insertion, deletion and even substitution playing significant role towards attaching with receptor proteins that could impact on the binding affinity of SARS-COV-2 protein. Several other studies on SARS-COV-2 have shown that a mutation can trigger protein structure alteration, dynamics and function while binding with human receptor protein (ACE2) [39, 40, 41, 42]. Therefore, these studies are important to understand the clinical presentation and spread of the disease, and also useful for antiviral drug design [43, 44].

In this study, we focus on sequence variability of worldwide SARS-COV-2 genome, particularly four structural proteins. For each protein, we cluster the genome by sequence variant. Then we identify and localize the amino acid substitution in each variant. Substitutions are then quantified and categorised according to their type and physico-chemical properties of amino acids. We then report country-specific unique variant and substitution type in each structural protein.

2. Methods and Materials

To the best of our knowledge no prior research analyse different SARS-COV-2 strains collected from 49 different countries across the global to understand the worldwide sequence variant due to mutation. In this section we report the data collected and various exploratory analysis performed during the study.

2.1. Sequence dataset

We collect around 9,587 SARS-COV-2 complete genomic (nucleotide) sequences of length $\approx 28kb$ from NCBI database reported till 15th July 2020. We exclude the incomplete and noisy sequences in each of the structural protein in our study and finally, we obtain 9058, 8954, 8271 and 7009 number of sequences for Spike (S), Envelope (E), Membrane (M), Nucleocapsid (N) protein, respectively.

2.2. Clustering analysis

Collected sequences from different patient samples are not unique. Many sequences are identical (exactly similar organizing pattern). We try to identify similar sequences and cluster them accordingly. To do that we perform string matching technique, which is applied on gene-wise collected whole list of sequences datasets and clustering is performed.

2.3. Multiple sequence alignment

The unique list of clustered sequences (one representative from each cluster) are then subject to multiple sequence alignment (MSA) mainly to observe the sequences with one or more inserted and deleted amino acid residues. We choose reference sequence which is the first reported sequence from Wuhan city of china (2019-nCoV/USA-WA1/2020, Accession number: $MN985325/NC\_045512$), E-protein (Accession number: QHO60596.1); M-protein (Accession number: QHO60597.1); N-protein (Accession number: QHO60601.1); S-protein (Accession number: QHO60594.1). We use MEGAX [45] for MSA. We obtain very few sequences (09 Spike and 03 Nucleocapsid proteins) which are observed either inserted or deleted some amino acid residues (*Supplementary-A (Figure S1 and S2)*). We exclude the these few sequences and all the remaining sequences (only with substitution type as compared with the taken reference sequence) are considered for subsequent analysis.

2.4. Amino acid substitution identification

Amino acids are macromolecule, which undergoing condensation reactions to form a protein through peptide bond. The amino acid substitutions in proteins is a common phenomenon in all living species and can vary across

doi:10.20944/preprints202008.0621.v1

protein families [46]. Amino acid substitution for a sequence is the change or alteration of one or more amino acids in some positions while compared with a reference sequence. The amino acid substitution (or non-synonymous mutation) is a common phenomenon in viruses genome and the rate of substitution is varying that mainly depend on the protein's expression level, functional category, metabolic costs, hydrophobicity and electrostatic, physicochemical properties, annotated active or binding site [47, 48, 49, 38].

To understand and observe any pattern during amino acid substitutions in the collected strains, we compare each and every aligned sequences with the reference sequence and report position-wise substitutions. In our study, we try to identify all possible substitutions from one amino acid to other ($20 \times 20 - 20 = 380$) for twenty (20) amino acids, without considering synonymous substitutions and categorizing them based on similar substitution patterns.

2.5. Investigating the change in biochemical properties

Amino acid substitutions (or insertion and deletion of amino acid residue) due to non-synonymous mutation may change the biochemical properties of the protein in some specific domain. As we discussed in introductory section, altering the biochemical properties of amino acid in some functional domain of proteins might have essential role in binding affinity, protein stability, substrate specificity of the protein.

We categorize amino acid substitutions (obtained from the last step) based on the change in the chemical properties. Here, we consider two kinds of broad groupings based on the biochemical properties of amino acid. One is eight chemical sub-groups based on side-chain structure [50], and the other is three Hydropathy classes of amino acids [51]. The sub-groups in each category are as follows:

- **Side-Chain based classes:** According to this grouping 20 amino acids are clustered as Acidic (D, E), Basic (R, H, K), Aromatic (F, W, Y), Aliphatic (A, G, I, L, V), Cyclic (P), Sulfur (C, M), Hydroxyl (S, T), and Amide (N, Q).

- **Hydropathy based classes:** Three such groups include Hydrophobic (A, C, I, L, M, F, W, V), Neutral (G, H, P, S, T, Y) and Hydrophilic (R, N, D, Q, E, K).

Here, we try to investigate all the substitutions observed in the last section according to the category of change of biochemical properties of amino acid. Our major goal is to highlight what kind of biochemical properties are majorly changed (quantitatively) due to substitution.

2.6. Functional domains of SARS-COV-2 structural proteins

The SARS-COV-2 structural protein (particularly S and N) encompasses several sub-domain responsible for specific functional activities. Few of them are:

- **Transmembrane (TM)** is a stretches of amino acids responsible for viral entry [52, 53].
- **Heptapeptide Repeat 1-2 (HR1, HR2)** are responsible for virus fusion [54].
- **Receptor-Binding domain (RBD)** is mainly responsible for binding of the virus to the receptor protein [53].
- **N-terminal (NTD) and C-terminal domain (CTD)** are two main RNA binding domains in SARS-COV N protein [26]. Both of them function as a receptor-binding entity. CTD recognizes the receptor and NTD engages the receptor [55].
- **Fusion peptides (FP)** are created fusing using two or more genes playing different functional roles. The FP play an important role in fusion of viral envelope with host cellular membranes [56].

The typical length of SARS-COV-2 spike (S) protein domain is 1273 amino acids. It consists primarily of three units: a) a signal peptide (amino acids 1–13) located at the N-terminus; b) the S1 subunit (14–685 residues), which is consisting of N-terminal domain (14–305 residues) and a receptor-binding domain (RBD, 319–541 residues); and c) the S2 subunit (686–1273 residues), which is consisting of the fusion peptide (FP) (788–806 residues), heptapeptide repeat sequence 1 (HR1) (912–984 residues), HR2 (1163–1213 residues), Transmembrane (TM) domain (1213–1237 residues), and cytoplasm domain (1237–1273 residues) [57].

Usually SARS-COV-2 nucleocapsid (N) protein domain consists of 419 amino acids. SARS-COV-2 N protein contains two distinct RNA-binding domains: the N-terminal domain (NTD, 44-179 residues) and the C-terminal domain (CTD, 247-363 residues) [58]. These two domains are linked by a poorly structured linkage region (LKR), and N-tail and C-tail domain at the beginning and end of the protein domain.

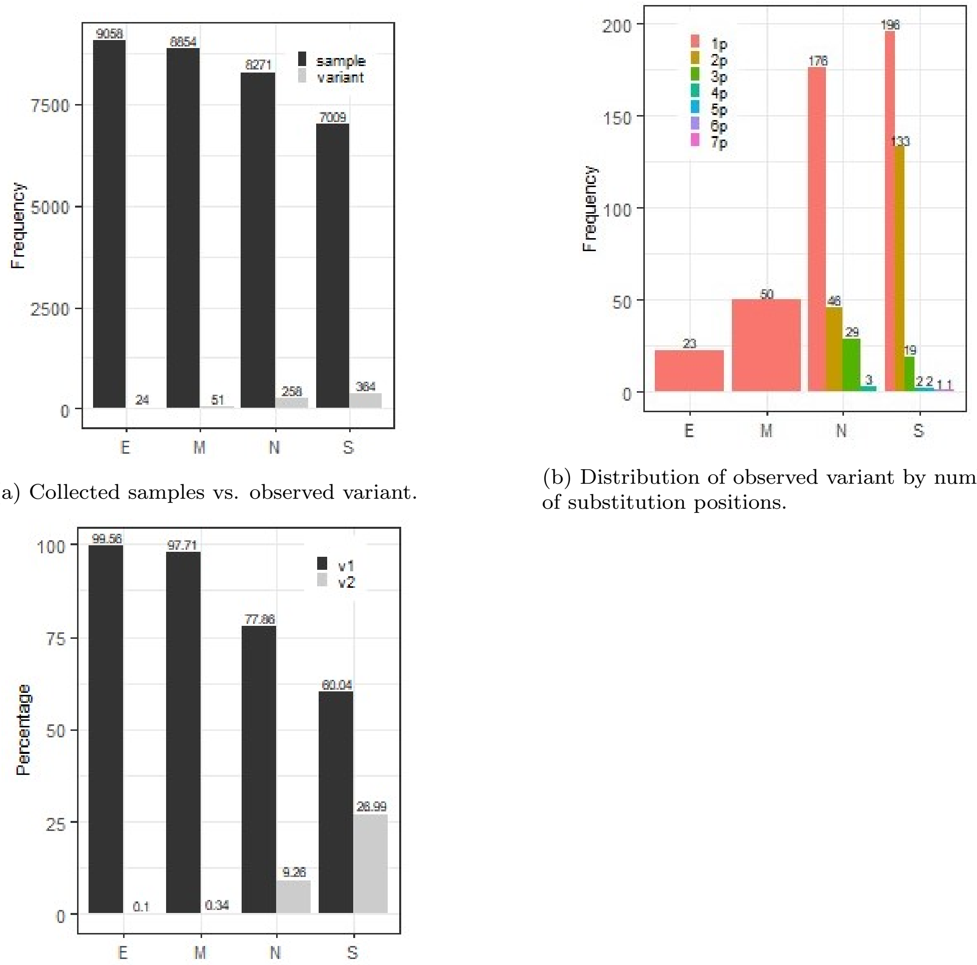
Identification of substitutions, particularly in functional domains are highly essential for understanding virulence power of SARS-COV-2 . Therefore, we try to highlight the substitutions in different functional domains in next section.

3. Results

3.1. Clustering structural protein variants

There are several sequences (or samples) which are exactly similar with regards to their constituent nucleotide organization. Therefore, all the collected structural protein sequences (E, M, N, S) are first grouped using simple string matching, such that each cluster comprises of exactly similar sequences. We observe 24, 51, 258, 364 number of clusters for E, M, N and S protein, respectively (Figure fig:clustering(a)). We set the representative as $v1, v2, \cdots, vk$ (k is the number of clusters or variants in each category of protein). We select any one sequence (or variant) from every group as the representative of that cluster for different analysis. Two variants are differ by at least one distinct substitution or same substitution in different positions. If we consider the cluster size for different proteins, we observe that more than 95% of the samples from E and M proteins shared in two clusters and 85% samples from N and S proteins (Figure fig:clustering(c)) shared in two groups. We observe a maximum variation in S protein followed by N, M and E proteins. It further reveals interesting fact that in terms of mutational changes E and M proteins are relatively stable than N and S proteins. A country-wise variants and its count (cardinality of the cluster) for each structural protein is listed in Table tab:con-specif-var.

Variants are usually differ by number of observed substitutions when compared with reference sequence. While studying the number of substitutions (n) occurs in each variants(np), interestingly we do not observe number of substitutions more than seven (07). Out of which single substitution occurring most commonly, whereas seven number of substitutions are rarely occurring in different candidate proteins (Figure fig:clustering(b)). In case of E and M protein, we observe mostly single amino acid substitutions. For N protein, substitutions varies from 1 to 4 numbers, whereas for S protein, count varies from 1 to 7.



(a) Collected samples vs. observed variant.

(b) Distribution of observed variant by number of substitution positions.

(c) The percentage of top two variants $v1$ and $v2$ by sample.

Figure 1: Variant clustering of collected samples/sequences for each structural protein. (a) collected sample with observed distinct variants; (b) Variants with number of substitutions (frequency by number of substitutions) . (c) Percentage of total samples in top (by sample frequency) two variant class.

3.2. Substitution patterns and locations

In this section, we first try to investigate the substitution patterns in comparison to our reference sequence. All possible substitutions (source to target amino acid) are shown in a 2D matrix representation in Figure fig:subtypy-matrix. In general, there could be $20 \times 20 (= 400) - 20$ (synonymous)=

Table 1: The table showing country-wise distinct variant count in each of structural proteins (Spike-S, Nucleocapsid -N, Mmembrane-M, Envelope-E).

	Variant count in					**Variant count in**			
Country	**S**	**N**	**M**	**E**	**Country**	**S**	**N**	**M**	**E**
AUS	40	42	11	4	LKA	2	3	2	1
BGD	22	14	6	3	MAR	1	3	2	1
BHR	3	2	1	1	MYS	1	1	1	1
BRA	4	3	1	1	Morocco	1	1	1	1
CHL	2	4	1	1	NGA	1	1	1	1
CHN	16	5	2	3	NLD	1	2	2	1
COL	2	2	1	1	NPL	1	1	1	1
CZE	1	2	2	1	NZL	1	1	1	1
DEU	7	6	2	1	PAK	1	1	1	1
EGY	7	6	1	1	PER	1	1	1	1
ESP	3	3	2	1	POL	2	3	1	1
FRA	15	4	4	1	RUS	2	2	2	1
GEO	1	1	1	1	SAU	5	6	2	2
GRC	11	9	1	2	SRB	2	4	2	1
GUM	1	2	1	1	SWE	1	1	1	1
Guangzhou	1	1	1	1	THA	6	2	2	1
HKG	6	4	2	1	TLS	1	2	1	1
IND	49	34	5	3	TUN	5	5	1	1
IRN	2	2	1	1	TUR	3	3	1	1
ISR	2	2	1	1	TWN	6	8	1	1
ITA	4	2	3	1	URY	1	1	1	1
JAM	2	2	1	1	USA	212	183	35	14
KAZ	2	4	2	1	VNM	1	1	1	1
KEN	1	2	2	2	ZAF	1	1	1	1
KOR	2	3	1	1					

380 possible unique substitutions considering twenty amino acids. The value in each cell of the matrix depicts the count for a particular substitution from a source (row) to target amino acid (column) occurs in different locations in different proteins, we consider.

Irrespective of variant and position of substitutions, we can observe from the Figure fig:sub-pos-type-count, a total 316, 217, 50, 23 number of amino acid substitutions in S, N, M, E protein, respectively (Figure fig:sub-pos-type-count). To understand how much random the location of substitutions in different proteins, we observe a variable number of substitution locations such as 261 ($\approx$ 21% of sequence length) for S, 167 ($\approx$ 40% of sequence length) for N, 45 ($\approx$ 20% of sequence length) for M, and 19 ($\approx$ 25% of sequence length) for E protein. Interesting, this findings shows that in regards to mutation location the N protein is most random and difficult to localize or predict its site of alternation. M protein relatively most stable in such scenario.

It is evident from our analysis that a total 91 types of substitutions are observed in S proteins. Similarly, for N, M and E proteins total 74, 33, and 19 types of substitutions observed. A position-wise substitutions in each proteins is shown in Figure fig:subtypepos. A number of substitutions in different proteins showing multiple target amino acids in the same location. For example, in case of N proteins, four target amino acids (K, M, S and G) are found in position 203 distributed across 36 variants (*Supplementary-A (Table S1.)*).

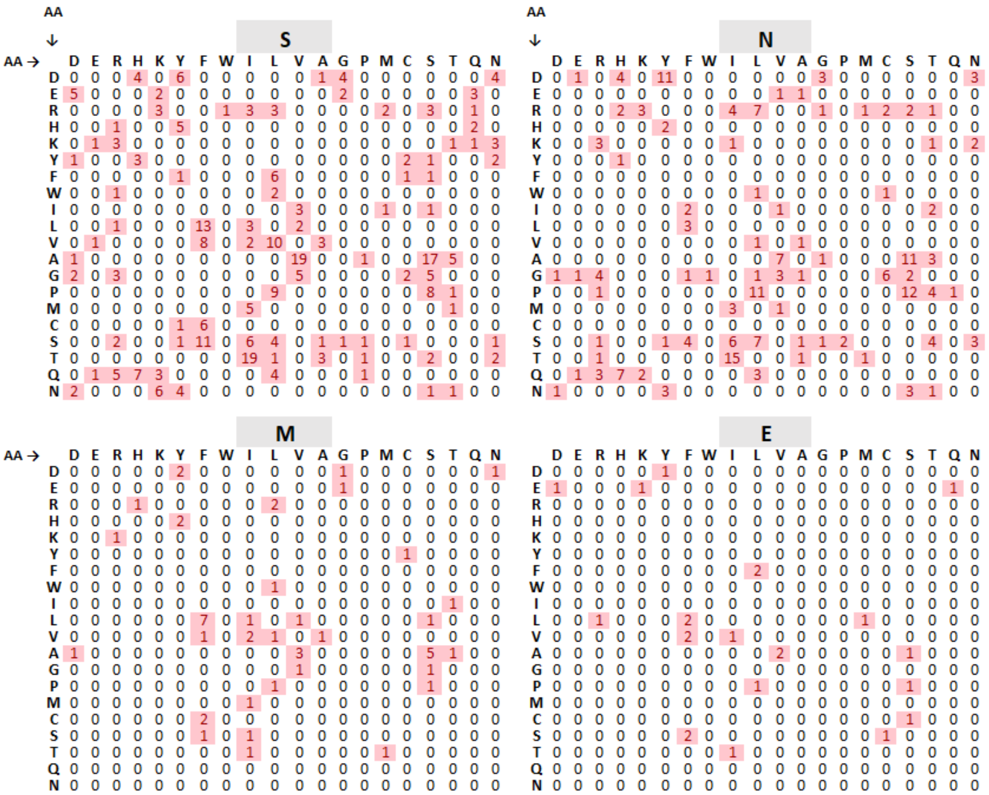
3.3. Trend of substitutions by change in amino acids

Amino acid substitution changes the linear organization of peptide bond, and hence some changes can lead to abnormal functionality disrupting its structure.

So, we focus on amino acid substitution by different types of amino acid change looking into substitutions in different positions. Majority of substitutions are observed distinct and occurring mostly in particular position. (Figure fig:sub-typy-matrix). We observe maximum substitutions for $A \rightarrow V$, which occurs in 19 different positions followed by $A \rightarrow S$ (17 different position) and $L \rightarrow F$ (13 different positions) in S protein. In case of N protein, we observe maximum substitution for $T \rightarrow I$, which occurs in 15 different positions followed by $P \rightarrow S$ (12 different position) and $P \rightarrow L$, $A \rightarrow S$ and $D \rightarrow Y$ (11 different positions). In case of M protein, we observe maximum substitution for $A \rightarrow S$ (15 different positions), and we do not observe more than two substitutions in any particular position in case of E protein.

Towards understanding amino acid change by substitutions type, we calculate the percentage by observing all substitutions occurring in all positions (discussed above). There are two categories of substitution we consider, one to many and many to one keeping a particular source or target amino acid. For both the categories, the total substitution count, the number of distinct substitution type (from Figure fig:sub-typy-matrix), and their percentages

doi:10.20944/preprints202008.0621.v1



AA ↓

S

AA →	D	E	R	H	K	Y	F	W	I	L	V	A	G	P	M	C	S	T	Q	N
D	0	0	0	4	0	6	0	0	0	0	0	1	4	0	0	0	0	0	0	4
E	5	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	0	0	3	0
R	0	0	0	0	3	0	0	1	3	3	0	0	0	0	2	0	3	0	1	0
H	0	0	1	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	2	0
K	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	3
Y	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	2
F	0	0	0	0	0	1	0	0	0	6	0	0	0	0	0	1	1	0	0	0
W	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	3	0	0	0	1	0	1	0	0	0
L	0	0	1	0	0	0	13	0	3	0	2	0	0	0	0	0	0	0	0	0
V	0	1	0	0	0	0	8	0	2	10	0	3	0	0	0	0	0	0	0	0
A	1	0	0	0	0	0	0	0	0	0	19	0	0	1	0	0	17	5	0	0
G	2	0	3	0	0	0	0	0	0	0	5	0	0	0	0	2	5	0	0	0
P	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	8	1	0	0
M	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	1	0	0
C	0	0	0	0	0	1	6	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	2	0	0	1	11	0	6	4	0	1	1	1	0	1	0	0	0	1
T	0	0	0	0	0	0	0	0	19	1	0	3	0	1	0	0	2	0	0	2
Q	0	1	5	7	3	0	0	0	0	4	0	0	0	1	0	0	0	0	0	0
N	2	0	0	0	6	4	0	0	0	0	0	0	0	0	0	0	1	1	0	0

AA ↓

N

	D	E	R	H	K	Y	F	W	I	L	V	A	G	P	M	C	S	T	Q	N
D	0	1	0	4	0	11	0	0	0	0	0	0	3	0	0	0	0	0	0	3
E	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
R	0	0	0	2	3	0	0	0	4	7	0	0	1	0	1	2	2	1	0	0
H	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	2
Y	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
I	0	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	0	2	0	0
L	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	7	0	1	0	0	0	11	3	0	0
G	1	1	4	0	0	0	1	1	0	1	3	1	0	0	0	6	2	0	0	0
P	0	0	1	0	0	0	0	0	0	11	0	0	0	0	0	0	12	4	1	0
M	0	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	1	0	0	1	4	0	6	7	0	1	1	2	0	0	0	4	0	3
T	0	0	1	0	0	0	0	0	15	0	0	1	0	0	1	0	0	0	0	0
Q	0	1	3	7	2	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
N	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	3	1	0	0

M

AA →	D	E	R	H	K	Y	F	W	I	L	V	A	G	P	M	C	S	T	Q	N
D	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	1
E	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
R	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
L	0	0	0	0	0	0	7	0	1	0	1	0	0	0	0	0	1	0	0	0
V	0	0	0	0	0	0	1	0	2	1	0	1	0	0	0	0	0	0	0	0
A	1	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	5	1	0	0
G	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
P	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
M	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

E

	D	E	R	H	K	Y	F	W	I	L	V	A	G	P	M	C	S	T	Q	N
D	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	1	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0
V	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
S	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2: The figure shows amino acid substitution by position count. The amino acid substitution type is shown by a 2D matrix relation. In figure, a substitution type, say $P->Q$, indicates left amino acid P (y-axis, left) (source) to substitution right right amino acid Q (x-axis, top) (target), and the cell value in the matrix indicates observed number of substitution positions for that particular substitution type (from Figure fig:subtypepos).

for each category of proteins are calculated. The top 5 such substitutions are shown in Table tab:top5-s-type. We observe a majority of amino acid change for L to other in both the E protein (≈ 17%, 3 distinct types) and M protein (≈ 20%, 4 distinct types). In case of N protein, majority of substitutions occurs for S to other (≈ 13%, 10 distinct types). For S protein, we observe A to other change is the majority (≈ 14%, 5 distinct types), but maximum distinct type observe for amino acid change S to other (10 distinct types).

3.4. Trend of substitutions by changing chemical properties of amino acids

A non-synonymous substitution that alters the amino acid, thereby changing the biochemical properties of amino acid. In this section, we categorise

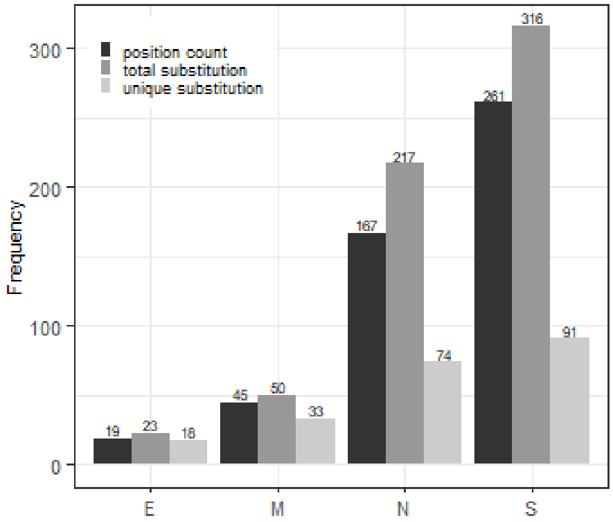
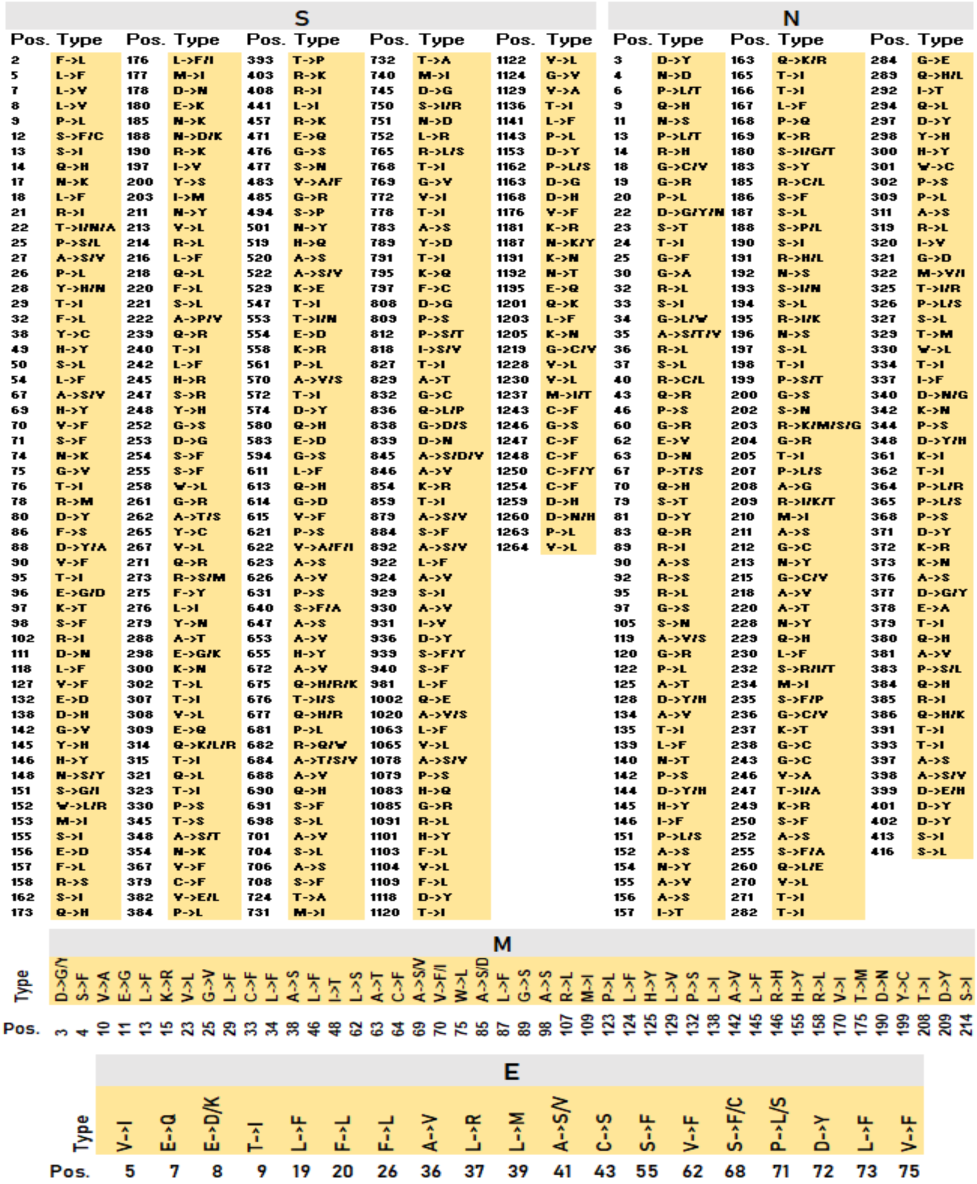


Figure 3: The figure shows observed total number of substitution irrespective of positions and variants, total substitution in all positions, observed substitutions by positions, and unique substitution count for each structural protein (S, N, M, E).

Table 2: The table shows top 5 substitution type and associative quantitative information for two categories. For a substitution $X \rightarrow Y$, in the the first category, X is fixed and Y can be any amino acid, and in second category, X is any amino acid and Y is a fixed.

$X \rightarrow any$					$Any \rightarrow Y$				
Protein	AA	Count	Distinct	Percentage	Protein	AA	Count	Distinct	Percentage
E	L	4	3	17.39	E	F	6	3	26.09
E	E	3	3	13.04	E	L	3	2	13.04
E	V	3	2	13.04	E	S	3	3	13.04
E	A	3	2	13.04	E	I	2	2	8.70
E	S	3	2	13.04	E	V	2	1	8.70
M	L	10	4	20.00	M	F	11	4	22.00
M	A	10	4	20.00	M	S	8	4	16.00
M	V	5	4	10.00	M	I	6	5	12.00
M	D	4	3	8.00	M	L	5	4	10.00
M	R	3	2	6.00	M	V	5	3	10.00
N	S	30	10	13.82	N	L	31	7	14.29
N	P	29	5	13.36	N	S	30	5	13.82
N	R	23	9	10.60	N	I	29	5	13.36
N	D	22	5	10.14	N	Y	17	4	7.83
N	A	22	4	10.14	N	T	16	7	7.37
S	A	43	5	13.61	S	L	39	8	12.34
S	S	29	10	9.18	S	S	39	9	12.34
S	T	28	6	8.86	S	F	38	4	12.03
S	V	24	5	7.59	S	I	38	6	12.03
S	Q	21	6	6.65	S	V	29	4	9.18

doi:10.20944/preprints202008.0621.v1



S

Pos.	Type	Pos.	Type	Pos.	Type	Pos.	Type	Pos.	Type
2	F->L	176	L->F/I	393	T->P	732	T->A	1122	V->L
5	L->F	177	M->I	403	R->K	740	M->I	1124	G->V
7	L->V	178	D->N	408	R->I	745	D->G	1129	V->A
8	L->V	180	E->K	441	L->I	750	S->I/R	1136	T->I
9	P->L	185	N->K	457	R->K	751	N->D	1141	L->F
12	S->F/C	188	N->D/K	471	E->Q	752	L->R	1143	P->L
13	S->I	190	R->K	476	G->S	765	R->L/S	1153	D->Y
14	Q->H	197	I->V	477	S->N	768	T->I	1162	P->L/S
17	N->K	200	Y->S	483	V->A/F	769	G->V	1163	D->G
18	L->F	203	I->M	485	G->R	772	V->I	1168	D->H
21	R->I	211	N->Y	494	S->P	778	T->I	1176	V->F
22	T->I/N/A	213	V->L	501	N->Y	783	A->S	1181	K->R
25	P->S/L	214	R->L	519	H->Q	789	Y->D	1187	N->K/Y
27	A->S/V	216	L->F	520	A->S	791	T->I	1191	K->N
26	P->L	218	Q->L	522	A->S/V	795	K->Q	1192	N->T
28	Y->H/N	220	F->L	529	K->E	797	F->C	1195	E->Q
29	T->I	221	S->L	547	T->I	808	D->G	1201	Q->K
32	F->L	222	A->P/V	553	T->I/N	809	P->S	1203	L->F
38	Y->C	239	Q->R	554	E->D	812	P->S/T	1205	K->N
49	H->Y	240	T->I	558	K->R	818	I->S/V	1219	G->C/V
50	S->L	242	L->F	561	P->L	827	T->I	1228	V->L
54	L->F	245	H->R	570	A->V/S	829	A->T	1230	V->L
67	A->S/V	247	S->R	572	T->I	832	G->C	1237	M->I/T
69	H->Y	248	Y->H	574	D->Y	836	Q->L/P	1243	C->F
70	V->F	252	G->S	580	Q->H	838	G->D/S	1246	G->S
71	S->F	253	D->G	583	E->D	839	D->N	1247	C->F
74	N->K	254	S->F	594	G->S	845	A->S/D/V	1248	C->F
75	G->V	255	S->F	611	L->F	846	A->V	1250	C->F/Y
76	T->I	258	W->L	613	Q->H	854	K->R	1254	C->F
78	R->M	261	G->R	614	G->D	859	T->I	1259	D->H
80	D->Y	262	A->T/S	615	V->F	879	A->S/V	1260	D->N/H
86	F->S	265	Y->C	621	P->S	884	S->F	1263	P->L
88	D->Y/A	267	V->L	622	V->A/F/I	892	A->S/V	1264	V->L
90	V->F	271	Q->R	623	A->S	922	L->F		
95	T->I	273	R->S/M	626	A->V	924	A->V		
96	E->G/D	275	F->Y	631	P->S	929	S->I		
97	K->T	276	L->I	640	S->F/A	930	A->V		
98	S->F	279	Y->N	647	A->S	931	I->V		
102	R->I	288	A->T	653	A->V	936	D->Y		
111	D->N	298	E->G/K	655	H->Y	939	S->F/Y		
118	L->F	300	K->N	672	A->V	940	S->F		
127	V->F	302	T->L	675	Q->H/R/K	981	L->F		
132	E->D	307	T->I	676	T->I/S	1002	Q->E		
138	D->H	308	V->L	677	Q->H/R	1020	A->V/S		
142	G->V	309	E->Q	681	P->L	1063	L->F		
145	Y->H	314	Q->K/L/R	682	R->Q/W	1065	V->L		
146	H->Y	315	T->I	684	A->T/S/V	1078	A->S/V		
148	N->S/Y	321	Q->L	688	A->V	1079	P->S		
151	S->G/I	323	T->I	690	Q->H	1083	H->Q		
152	W->L/R	330	P->S	691	S->F	1085	G->R		
153	M->I	345	T->S	698	S->L	1091	R->L		
155	S->I	348	A->S/T	701	A->V	1101	H->Y		
156	E->D	354	N->K	704	S->L	1103	F->L		
157	F->L	367	V->F	706	A->S	1104	V->L		
158	R->S	379	C->F	708	S->F	1109	F->L		
162	S->I	382	V->E/L	724	T->A	1118	D->Y		
173	Q->H	384	P->L	731	M->I	1120	T->I		

N

Pos.	Type	Pos.	Type	Pos.	Type
3	D->Y	163	Q->K/R	284	G->E
4	N->D	165	T->I	289	Q->H/L
6	P->L/T	166	T->I	292	I->T
9	Q->H	167	L->F	294	Q->L
11	N->S	168	P->Q	297	D->Y
13	P->L/T	169	K->R	298	Y->H
14	R->H	180	S->I/G/T	300	H->Y
18	G->C/V	183	S->Y	301	W->C
19	G->R	185	R->C/L	302	P->S
20	P->L	186	S->F	309	P->L
22	D->G/Y/N	187	S->L	311	A->S
23	S->T	188	S->P/L	319	R->L
24	T->I	190	S->I	320	I->V
25	G->F	191	R->H/L	321	G->D
30	G->A	192	N->S	322	M->V/I
32	R->L	193	S->I/N	325	T->I/R
33	S->I	194	S->L	326	P->L/S
34	G->L/W	195	R->I/K	327	S->L
35	A->S/T/V	196	N->S	329	T->M
36	R->L	197	S->L	330	W->L
37	S->L	198	T->I	334	T->I
40	R->C/L	199	P->S/T	337	I->F
43	Q->R	200	G->S	340	D->N/G
46	P->S	202	S->N	342	K->N
60	G->R	203	R->K/M/S/G	344	P->S
62	E->V	204	G->R	348	D->Y/H
63	D->N	205	T->I	361	K->I
67	P->T/S	207	P->L/S	362	T->I
70	Q->H	208	A->G	364	P->L/R
79	S->T	209	R->I/K/T	365	P->L/S
81	D->Y	210	M->I	368	P->S
83	Q->R	211	A->S	371	D->Y
89	R->I	212	G->C	372	K->R
90	A->S	213	N->Y	373	K->N
92	R->S	215	G->C/V	376	A->S
95	R->L	218	A->V	377	D->G/Y
97	G->S	220	A->T	378	E->A
105	S->N	228	N->Y	379	T->I
119	A->V/S	229	Q->H	380	Q->H
120	G->R	230	L->F	381	A->V
122	P->L	232	S->R/I/T	383	P->S/L
125	A->T	234	M->I	384	Q->H
128	D->Y/H	235	S->F/P	385	R->I
134	A->V	236	G->C/V	386	Q->H/K
135	T->I	237	K->T	391	T->I
139	L->F	238	G->C	393	T->I
140	N->T	243	G->C	397	A->S
142	P->S	246	V->A	398	A->S/V
144	D->Y/H	247	T->I/A	399	D->E/H
145	H->Y	249	K->R	401	D->Y
146	I->F	250	S->F	402	D->Y
151	P->L/S	252	A->S	413	S->I
152	A->S	255	S->F/A	416	S->L
154	N->Y	260	Q->L/E		
155	A->V	270	V->L		
156	A->S	271	T->I		
157	I->T	282	T->I		

M

Type	D->G/Y	S->F	V->A	E->G	L->F	K->R	V->L	G->V	L->F	C->F	L->F	A->S	L->F	I->T	L->S	A->T	C->F	A->S/V	V->F/I	W->L	A->S/D	L->F	G->S	A->S	R->L	M->I	P->L	L->F	H->Y	L->V	P->S	L->I	A->V	L->F	R->H	H->Y	R->L	V->I	T->M	D->N	Y->C	T->I	D->Y	S->I
Pos.	3	4	10	11	13	15	23	25	29	33	34	38	46	48	62	63	64	69	70	75	85	87	89	98	107	109	123	124	125	129	132	138	142	145	146	155	158	170	175	190	199	208	209	214

E

Type	V-->I	E-->Q	E-->D/K	T-->I	L-->F	F-->L	F-->L	A-->V	L-->R	L-->M	A-->S/V	C-->S	S-->F	V-->F	S-->F/C	P-->L/S	D-->Y	L-->F	V-->F
Pos.	5	7	8	9	19	20	26	36	37	39	41	43	55	62	68	71	72	73	75

Figure 4: The figure shows observed amino acid substitution by position and type, which are obtained from across all the variants and shown by each structural protein shaded at the top (S, N. M, E).

the substitution types by observing their chemical properties of amino acids. We consider two kinds of biochemical properties, eight chemical properties

doi:10.20944/preprints202008.0621.v1

of amino acids and three Hydropathy class of amino acid as discussed in Section sec:biochem.

In the case of side-chain structure change (using eight chemical properties of amino acids), most of the substitutions change its chemical properties. We see 77 out of 91 ($\approx$ 85%), 61 out of 74 ($\approx$ 82%), 24 out of 33 ($\approx$ 73%), 15 out of 18 ($\approx$ 83%) of substitutions in S, N, M, and E protein, respectively, that transit to different chemical groups. In case of E and M proteins, majority of the substitutions leads change in chemical properties from Aromatic to Aliphatic, Aliphatic to Hydroxyl-containing for M protein, Hydroxyl-containing to Aliphatic for N protein, Aliphatic to Aromatic and Hydroxyl-containing for S protein (Figure fig:sub-by-amino-chem(a)).

More than 50% substitutions (except E protein), we observe a change of one Hydropathy class to another. We observe 50 out of 91 ($\approx$ 55%), 46 out of 74 ($\approx$ 62%), 17 out of 33 ($\approx$ 52%), 8 out of 18 ($\approx$ 44%) of substitutions in S, N, M, and E protein, respectively, that changes the Hydropathy class. The significant change in Hydropathy classes from Hydrophilic to Hydrophobic (S, N and E protein), and Hydrophobic to Hydrophilic (M protein) (Figure fig:sub-by-amino-chem(b)).

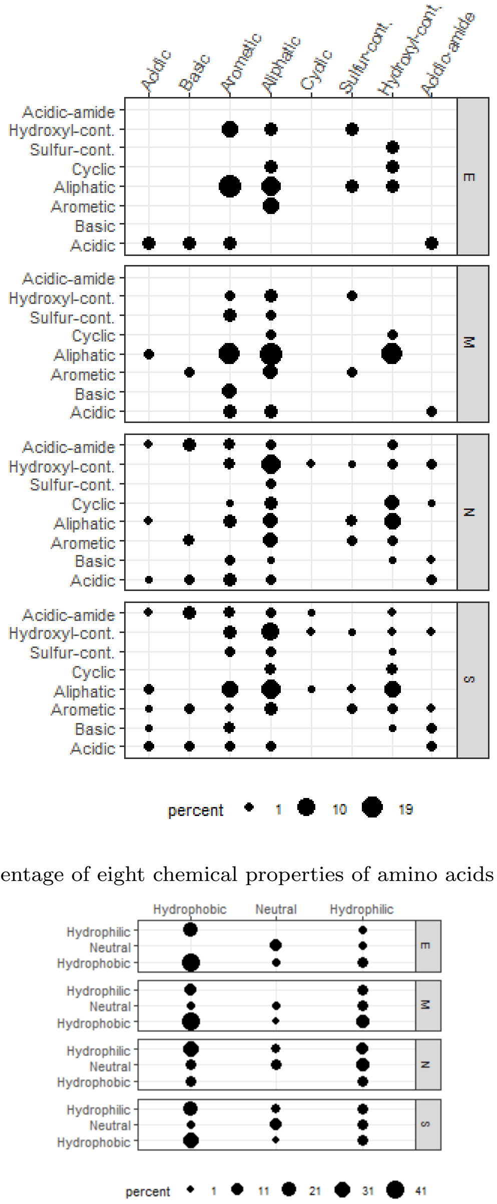
3.5. Substitution in different functional sub-domains of structural proteins

In this section, we try to highlight and quantify the substitutions in various functional sub-domains of two structural proteins, S and N. We report domain specific substitutions in S and N proteins in Table tab:sub-func.

Observed substitutions are mostly unique. In a particular location not more than two substations are observed. In case of S protein, we observe maximum substitutions in HR2 and FP domains ($\approx$ 22% locations) followed by Transmembrane (TM) domain ($\approx$ 17% locations). From our study we may assume that SARS-COV-2 favours mutations in HR1 and FP to become more virulent by making the mechanism of host cell membrane fusion and entry to host cell more improved. Although receptor binding domain in S protein is equally important for viral entry to host [53], we observe comparatively few substitutions ($\approx$ 12% locations) in this domain.

In case of N protein, we observe substitutions in both the NTD and CTD domains ($\approx$ 30% locations). We observe few substitutions in NTD and CTD domain are related to charged amino acids. This may leads to issues in RNA packaging in the virus [59]. Although the molecular mechanism of SARS-COV-2 N protein is yet to explore thoroughly [60].

doi:10.20944/preprints202008.0621.v1



(a) Change percentage of eight chemical properties of amino acids by substitution.

(b) Change percentage of Hydropathy classes of amino acids by substitution.

Figure 5: The trends of substitution type (percentage) shown for two categories of amino acid biochemical property.

doi:10.20944/preprints202008.0621.v1

Table 3: Observed substitutions and quantitative information in different functional domain of S and N protein.

Protein	Domain	Substitution	Count	Percentage
S	RBD	Q->L, T->I, P->S, T->S, A->S, N->K, V->F, C->F, V->L, V->E, P->L, T->P, R->K, R->I, L->I, E->Q, G->S, S->N, V->A, G->R, S->P, N->Y, H->Q, K->E, A->T, A->V	26	0.12
	HR1	L->F, A->V, S->I, I->V, D->Y, S->F,S->Y	7	0.13
	HR2	D->G, D->H, V->F, K->R, N->K, K->N, N->T, E->Q, Q->K, L->F, N->Y	11	0.22
	FP	Y->D, T->I, K->Q, F->C	4	0.22
	TM	G->C, G->V, V->L, M->I, M->T	5	0.17
N	NTD	P->S, G->R, E->V, D->N, P->T, Q->H, S->T, D->Y, Q->R, R->I, A->S, R->S, R->L, G->S, S->N, A->V, P->L, A->T, T->I, L->F, N->T, H->Y, I->F, N->Y, I->T, Q->K, P->Q, K->R, D->H	29	0.33
	CTD	T->I, K->R, S->F, A->S, Q->L, V->L, G->E, Q->H, I->T, D->Y, Y->H, H->Y, W->C, P->S, P->L, R->L, I->V, G->D, M->V, S->L, T->M, W->L, I->F, D->N, K->N, K->I, T->A, S->A, Q->E, D->G, D->H, M->I, T->R	33	0.27

3.6. Country specific unique variant and substitution study

An unique variant in more than one sample may be highly significant. Therefore, we search for those unique variants and substitutions type in 26 countries. Among 49 countries, we find at least one unique variants in 26 countries associated in all four proteins (E-6 country, M-7 country, N-18 country, S-22 country). Some of the countries (including USA, AUS, IND, FRA, GRC, HKG, EGY) having more than 50% of variants are unique variant in each category (Figure fig:percent-unique-variant), mostly in N and S protein. We highlight the unique variants with at least two associated samples, which is observed only in ten countries (Table tab:con-specif-var). Few country specific observations are listed below.

- **Australia (AUS) :** We observe unique variant 1 in E protein, 2 in M protein and several variants in N and S protein. The top variants are $v4$ (substitution is $L \rightarrow F(73)$[Change the chemical property Aliphatic to Arometic, no Hydropathy class change]) in E protein, $v12$ (substitution is $v \rightarrow I(214)$, no change in chemical property, no Hydropathy class change) in M protein, $v17$ (substitution is $D \rightarrow G(22)$ [Change the chemical property Acidic to Aliphatic, Neutral to Hydrophilic]) in N protein, and $v5$ (substitution is $S \rightarrow N(477)$ [Change the chemical property Hydroxyl-containing to Acidic-amide, Hydrophilic to Neutral] in S protein.

doi:10.20944/preprints202008.0621.v1

- **Bangladesh (BGD) :** Observed unique variant 1 in N protein, 2 in S protein. The top variants are $v39$ (substitution are $R \rightarrow K(203)$ [no change in chemical property], $G \rightarrow R(204)$ [Change the chemical property Aliphatic to Basic, Hydrophilic to Neutral], $D \rightarrow G(377)$ Change the chemical property Acidic to Aliphatic, Neutral to Hydrophilic]) in N protein, $v4$ (substitution are $L \rightarrow F(73)$[Change the chemical property Aliphatic to Arometic, no Hydropathy class change], $T \rightarrow I(95)$[Change the chemical property Hydroxyl-containing to Aliphatic, Hydrophilic to Hydrophobic]).

- **Egypt (EGY):** Observed only 1 unique variant in S protein. The variant is $v81$ (substitution is $R \rightarrow I(408)$ [Change the chemical property Basic to Aliphitic, Neutral to Hydrophobic].

- **French (FRA):** Only 1 unique variant with multiple substitutions in S protein. The variant is $v88$ (substitution are $T \rightarrow I(95)$[Change the chemical property Hydroxyl-containgni to Aliphatic, Hydrophilic to Hydrophobic], $P \rightarrow L(1162)$[Change the chemical property Cyclic to Aliphatic, Hydrophilic to Hydrophobic]).

- **Greece (GRC):** Only 1 unique variant with multiple substitutions in N protein. The variant is $v34$ (substitution are $R \rightarrow K(203)$ [no change in chemical property], $G \rightarrow R(204)$[Change the chemical property Aliphatic to Basic, Hydrophilic to Neutral], $T \rightarrow I(95)$[Change the chemical property Hydroxyl-containgni to Aliphatic, Hydrophilic to Hydrophobic]).

- **Hongkong (HKG) :** Only 1 unique variant with multiple substitutions in S protein. The variant is $v63$ (substitution are $L \rightarrow V(8)$[no change in chemical property], $G \rightarrow D(614)$[Change the chemical property Aliphatic to Acidic, Hydrophilic to Neutral]).

- **India (IND):** We observe 6 unique variant in N and S protein. The top variants are $v55$ (substitution are $A \rightarrow S(156)$[Change the chemical Aliphatic to Hydroxyl-containing, Hydrophobic to Hydrophilic], $S \rightarrow L(194)$[Change the chemical property Hydroxyl-containing to Aliphatic, Hydrophilic to Hydrophobic]), $v19$ (substitution is $R \rightarrow M(78)$ [Change the chemical property Basic to Sulfur-containing, Neutral to Hydrophobic]).

doi:10.20944/preprints202008.0621.v1

- **Thailand (THA):** We observe only 1 unique variant with multiple substitutions in S protein. The variant is $v6$ (substitution are $G \rightarrow D(614)$[Change the chemical property Aliphatic to Acidic, Hydrophilic to Neutral], $A \rightarrow T(614)$[Change the chemical property Aliphatic to Hydroxyl-containing, Hydrophobic to Hydrophilic]).

- **Taiwan (TWN) :** Only 1 unique variant with multiple substitutions in S protein. The variant is $v26$ (substitutions are $G \rightarrow D(614)$[Change the chemical property Aliphatic to Acidic, Hydrophilic to Neutral], $T \rightarrow I(95)$[Change the chemical property Hydroxyl-containgni to Aliphatic, Hydrophilic to Hydrophobic]).

- **United States (USA):** Several unique variants with multiple substitutions in all four structural proteins. The top variants are $v2$ (substitution is $P \rightarrow L(71)$[Change the chemical property Cyclic to Apliphatic, Hydrophilic to Hydrophobic] in E protein, $v5$ (substitution is $K \rightarrow R(15)$[no change in chemical property]) in M protein, $v8$ (substitution is $E \rightarrow V(62)$[change chemical property Acidic to Aliphatic, Neutral to Hydrophobic]) in N protein, $v8$ (substitution are $v \rightarrow A(483)$[no change chemical property], $G \rightarrow D(614)$[change chemical property Aliphatic to Acidic, Hydrophilic to Neutral]) in S protein.

Table 4: Country-wise unique variants, order by sample frequency (minimum of two samples) in each category of structural proteins (S, N, M, E). For each variant, amino acid (AA) substitution type and position are also shown. var-variant, sam.-sample, pos.-position

Country	Protein	Var.	#Sam.	AA substitution (pos.)
AUS	E	v4	3	L->F(73)
	M	v12	4	V->I(170)
	M	v13	4	S->I(214)
	N	v17	13	D->G(22)
	N	v36	4	L->F(139)
	N	v46	3	M->I(234),T->I(334)
	N	v52	3	N->Y(228)
	N	v70	2	S->F(255)
	S	v5	37	S->N(477)
	S	v14	13	N->Y(501)
	S	v22	7	G->D(614),G->V(1124)
	S	v37	5	G->V(1124)

doi:10.20944/preprints202008.0621.v1

Table 4 – continued from previous page

Country	Protein	Var.	#Sam.	AA substitution (pos.)
	S	v40	4	S->L(50)
	S	v46	4	A->T(262)
	S	v54	3	G->D(614),D->N(1260)
	S	v68	2	T->I(22),E->K(180),G->D(614)
	S	v98	2	I->V(931)
BGD	N	v39	4	R->K(203),G->R(204),D->G(377)
	S	v64	2	L->F(5),T->I(95)
	S	v70	2	P->L(26)
EGY	S	v81	2	R->I(408)
FRA	S	v88	2	T->I(676),P->L(1162)
GRC	N	v34	5	R->K(203),G->R(204),T->I(205)
HKG	S	v63	3	L->V(8),G->D(614)
IND	N	v55	3	A->S(156),S->L(194)
	N	v60	2	S->I(33)
	S	v19	8	R->M(78)
	S	v66	2	S->F(12)
	S	v76	2	W->L(152)
	S	v77	2	M->I(177)
THA	S	v6	37	G->D(614),A->T(829)
TWN	S	v26	6	G->D(614),T->I(791)
USA	E	v2	9	P->L(71)
	E	v5	2	S->F(55)
	E	v6	2	P->S(71)
	E	v7	2	F->L(26)
	M	v5	18	K->R(15)
	M	v7	12	V->I(70)
	M	v10	6	L->V(129)
	M	v14	3	M->I(109)
	M	v15	3	L->F(145)
	M	v16	3	H->Y(155)
	M	v17	3	C->F(64)
	M	v19	2	A->S(85)
	N	v8	34	E->V(62)
	N	v13	17	E->A(378)
	N	v15	16	T->I(362)
	N	v18	11	D->Y(371)

Table 4 – continued from previous page

Country	Protein	Var.	#Sam.	AA substitution (pos.)
	N	v19	10	G->C(18)
	N	v22	10	D->Y(144)
	N	v23	9	P->S(368)
	N	v28	6	S->L(188)
	N	v31	6	E->V(62),S->L(194)
	N	v32	5	R->I(89)
	N	v35	5	Q->H(229)
	N	v37	4	R->L(185)
	N	v41	4	K->R(169),R->K(203),G->R(204)
	N	v42	4	A->V(155)
	N	v49	3	P->L(365)
	N	v50	3	Q->H(384)
	N	v56	3	P->S(142)
	N	v57	3	D->Y(128)
	N	v58	3	P->T(67)
	N	v59	2	T->I(24)
	N	v61	2	I->F(146)
	N	v62	2	T->I(166)
	N	v64	2	S->N(193)
	N	v65	2	R->K(203)
	N	v66	2	R->K(203),G->R(204),G->C(243)
	N	v67	2	G->C(212)
	N	v71	2	Q->H(289)
	N	v72	2	Y->H(298)
	N	v73	2	T->I(334)
	N	v74	2	P->L(364)
	N	v75	2	Q->H(386)
	N	v76	2	T->I(391)
	N	v78	2	S->L(416)
	N	v80	2	A->S(397)
	N	v81	2	D->Y(377)
	N	v82	2	P->S(365)
	N	v84	2	D->N(340)
	N	v85	2	G->S(200)
	N	v87	2	A->S(156),R->K(203),G->R(204)
	N	v88	2	A->S(156)

Table 4 – continued from previous page

Country	Protein	Var.	#Sam.	AA substitution (pos.)
	N	v89	2	A->S(152)
	N	v90	2	D->H(144)
	N	v91	2	P->S(67)
	N	v92	2	A->S(35)
	N	v93	2	G->L(34)
	N	v95	2	D->Y(22)
	N	v96	2	G->R(19)
	N	v97	2	P->T(13)
	S	v8	22	V->A(483),G->D(614)
	S	v9	21	H->Y(146)
	S	v11	16	P->L(681)
	S	v17	10	D->H(138),E->D(554)
	S	v20	8	V->L(308)
	S	v21	8	R->K(403)
	S	v25	7	A->D(845)
	S	v27	6	G->D(614),G->D(838)
	S	v29	6	T->I(859),P->L(1263)
	S	v30	6	S->F(939)
	S	v31	6	G->S(476),G->D(614)
	S	v32	6	F->L(220)
	S	v36	5	L->F(1203)
	S	v38	5	A->S(520),G->D(614)
	S	v39	4	T->I(29),G->D(614)
	S	v41	4	L->F(54),G->D(614)
	S	v42	4	Y->H(145)
	S	v43	4	L->F(216)
	S	v47	3	L->F(5),H->Y(146)
	S	v49	3	S->F(71),G->D(614)
	S	v50	3	S->P(494)
	S	v51	3	T->I(547)
	S	v52	3	G->D(614),S->L(704)
	S	v53	3	G->D(614),S->F(940)
	S	v55	3	G->D(614),G->R(1085)
	S	v56	3	Q->L(836)
	S	v57	3	S->I(929)
	S	v58	3	H->Y(1101)

Table 4 – continued from previous page

Country	Protein	Var.	#Sam.	AA substitution (pos.)
	S	v59	3	A->S(1078)
	S	v61	3	A->S(262),G->D(614)
	S	v65	2	L->F(5),G->S(476),G->D(614)
	S	v67	2	L->F(18)
	S	v69	2	T->N(22),S->L(698)
	S	v73	2	T->I(95)
	S	v74	2	T->I(95),D->H(138),E->D(554)
	S	v75	2	S->F(98)
	S	v78	2	R->L(214)
	S	v80	2	P->L(384)
	S	v82	2	T->I(553)
	S	v83	2	E->D(583),Q->R(675)
	S	v86	2	G->D(614),T->I(1136)
	S	v87	2	S->F(640)
	S	v89	2	Q->H(677),V->L(1230)
	S	v90	2	G->D(838)
	S	v91	2	L->F(922)
	S	v92	2	V->L(1065)
	S	v94	2	V->L(1122)
	S	v96	2	A->V(1078)
	S	v100	2	P->S(812)
	S	v101	2	A->S(783)
	S	v102	2	A->V(570)
	S	v103	2	P->S(330)
	S	v104	2	R->S(273)
	S	v105	2	I->M(203)
	S	v106	2	D->H(138),E->D(554),V->A(1129)
	S	v107	2	D->N(111),G->D(614)
	S	v108	2	D->Y(88),C->F(1250)
	S	v109	2	D->Y(80),G->D(614)
	S	v110	2	H->Y(69)
	S	v112	2	F->L(32)
	S	v113	2	A->S(27)
	S	v114	2	P->S(25)

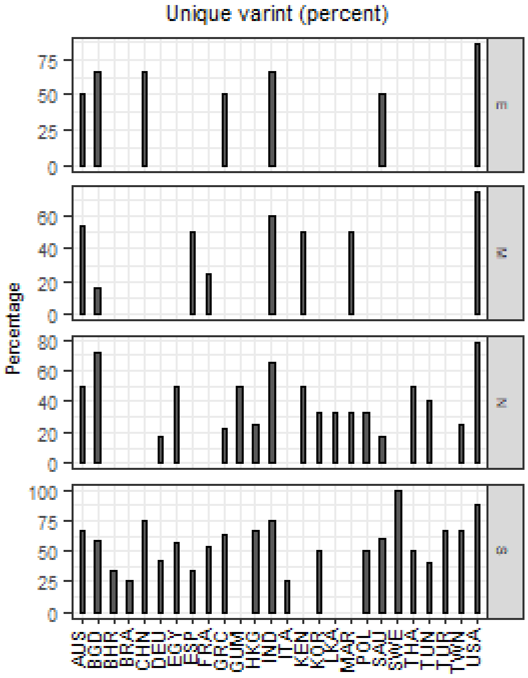


Figure 6: Country-wise percentage of unique variant observed for each structural protein (S, N, M, E).

4. Conclusion

In this work, we performed a statistical study of the structural proteins of SARS-COV-2 based on 9587 sequences, reported from different parts of the world. All the similar sequences are grouped together, and variations are analysed based on Wuhan SARS-COV-2 sequence (as reference). We highlighted various commonly and rarely occurring amino acids substitutions in four structural proteins. We reported location-wise amino acid substitution patterns in the structural proteins. The change in biochemical groupings as a result of substitutions are also reported for the candidate variants. We even highlighted above variations specific to any particular country.

Although we observed much more unique variants in S proteins as compared to N protein, in terms of substitutions changes, N protein is much vulnerable as it showing the substitution in 40% positions. Majority of substitutions type for all four proteins are observed distinct, and they occur mostly in a particular location. The substitution type $A \rightarrow V$ is very common for S, N and M protein. We also did not observe more than two substitutions in any particular position in the case of E protein. When considering biochemical properties, it is notated that the majority of the substitutions change from Aromatic ⇔ Aliphatic, Aliphatic ⇔ Hydroxyl-containing and Hydrophilic → Hydrophobic. We also found that more than 50% of variants are unique variant in each country for each category of structural proteins.

We believe that the current findings will be helpful in better understanding the disease pathogenesis of COVID-19 followed by suitable and relatively stable small molecules idendification that may binds susceptible structural proteins like Spike (S) protein that are changing frequently.

Supplementary Materials

Supplementary-A

References

[1] S. Herlitze, M. Koenen, A general and rapid mutagenesis method using polymerase chain reaction, Gene 91 (1990) 143–147.

[2] M. C. Zambon, The pathogenesis of influenza in humans, Reviews in medical virology 11 (2001) 227–241.

[3] P. J. Walker, J. A. Cowley, Viral genetic variation: implications for disease diagnosis and detection of shrimp pathogens, FAO fisheries. Technical paper (2000) 54–9.

[4] J. B. Watney, P. K. Agarwal, S. Hammes-Schiffer, Effect of mutation on enzyme motion in dihydrofolate reductase, Journal of the American Chemical Society 125 (2003) 3745–3750.

[5] L.-L. Xue, Y.-H. Wang, Y. Xie, P. Yao, W.-H. Wang, W. Qian, Z.-X. Huang, J. Wu, Z.-X. Xia, Effect of mutation at valine 61 on the three-dimensional structure, stability, and redox potential of cytochrome b 5, Biochemistry 38 (1999) 11961–11972.

[6] A. E. Eriksson, W. A. Baase, X.-J. Zhang, D. W. Heinz, M. Blaber, E. P. Baldwin, B. W. Matthews, Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect, Science 255 (1992) 178–183.

[7] E. Cota, S. J. Hamill, S. B. Fowler, J. Clarke, Two proteins with the same structure respond very differently to mutation: the role of plasticity in protein stability, Journal of molecular biology 302 (2000) 713–725.

[8] I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using polyphen-2, Current protocols in human genetics 76 (2013) 7–20.

[9] Y. Sergeev, N. Smaoui, R. Sui, D. Stiles, N. Gordiyenko, N. Strunnikova, I. Macdonald, The functional effect of pathogenic mutations in rab escort protein 1, Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 665 (2009) 44–50.

[10] J. D. Ramírez, M. Muñoz, C. Hernández, C. Florez, S. Gomez, A. Rico, L. Pardo, E. C. Barros, A. E. Paniz-Mondolfi, Genetic diversity among sars-cov2 strains in south america may impact performance of molecular detection, Pathogens 9 (2020) 580.

[11] V. D. Menachery, K. Debbink, R. S. Baric, Coronavirus non-structural protein 16: evasion, attenuation, and possible treatments, Virus research 194 (2014) 191–199.

[12] L. R. Lopes, G. de Mattos Cardillo, P. B. Paiva, Molecular evolution and phylogenetic analysis of sars-cov-2 and hosts ace2 protein suggest malayan pangolin as intermediary host, Brazilian Journal of Microbiology (2020) 1–7.

[13] M. Uddin, F. Mustafa, T. A. Rizvi, T. Loney, H. A. Suwaidi, A. H. H. Al-Marzouqi, A. K. Eldin, N. Alsabeeha, T. E. Adrian, C. Stefanini, et al., Sars-cov-2/covid-19: Viral genomics, epidemiology, vaccines, and therapeutic interventions, Viruses 12 (2020) 526.

[14] X. Li, Y. Song, G. Wong, J. Cui, Bat origin of a new human coronavirus: there and back again, Science China Life Sciences 63 (2020) 461–462.

[15] A. Chaudhuri, Comparative analysis of non structural protein 1 of sars-cov2 with sars-cov1 and mers-cov: An in silico study, bioRxiv (2020).

[16] A. Joshi, S. Paul, Phylogenetic analysis of the novel coronavirus reveals important variants in indian strains, BioRxiv (2020).

[17] R. Sardar, D. Satish, S. Birla, D. Gupta, Comparative analyses of sar-cov2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis, bioRxiv (2020).

[18] T.-J. Chang, D.-M. Yang, M.-L. Wang, K.-H. Liang, P.-H. Tsai, S.-H. Chiou, T.-H. Lin, C.-T. Wang, Genomic analysis and comparative multiple sequences of sars-cov2, Journal of the Chinese Medical Association 83 (2020) 537–543.

[19] Y. Ruan, C. L. Wei, A. E. Ling, V. B. Vega, H. Thoreau, S. Y. S. Thoe, J.-M. Chia, P. Ng, K. P. Chiu, L. Lim, et al., Comparative full-length genome sequence analysis of 14 sars coronavirus isolates and common mutations associated with putative origins of infection, The Lancet 361 (2003) 1779–1785.

[20] S. Perlman, J. Netland, Coronaviruses post-sars: update on replication and pathogenesis, Nature reviews microbiology 7 (2009) 439–450.

[21] W. Song, M. Gui, X. Wang, Y. Xiang, Cryo-em structure of the sars coronavirus spike glycoprotein in complex with its host cell receptor ace2, PLoS pathogens 14 (2018) e1007236.

[22] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, et al., Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor, Nature 581 (2020) 215–220.

[23] S. Xia, L. Yan, W. Xu, A. S. Agrawal, A. Algaissi, C.-T. K. Tseng, Q. Wang, L. Du, W. Tan, I. A. Wilson, et al., A pan-coronavirus fusion inhibitor targeting the hr1 domain of human coronavirus spike, Science advances 5 (2019) eaav4580.

[24] R. Hulswit, C. De Haan, B.-J. Bosch, Coronavirus spike protein and tropism changes, in: Advances in virus research, volume 96, Elsevier, 2016, pp. 29–57.

[25] M. Surjit, B. Liu, P. Kumar, V. T. Chow, S. K. Lal, The nucleocapsid protein of the sars coronavirus is capable of self-association through a c-terminal 209 amino acid interaction domain, Biochemical and biophysical research communications 317 (2004) 1030–1036.

[26] Q. Huang, L. Yu, A. M. Petros, A. Gunasekera, Z. Liu, N. Xu, P. Hajduk, J. Mack, S. W. Fesik, E. T. Olejniczak, Structure of the n-terminal rna-binding domain of the sars cov nucleocapsid protein, Biochemistry 43 (2004) 6059–6063.

[27] X. Yan, Q. Hao, Y. Mu, K. A. Timani, L. Ye, Y. Zhu, J. Wu, Nucleocapsid protein of sars-cov activates the expression of cyclooxygenase-2 by binding directly to regulatory elements for nuclear factor-kappa b and ccaat/enhancer binding protein, The international journal of biochemistry & cell biology 38 (2006) 1417–1428.

[28] Y. Hu, J. Wen, L. Tang, H. Zhang, X. Zhang, Y. Li, J. Wang, Y. Han, G. Li, J. Shi, et al., The m protein of sars-cov: basic structural and immunological properties, Genomics, proteomics & bioinformatics 1 (2003) 118–130.

[29] S. Xia, M. Liu, C. Wang, W. Xu, Q. Lan, S. Feng, F. Qi, L. Bao, L. Du, S. Liu, et al., Inhibition of sars-cov-2 (previously 2019-ncov) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion, Cell research 30 (2020) 343–355.

[30] M.-S. Chang, Y.-T. Lu, S.-T. Ho, C.-C. Wu, T.-Y. Wei, C.-J. Chen, Y.-T. Hsu, P.-C. Chu, C.-H. Chen, J.-M. Chu, et al., Antibody detection of sars-cov spike and nucleocapsid protein, Biochemical and biophysical research communications 314 (2004) 931–936.

[31] A. B. Gussow, N. Auslander, G. Faure, Y. I. Wolf, F. Zhang, E. V. Koonin, Genomic determinants of pathogenicity in sars-cov-2 and other human coronaviruses, Proceedings of the National Academy of Sciences (2020).

[32] M. Bianchi, D. Benvenuto, M. Giovanetti, S. Angeletti, M. Ciccozzi, S. Pascarella, Sars-cov-2 envelope and membrane proteins: Structural differences linked to virus characteristics?, BioMed Research International 2020 (2020).

[33] R. E. Berry, M. N. Shokhirev, A. Y. Ho, F. Yang, T. K. Shokhireva, H. Zhang, A. Weichsel, W. R. Montfort, F. A. Walker, Effect of mutation of carboxyl side-chain amino acids near the heme on the midpoint

potentials and ligand binding constants of nitrophorin 2 and its no, histamine, and imidazole complexes, Journal of the American Chemical Society 131 (2009) 2313–2327.

[34] C. N. Pace, Contribution of the hydrophobic effect to globular protein stability, Journal of molecular biology 226 (1992) 29–35.

[35] M. J. Betts, R. B. Russell, Amino acid properties and consequences of substitutions, Bioinformatics for geneticists 317 (2003) 289.

[36] A. Haimeur, G. Conseil, R. G. Deeley, S. P. Cole, Mutations of charged amino acids in or near the transmembrane helices of the second membrane spanning domain differentially affect the substrate specificity and transport activity of the multidrug resistance protein mrp1 (abcc1), Molecular pharmacology 65 (2004) 1375–1385.

[37] B. E. Bowler, K. May, T. Zaragoza, P. York, A. Dong, W. S. Caughey, Destabilizing effects of replacing a surface lysine of cytochrome c with aromatic amino acids: implications for the denatured state, Biochemistry 32 (1993) 183–190.

[38] P. Basak, S. Maitra-Majee, J. K. Das, A. Mukherjee, S. Ghosh Dastidar, P. Pal Choudhury, A. Lahiri Majumder, An evolutionary analysis identifies a conserved pentapeptide stretch containing the two essential lysine residues for rice l-myo-inositol 1-phosphate synthase catalytic activity, PloS one 12 (2017) e0185351.

[39] J. Ogawa, W. Zhu, N. Tonnu, O. Singer, T. Hunter, A. L. Ryan, G. M. Pao, The d614g mutation in the sars-cov2 spike protein increases infectivity in an ace2 receptor dependent manner, bioRxiv (2020).

[40] F. Begum, D. Mukherjee, S. Das, D. Thagriki, P. P. Tripathi, A. K. Banerjee, U. Ray, Specific mutations in sars-cov2 rna dependent rna polymerase and helicase alter protein structure, dynamics and thus function: Effect on viral rna replication, bioRxiv (2020).

[41] C. Shu, X. Huang, J. Brosius, C. Deng, Exploring potential super infection in sars-cov2 by genome-wide analysis and receptor–ligand docking (2020).

[42] S. Saurabh, S. S. Purohit, A modified ace2 peptide mimic to block sars-cov2 entry, bioRxiv (2020).

[43] M. Tiwari, D. Mishra, Investigating the genomic landscape of novel coronavirus (2019-ncov) to identify non-synonymous mutations for use in diagnosis and drug design, Journal of Clinical Virology (2020) 104441.

[44] N. Chitranshi, V. K. Gupta, R. Rajput, A. Godinez, K. Pushpitha, T. Sheng, M. Mirzaei, Y. You, D. Basavarajappa, V. Gupta, et al., Evolving geographic diversity in sars-cov2 and in silico analysis of replicating enzyme 3clpro targeting repurposed drug candidates (2020).

[45] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, Mega x: molecular evolutionary genetics analysis across computing platforms, Molecular biology and evolution 35 (2018) 1547–1549.

[46] R. E. Dickerson, The structure of cytochromec and the rates of molecular evolution, Journal of Molecular Evolution 1 (1971) 26–45.

[47] E. P. Rocha, A. Danchin, An analysis of determinants of amino acids substitution rates in bacterial proteins, Molecular biology and evolution 21 (2004) 108–116.

[48] P. C. Ng, S. Henikoff, Sift: Predicting amino acid changes that affect protein function, Nucleic acids research 31 (2003) 3812–3814.

[49] E. A. Stone, A. Sidow, Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity, Genome research 15 (2005) 978–986.

[50] J. K. Das, P. Das, K. K. Ray, P. P. Choudhury, S. S. Jana, Mathematical characterization of protein sequences using patterns as chemical group combinations of amino acids, PloS one 11 (2016) e0167651.

[51] C. Pommié, S. Levadoux, R. Sabatier, G. Lefranc, M.-P. Lefranc, Imgt standardized criteria for statistical analysis of immunoglobulin v-region amino acid properties, Journal of Molecular Recognition 17 (2004) 17–32.

[52] M. Gui, W. Song, H. Zhou, J. Xu, S. Chen, Y. Xiang, X. Wang, Cryo-electron microscopy structures of the sars-cov spike glycoprotein reveal

doi:10.20944/preprints202008.0621.v1

a prerequisite conformational state for receptor binding, Cell research 27 (2017) 119–129.

[53] A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, D. Veesler, Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein, Cell (2020).

[54] J. Cui, F. Li, Z.-L. Shi, Origin and evolution of pathogenic coronaviruses, Nature Reviews Microbiology 17 (2019) 181–192.

[55] Q. Wang, Y. Zhang, L. Wu, S. Niu, C. Song, Z. Zhang, G. Lu, C. Qiao, Y. Hu, K.-Y. Yuen, et al., Structural and functional basis of sars-cov-2 entry by using human ace2, Cell (2020).

[56] X. Ou, W. Zheng, Y. Shan, Z. Mu, S. R. Dominguez, K. V. Holmes, Z. Qian, Identification of the fusion peptide-containing region in betacoronavirus spike glycoproteins, Journal of virology 90 (2016) 5586–5600.

[57] Y. Huang, C. Yang, X.-f. Xu, W. Xu, S.-w. Liu, Structural and functional properties of sars-cov-2 spike protein: potential antivirus drug development for covid-19, Acta Pharmacologica Sinica (2020) 1–9.

[58] W. Zeng, G. Liu, H. Ma, D. Zhao, Y. Yang, M. Liu, A. Mohammed, C. Zhao, Y. Yang, J. Xie, et al., Biochemical characterization of sars-cov-2 nucleocapsid protein, Biochemical and biophysical research communications (2020).

[59] C.-Y. Chen, C.-k. Chang, Y.-W. Chang, S.-C. Sue, H.-I. Bai, L. Riang, C.-D. Hsiao, T.-h. Huang, Structure of the sars coronavirus nucleocapsid protein rna-binding dimerization domain suggests a mechanism for helical packaging of viral rna, Journal of molecular biology 368 (2007) 1075–1086.

[60] S. M. Cascarina, E. D. Ross, A proposed role for the sars-cov-2 nucleocapsid protein in the formation and regulation of biomolecular condensates, The FASEB Journal (2020).

Competing interests

The authors declare no competing interests.