doi:10.20944/preprints202008.0621.v2

A study on non-synonymous mutational patterns in structural proteins of SARS-CoV-2

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

SARS-CoV-2 is mutating and creating divergent variants across the world. An in-depth investigation of the amino acid substitution in the genomic signature of SARS-CoV-2 proteins is highly essential for understanding its host adaptation and infection biology. A total of 9587 SARS-CoV-2 structural protein sequences collected from 49 different countries are used to characterize protein-wise variants, substitution pattern (type and location), and major substitution changes. The majority of the substitutions are distinct, occurred mostly in a particular location, and leads to a change in amino acid's biochemical properties. In terms of mutational changes, Envelope (E) and Membrane (M) proteins are relatively stable than Nucleocapsid (N) and Spike (S) proteins. Several co-occurrence substitutions are observed, particularly in S and N proteins. Substitution specific to active sub-domains reveals that Heptapeptide Repeat, Fusion peptides, Transmembrane in S protein, and N-terminal and C-terminal domains in N protein are remarkably mutated, and also found few deleterious mutations in these domains.

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

1. Introduction

The outbreak of novel Severe Acute Respiratory Syndrome Corona Virus (SARS-CoV-2), causing the disease COVID-19. With the passage of time a good number of variants are reported so far across the globe. Though the rate of mutation in SARS-CoV-2 observed to be relatively low, even then the variants of SARS-CoV-2 are differed by a number of 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 variants too [2, 3]. Couple of studies reveal that mutation can triggers enzyme motion in dihydrofolate reductase [4], impact on three-dimensional structure, stability and redox potential [5], hydrophobic effect [6], functional diversity in the proteins. The same may be observed even in the structurally similar proteins [7, 8, 9].

The genetic diversity in SARS-CoV-2 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 the SARS-CoV-2 proteins, which ultimately helps in designing more stable small molecules that may binds viral proteins.

Mutation is the key factor that trigger a virus to switch hosts [11]. SARS-CoV-2 genome shows different degree of similarity with other coronaviruses. SARS-CoV-2 genome is highly

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

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, such as 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 which encodes structural and accessory proteins [19]. SARS-CoV-2 proteins play a diverse functional role. Mutations in structural proteins can alter various functionality of the virus, more importantly protein from this class initiate host jumping mechanism. 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 Membrane protein include promoting membrane fusion, regulating viral replication, packing genomic RNA into viral particles, interaction with other proteins [28, 29, 30]. Both the spike and nucleocapsid proteins are in the main target for antibody detection [31]. Many of the sequence variability features of structural proteins are still unknown and need 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 [32]. 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 [32] 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 [33]. 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 [34, 35, 36, 37, 38, 39]. 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) [40, 41, 42, 43]. Therefore, these studies are important to understand the clinical presentation and spread of the disease, and also useful for antiviral drug design [44, 45].

In this study, we focus on sequence variability of worldwide SARS-CoV-2 genome, particularly four structural proteins. For each protein, we identify unique variants. 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.

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

2. Materials and methods

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 variation due to mutation. In this section we report the sequence used and various exploratory analysis performed on the sequences.

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 sequences with undetermined character present within the sequence for our analysis, and obtain 9058, 8954, 8271 and 7009 number of sequences for Spike (S), Envelope (E), Membrane (M), Nucleocapsid (N) protein, respectively. Collected sequences are then grouped based on similarity to isolate unique variants for further analysis.

2.2. Multiple sequence alignment

The unique list of sequences (one representative from each group) are aligned using multiple sequence aligner (MSA) to observe possible insertion or deletion in amino acid residues. We select 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) as a reference sequence. We use MEGAX [46] for MSA. We observe few sequences with indels in Spike and Nucleocapsid proteins, which we separately analyze as reported in *Supplementary-A (Figure S1 and S2)*, and all the remaining sequences are considered for subsequent analysis.

2.3. Amino acid substitution identification

The amino acid substitution (or non-synonymous mutation) is a common phenomenon in virus genome and the rate of substitution is mainly depends on the protein's expression level, functional category, metabolic costs, hydrophobicity and electrostatic, physicochemical properties, annotated active or binding site [47, 48, 49, 39].

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 for twenty (20) amino acids ($20 \times 20 - 20 = 380$) without considering synonymous substitutions and categorizing them based on similar substitution patterns.

2.4. Investigating the change in biochemical properties

Amino acid substitutions due to non-synonymous mutation may change the biochemical properties of the target proteins. We categorize observed amino acid substitutions 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).

Our major goal is to highlight what kind of biochemical properties are majorly changed (quantitatively) due to substitution.

2.5. Functional domains of SARS-CoV-2 structural proteins

The SARS-CoV-2 structural protein (particularly S and N) encompasses several sub-domains responsible for a 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-CoVN 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. Primarily it consists 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, might help understand the virulence power of SARS-CoV-2 . Therefore, we try to highlight the substitutions in different functional domains in the next section.

2.6. Software tools and programming

For predicting whether a mutation is deleterious or neutral, we use software tool *PROVEAN* [2]. *PROVEAN* (Protein Variation Effect Analyzer), a web server software tool, is used to predict any non-synonymous amino acid substitution or indel impacts on the biological function of a protein [59]. We utilize a recently developed mutation simulation tool [60] which generates fine-grained simulated random mutations in any genome. We use Python scripting 3.7 for quantitative analysis.

3. Results and Discussion

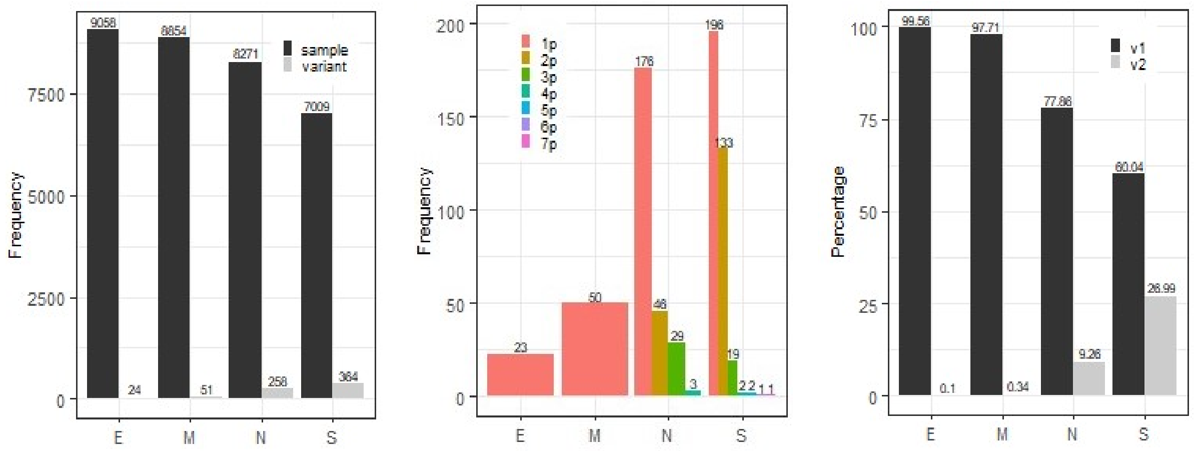
3.1. Grouping identical variants

At first, similar sequences are grouped for each structural protein. We observe 24, 51, 258, 364 number of groups (distinct variants) for E, M, N and S protein, respectively (Figure 1(a)). The distinct variants are represented as $v1, v2, \cdots, vk$ (k is the number of distinct variants

[2] http://provean.jcvi.org/index.php

observed in each protein category). If we consider the group size for different proteins, we observe that more than 95% of the samples from E and M proteins sharing two groups and 85% samples from N and S proteins distributed in two groups (Figure 1(c)). 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 (sample frequency) for each structural protein is listed in *Supplementary-A (Table S1)*.

While studying the number of substitutions (or substituted positions) occurs in each variant, interestingly we observe maximum of seven (07) substituted positions. We represent number of substituted positions by numbering, for example, one substituted positions by $1p$, two substituted positions by $2p$, and so on. Out of which single substitution occurring most commonly, whereas seven number of substitutions are rarely occurring in different candidate proteins (Figure 1(b)). In case of E and M protein, we observe only single amino acid substitution position in each variant. For N protein, number of substitutions varies from $1p$ to $4p$ numbers, whereas for S protein, count varies from $1p$ to $7p$.



(a) Collected samples vs. distinct variant.

(b) Distribution of substitutions by positions.

(c) The percentage of sample frequency of top two variants.

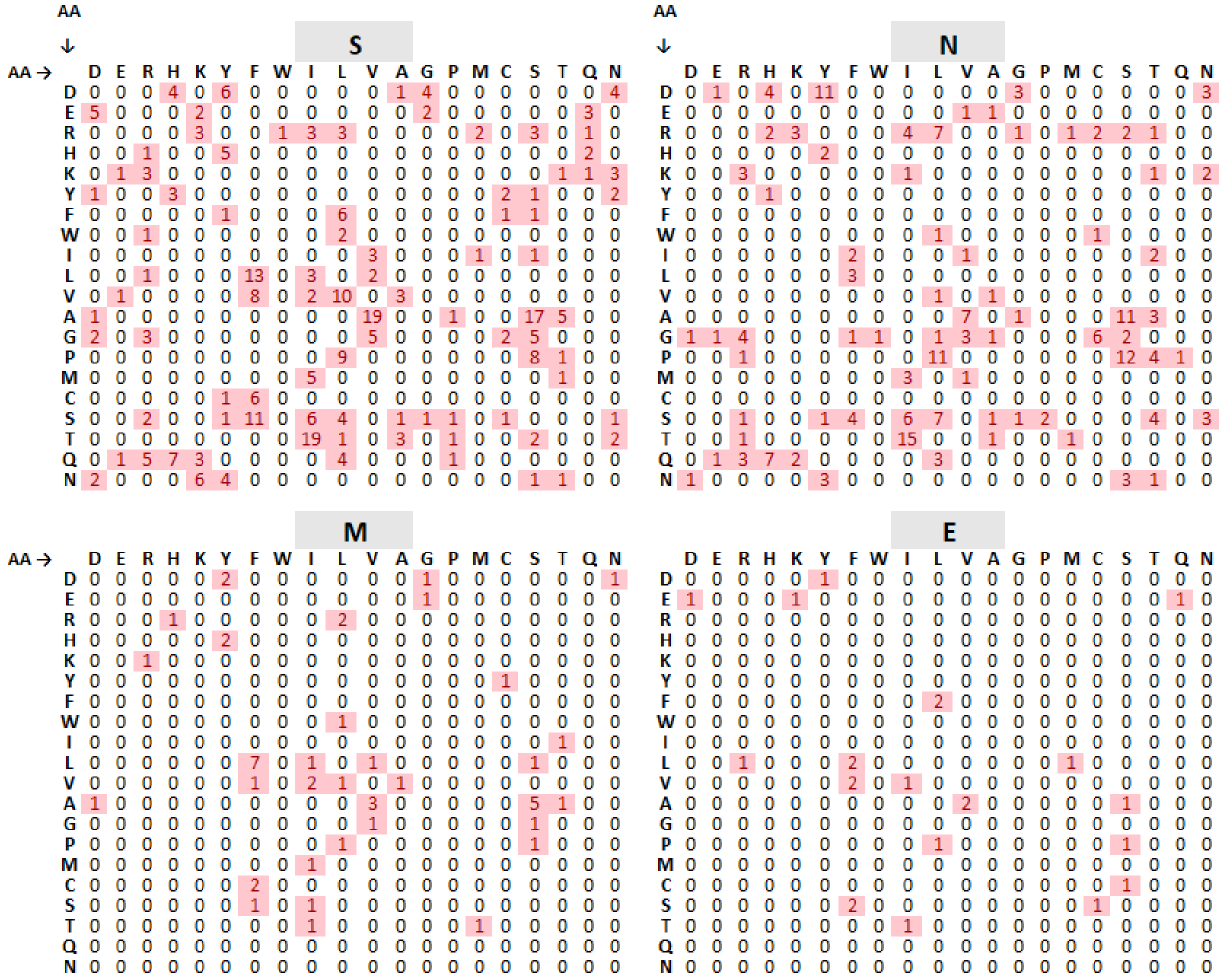
Figure 1: The quantification of observed variant and non-synonymous substitutions positions in each SARS-CoV-2 structural protein. (a) collected sample with observed distinct variants; (b) Variant frequency with number of substituted positions; (c) Percentage of sample frequency for the top two variant class ($v1$ and $v2$).

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 2. In general, there could be $20 \times 20 - 20 = 380$ possible unique non synonymous 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 3, a total 316, 217, 50, 23 number of amino acid substitutions in S, N, M, E protein, respectively. 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, these findings show that mutation location in the N protein is most random and difficult to localize or predict its site of alternation. M protein relatively most stable in such scenario.

doi:10.20944/preprints202008.0621.v2



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 4).

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 are observed. A position-wise substitutions in each proteins is shown in Figure 4. 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) as substitutions are found in position 203 distributed across 36 variants (*Supplementary-B*).

The top few substitutions (based on sample frequency) are observed common in the majority of the countries in different SARS-CoV-2 structural proteins (Figure 5), which are $P > L(71)$ in E protein; $T > M(175)$ in M protein; $R > K(203)$, $G > R(204)$ and $S > L(194)$ in N protein; $D > G(614)$ and $L > F(54)$ in S protein. Several substitutions are observed, mostly specific to in different countries, with sample frequency 3 or more as depicted in the Figure 5. The complete list of substitutions patterns observed in different countries (with sample frequency ≥ 1) can be seen from *Supplementary-C*.

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 and 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

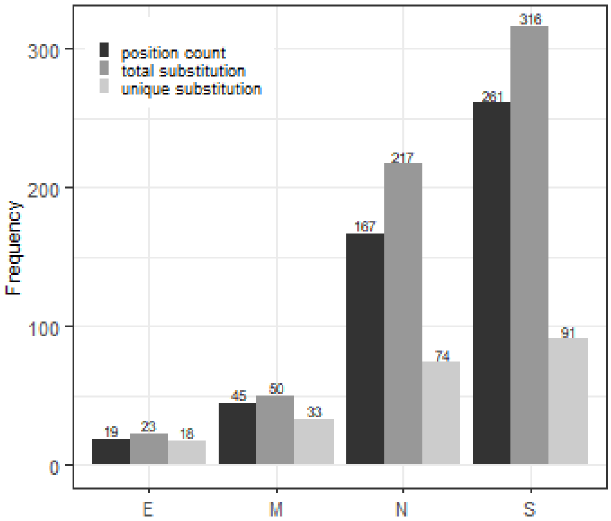


Figure 3: The figure shows observed total number of substitution irrespective of positions and variants, total substitution in all positions, observed substitutions by positions, and unique substitution count for each structural protein (S, N, M, E).

position (Figure 2). We observe maximum substitutions for $A > V$, which occurs in 19 different positions followed by $A > S$ (17 different position) and $L > F$ (13 different positions) in S protein. In case of N protein, we observe maximum substitution for $T > I$, which occurs in 15 different positions followed by $P > S$ (12 different position) and $P > L$, $A > S$ and $D > Y$ (11 different positions). In case of M protein, we observe maximum substitution for $A > S$ (15 different positions), and we do not observe more than two substitutions in any particular position in case of E protein.

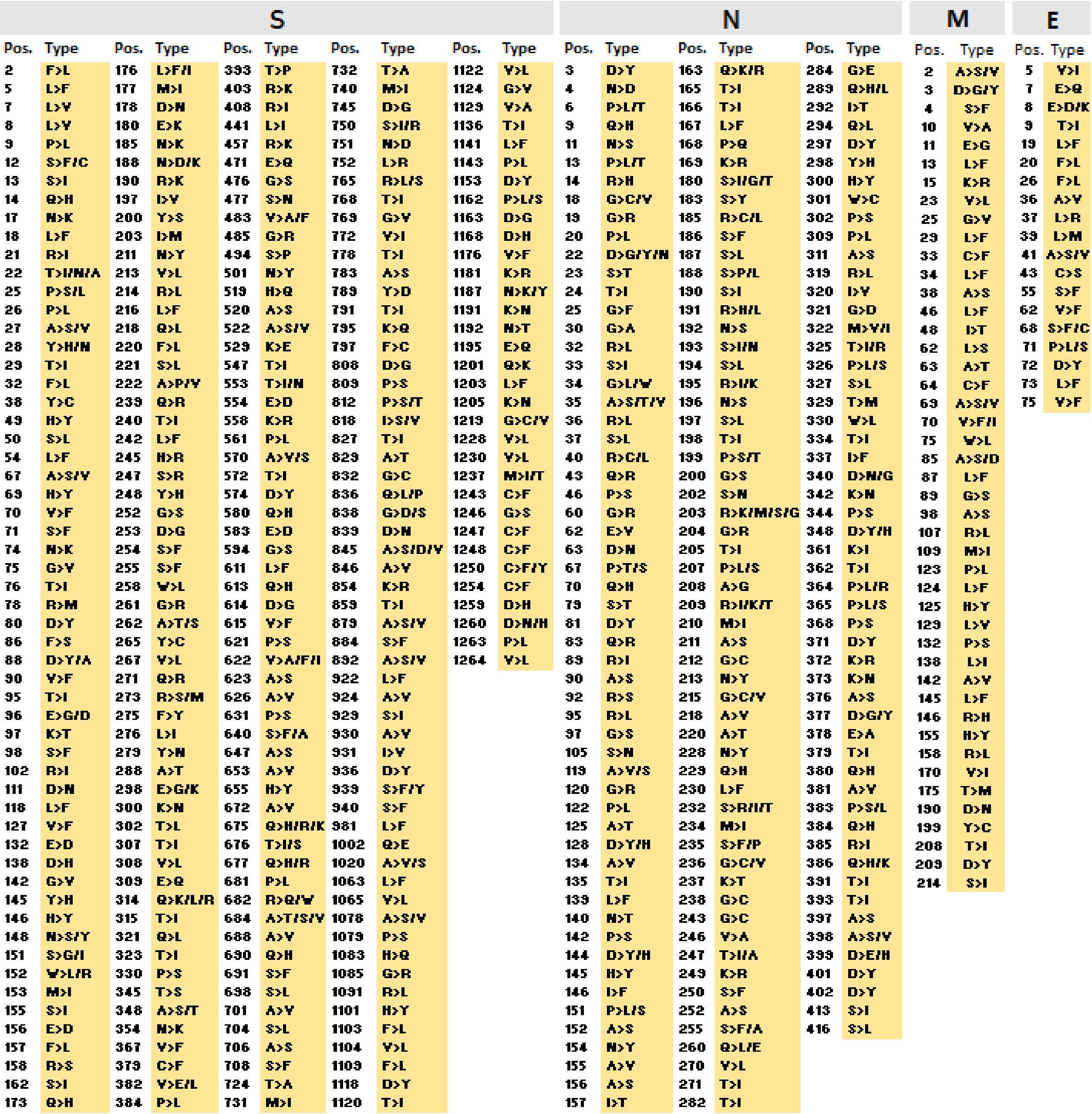
Towards understanding amino acid change by substitutions type, we calculate the percentage by observing all substitutions occurring in all positions (discussed above). There are two categories of substitution we consider, one to many and many to one fixing a particular source or target amino acid. For both the categories, the total substitution count, the number of distinct substitution type (from Figure 2), and their percentages for each category of proteins are calculated. The top 5 substitutions are reported in Table 1. Majority of the amino acid changes from L to other in both the envelope (E) protein ($\approx 17\%$, 3 distinct types) and membrane (M) protein ($\approx 20\%$, 4 distinct types). In case of nucleocapsid (N) protein, substitutions occur for S to other ($\approx 13\%$, 10 distinct types). For spike (S) protein, we observe A to other change is the maximum ($\approx 14\%$, 5 distinct types). However, maximum distinct substitution are observed in S to others (10 distinct types).

We also observe several co-occurrence substitutions in Spike (S) and Nucleocapsid (N) proteins (Table 2). For example, dominant substitution $D > G(614)$ in S protein is co-occurred with $L > F(54)$ in 53 samples and $L > F(5)$ in 39 samples. Similarly, in N protein, dominant substitution $R > K(203)$ is co-occurred with $G > R(204)$ in 766 samples.

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

A non-synonymous substitution that alters the amino acid, in tern alter the biochemical properties of the amino acid. In this section, we categorise the substitution type by observing their chemical properties of amino acids. We consider two kinds of biochemical properties, eight chemical properties of amino acids and three Hydropathy class of amino acid as discussed in Section 2.4.

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 19 ($\approx 83\%$) of substitutions in S, N,



S										N						M		E	
Pos.	Type	Pos.	Type	Pos.	Type	Pos.	Type	Pos.	Type	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	3	D>Y	163	Q>K/R	284	G>E	2	A>S/V	5	V>I
5	L>F	177	M>I	403	R>K	740	M>I	1124	G>V	4	N>D	165	T>I	289	Q>H/L	3	D>G/Y	7	E>Q
7	L>V	178	D>N	408	R>I	745	D>G	1129	V>A	6	P>L/T	166	T>I	292	I>T	4	S>F	8	E>D/K
8	L>V	180	E>K	441	L>I	750	S>I/R	1136	T>I	9	Q>H	167	L>F	294	Q>L	10	V>A	9	T>I
9	P>L	185	N>K	457	R>K	751	N>D	1141	L>F	11	N>S	168	P>Q	297	D>Y	11	E>G	19	L>F
12	S>F/C	188	N>D/K	471	E>Q	752	L>R	1143	P>L	13	P>L/T	169	K>R	298	Y>H	13	L>F	20	F>L
13	S>I	190	R>K	476	G>S	765	R>L/S	1153	D>Y	14	R>H	180	S>I/G/T	300	H>Y	15	K>R	26	F>L
14	Q>H	197	I>V	477	S>N	768	T>I	1162	P>L/S	18	G>C/V	183	S>Y	301	W>C	23	V>L	36	A>V
17	N>K	200	Y>S	483	V>A/F	769	G>V	1163	D>G	19	G>R	185	R>C/L	302	P>S	25	G>V	37	L>R
18	L>F	203	I>M	485	G>R	772	V>I	1168	D>H	20	P>L	186	S>F	309	P>L	29	L>F	39	L>M
21	R>I	211	N>Y	494	S>P	778	T>I	1176	V>F	22	D>G/Y/N	187	S>L	311	A>S	33	C>F	41	A>S/V
22	T>I/N/A	213	V>L	501	N>Y	783	A>S	1181	K>R	23	S>T	188	S>P/L	319	R>L	34	L>F	43	C>S
25	P>S/L	214	R>L	519	H>Q	789	Y>D	1187	N>K/Y	24	T>I	190	S>I	320	I>V	38	A>S	55	S>F
26	P>L	216	L>F	520	A>S	791	T>I	1191	K>N	25	G>F	191	R>H/L	321	G>D	46	L>F	62	V>F
27	A>S/V	218	Q>L	522	A>S/V	795	K>Q	1192	N>T	30	G>A	192	N>S	322	M>V/I	48	I>T	68	S>F/C
28	Y>H/N	220	F>L	529	K>E	797	F>C	1195	E>Q	32	R>L	193	S>I/N	325	T>I/R	62	L>S	71	P>L/S
29	T>I	221	S>L	547	T>I	808	D>G	1201	Q>K	33	S>I	194	S>L	326	P>L/S	63	A>T	72	D>Y
32	F>L	222	A>P/V	553	T>I/N	809	P>S	1203	L>F	34	G>L/W	195	R>I/K	327	S>L	64	C>F	73	L>F
38	Y>C	239	Q>R	554	E>D	812	P>S/T	1205	K>N	35	A>S/T/V	196	N>S	329	T>M	69	A>S/V	75	V>F
49	H>Y	240	T>I	558	K>R	818	I>S/V	1219	G>C/V	36	R>L	197	S>L	330	W>L	70	V>F/I		
50	S>L	242	L>F	561	P>L	827	T>I	1228	V>L	37	S>L	198	T>I	334	T>I	75	W>L		
54	L>F	245	H>R	570	A>V/S	829	A>T	1230	V>L	40	R>C/L	199	P>S/T	337	I>F	85	A>S/D		
67	A>S/V	247	S>R	572	T>I	832	G>C	1237	M>I/T	43	Q>R	200	G>S	340	D>N/G	87	L>F		
69	H>Y	248	Y>H	574	D>Y	836	Q>L/P	1243	C>F	46	P>S	202	S>N	342	K>N	89	G>S		
70	V>F	252	G>S	580	Q>H	838	G>D/S	1246	G>S	60	G>R	203	R>K/M/S/G	344	P>S	98	A>S		
71	S>F	253	D>G	583	E>D	839	D>N	1247	C>F	62	E>V	204	G>R	348	D>Y/H	107	R>L		
74	N>K	254	S>F	594	G>S	845	A>S/D/V	1248	C>F	63	D>N	205	T>I	361	K>I	109	M>I		
75	G>V	255	S>F	611	L>F	846	A>V	1250	C>F/Y	67	P>T/S	207	P>L/S	362	T>I	123	P>L		
76	T>I	258	W>L	613	Q>H	854	K>R	1254	C>F	70	Q>H	208	A>G	364	P>L/R	124	L>F		
78	R>M	261	G>R	614	D>G	859	T>I	1259	D>H	79	S>T	209	R>I/K/T	365	P>L/S	125	H>Y		
80	D>Y	262	A>T/S	615	V>F	879	A>S/V	1260	D>N/H	81	D>Y	210	M>I	368	P>S	129	L>V		
86	F>S	265	Y>C	621	P>S	884	S>F	1263	P>L	83	Q>R	211	A>S	371	D>Y	132	P>S		
88	D>Y/A	267	V>L	622	V>A/F/I	892	A>S/V	1264	V>L	89	R>I	212	G>C	372	K>R	138	L>I		
90	V>F	271	Q>R	623	A>S	922	L>F			90	A>S	213	N>Y	373	K>N	142	A>V		
95	T>I	273	R>S/M	626	A>V	924	A>V			92	R>S	215	G>C/V	376	A>S	145	L>F		
96	E>G/D	275	F>Y	631	P>S	929	S>I			95	R>L	218	A>V	377	D>G/Y	146	R>H		
97	K>T	276	L>I	640	S>F/A	930	A>V			97	G>S	220	A>T	378	E>A	155	H>Y		
98	S>F	279	Y>N	647	A>S	931	I>V			105	S>N	228	N>Y	379	T>I	158	R>L		
102	R>I	288	A>T	653	A>V	936	D>Y			119	A>V/S	229	Q>H	380	Q>H	170	V>I		
111	D>N	298	E>G/K	655	H>Y	939	S>F/Y			120	G>R	230	L>F	381	A>V	175	T>M		
118	L>F	300	K>N	672	A>V	940	S>F			122	P>L	232	S>R/I/T	383	P>S/L	190	D>N		
127	V>F	302	T>L	675	Q>H/R/K	981	L>F			125	A>T	234	M>I	384	Q>H	199	Y>C		
132	E>D	307	T>I	676	T>I/S	1002	Q>E			128	D>Y/H	235	S>F/P	385	R>I	208	T>I		
138	D>H	308	V>L	677	Q>H/R	1020	A>V/S			134	A>V	236	G>C/V	386	Q>H/K	209	D>Y		
142	G>V	309	E>Q	681	P>L	1063	L>F			135	T>I	237	K>T	391	T>I	214	S>I		
145	Y>H	314	Q>K/L/R	682	R>Q/W	1065	V>L			139	L>F	238	G>C	393	T>I				
146	H>Y	315	T>I	684	A>T/S/V	1078	A>S/V			140	N>T	243	G>C	397	A>S				
148	N>S/Y	321	Q>L	688	A>V	1079	P>S			142	P>S	246	V>A	398	A>S/V				
151	S>G/I	323	T>I	690	Q>H	1083	H>Q			144	D>Y/H	247	T>I/A	399	D>E/H				
152	W>L/R	330	P>S	691	S>F	1085	G>R			145	H>Y	249	K>R	401	D>Y				
153	M>I	345	T>S	698	S>L	1091	R>L			146	I>F	250	S>F	402	D>Y				
155	S>I	348	A>S/T	701	A>V	1101	H>Y			151	P>L/S	252	A>S	413	S>I				
156	E>D	354	N>K	704	S>L	1103	F>L			152	A>S	255	S>F/A	416	S>L				
157	F>L	367	V>F	706	A>S	1104	V>L			154	N>Y	260	Q>L/E						
158	R>S	379	C>F	708	S>F	1109	F>L			155	A>V	270	V>L						
162	S>I	382	V>E/L	724	T>A	1118	D>Y			156	A>S	271	T>I						
173	Q>H	384	P>L	731	M>I	1120	T>I			157	I>T	282	T>I						

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).

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 6(a)).

In case of, more than 50% substitutions (except E protein), we observe a change of one Hydropathy class to another. We observe 50 out of 91 (≈ 55%), 46 out of 74 (≈ 62%), 17 out of 33 (≈ 52%), 8 out of 18 (≈ 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 6(b)).

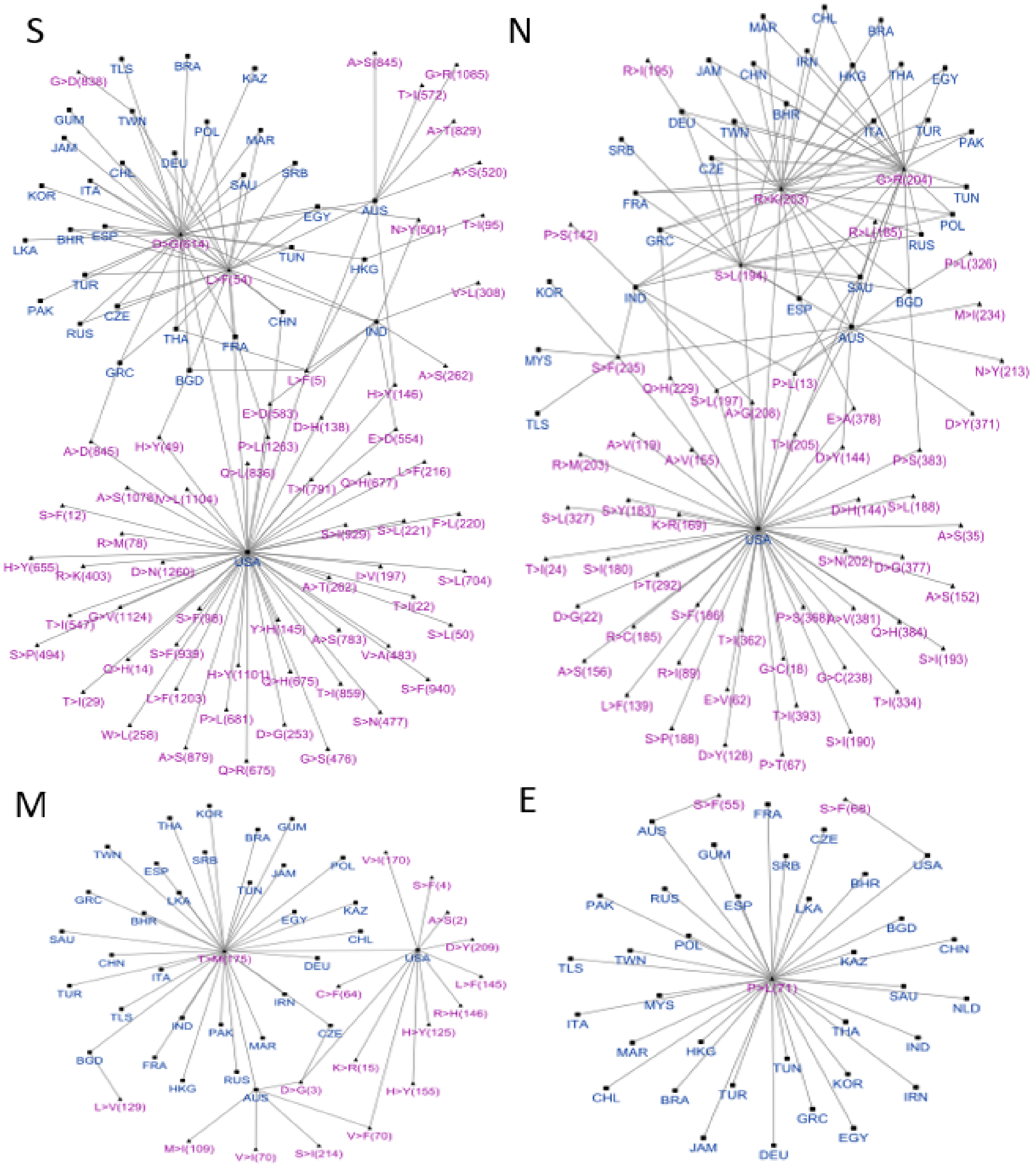


Figure 5: The bipartite graph shows the various substitutions (with sample frequency ≥3) observed in four structural proteins (S, N, M, E) in the samples from different countries. There are two types of node in the figure. The blue node represents country (labeled with three letters country code), whereas the light red node represents substitutions with position indicated within parenthesis. The edge between two kinds of nodes depicts existence of a particular substitution in the sequence from a country.

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

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 3.

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 (≈ 22% locations) followed by Transmembrane (TM) domain (≈ 17% locations).

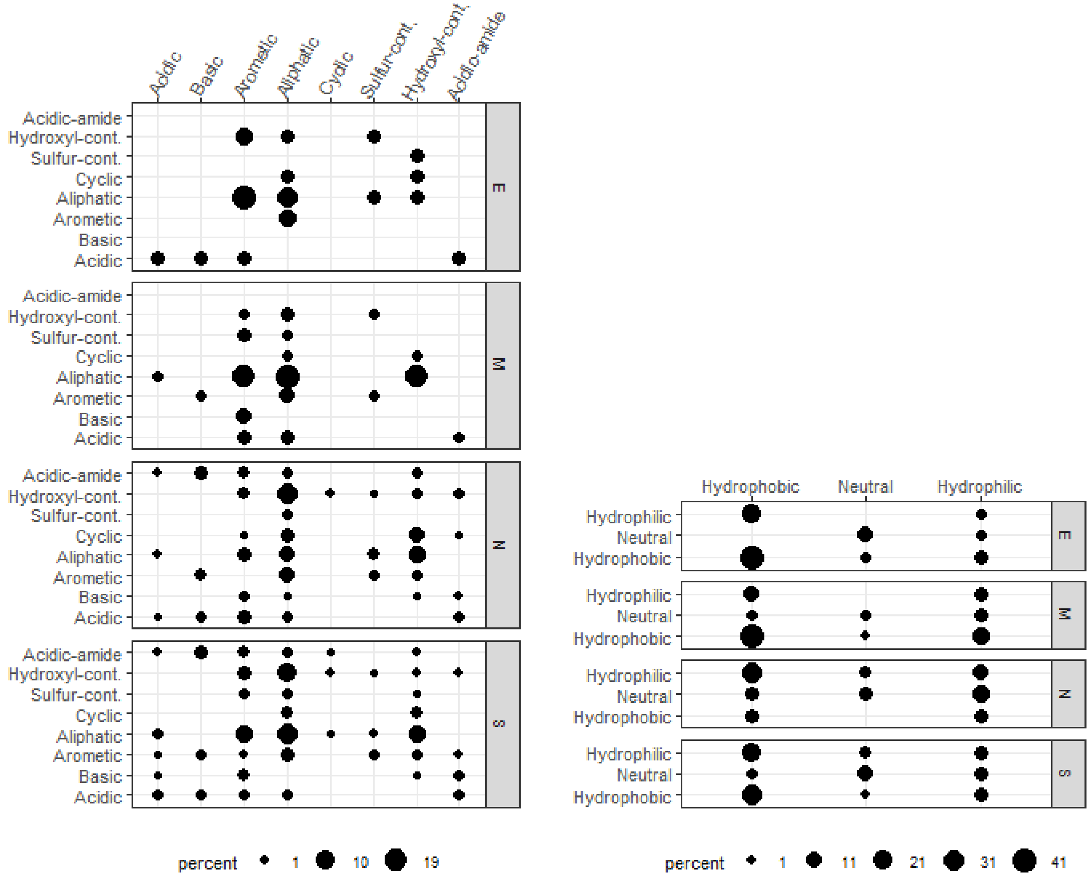
Table 1: The table shows top 5 substitution type and associative quantitative information for two categories. For a substitution $X > Y$, in the the first category, X is fixed and Y can be any amino acid, and in the second category, X is any amino acid and Y is a fixed.

X> any					Any >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

Table 2: The dominant substitutions that are co-occurred with other substitutions (with sample frequency 5 or more) in Spike(S) and Nucleocapsid (N) protein.

Protein	**Substitution (a)**	**Co-occurred subst-itution with (a)**	**Sample frequency**
S	D>G(614)	L>F(54)	53
		L>F(5)	39
		S>N(477)	37
		H>Y(146)	21
		S>L(221)	17
		P>L(681)	16
		D>G(253)	14
		N>Y(501)	13
		Q>H(677)	13
		T>I(572)	11
		E>D(583)	11
		Q>H(14)	8
		R>M(78)	8
		V>L(308)	8
		R>K(403)	8
		A>D(845)	7
		P>L(1263)	7
		F>L(220)	6
		D>H(138)	6
		H>Y(49)	6
		S>F(939)	6
		T>I(859)	6
		M>I(153)	5
		G>V(1124)	5
		L>F(1203)	5
		E>D(554);D>H(138)	10
N	R>K(203)	G>R(204)	766
		G>R(204);I>T(292)	16
		G>R(204);Q>H(229)	10
		G>R(204);T>I(205)	5
	A>G(208)	T>I(393)	6
	E>V(62)	S>L(194)	6
	P>L(13)	S>L(197)	28

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. We then predict whether the mutations/substitutions in functional domains are deleterious or neutral. We find



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

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

Figure 6: The trends of substitution type (percentage) shown by percentage for two categories of amino acid biochemical property. The circle in the figure indicates the proportion of changes percentage from one group to another.

eleven (11) deleterious mutations in functional domain of S protein (*Supplementary-A(Table S2)*), which are distributed in RBD domain ($C379F, V382E$); FP domain ($F797C$); HR1 domain ($A924V, S929I, A930V, D936Y, S939F, S939Y, S940F$); HR2 domain ($L1203F$). 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.

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 [61]. Although the molecular mechanism of SARS-CoV-2 N protein is yet to explore thoroughly [62]. There are total of twenty (25) deleterious mutations are observed, 13 in CTD domain and 12 in NTD domain (*Supplementary-A (Table S3)*).

3.6. Comparison of real vs. simulated mutations

Various simulation tools are developed assuming certain probabilistic distributions followed during genetic mutations. It may be interesting to verify whether mutations in SARS-CoV-2 follow similar distributions or not. We use very recently developed mutation simulation tool [60] which generates fine-grained simulated random mutations in any genome. We consider reference sequence (*Wuhan-Hu-1*) and mutation rate (per base) (Table 4) and use *ARGS mode* as defined in [60] to generate simulated sequences.

Table 3: Quantification of non-synonymous amino acid substitutions observed in different functional domains of S and N protein.

Protein	Domain	Substitution (freq)	# mutated position	Domain (%)	Overall (%)
S	RBD	Q314K(1), Q314L(1), Q314R(1), T315I(1), Q321L(1), T323I(2), P330S(1), T345S(1), A348S(1), A348T(1), N354K(1), V367F(4), C379F(1), V382E(1), V382L(1), P384L(1), T393P(1), R403K(1), R408I(1), L441I(1), R457K(1), E471Q(1), G476S(2), S477N(1), V483A(2), V483F(1), G485R(2), S494P(1), N501Y(1), H519Q(2), A520S(2), A522S(1), A522V(1), K529E(1)	28	12.55	10.72
	HR1	L922F(1), A924V(1), S929I(1), A930V(1), I931V(1), D936Y(1), S939F(2), S939Y(1), S940F(1), L981F(1)	9	12.85	3.44
	HR2	D1163G(1), D1168H(1), V1176F(1), K1181R(1), N1187K(1), N1187Y(1), K1191N(2), N1192T(1), E1195Q(1), Q1201K(1), L1203F(1), K1205N(1)	11	22.0	4.21
	FP	Y789D(1), T791I(5), K795Q(1), F797C(1)	4	22.22	1.53
	TM	G1219C(1), G1219V(1), V1228L(2), V1230L(1), M1237I(1), M1237T(1)	4	16.52	1.53
N	NTD	P46S(1), G60R(1), E62V(2), D63N(1), P67T(1), P67S(1), Q70H(1), S79T(1), D81Y(1), Q83R(1), R89I(1), A90S(1), R92S(1), R95L(1), G97S(1), S105N(1), A119V(2), A119S(1), G120R(1), P122L(1), A125T(1), D128Y(1), D128H(1), A134V(1), T135I(1), L139F(2), N140T(1), P142S(1), D144Y(3), D144H(2), H145Y(1), I146F(1), P151L(1), P151S(1), A152S(3), N154Y(1), A155V(1), A156S(3), I157T(1), Q163K(1), Q163R(1), T165I(1), T166I(1), L167F(1), P168Q(1), K169R(1)	40	29.6	23.95
	CTD	T247I(1), T247A(1), K249R(1), S250F(1), A252S(1), S255F(1), S255A(1), Q260L(1), Q260E(1), V270L(2), T271I(2), T282I(2), G284E(1), Q289H(1), Q289L(1), I292T(2), Q294L(1), D297Y(1), Y298H(1), H300Y(1), W301C(1), P302S(1), P309L(1), A311S(1), R319L(1), I320V(1), G321D(1), M322V(1), M322I(1), T325I(1), T325R(1), P326L(1), P326S(1), S327L(2), T329M(1), W330L(1), T334I(2), I337F(1), D340N(1), D340G(1), K342N(1), P344S(1), D348Y(4), D348H(1), K361I(1), T362I(2)	37	31.89	22.15

The mutation rate is the number of mutations per genetic unit at (gene, base) per unit time (year, million years, or generation) [63]. We calculate mutation rate (μ) based on number of mutations and distinct variants for each structural protein, which is calculated as follows:[3]

$$\mu = \frac{m}{t}.$$

where, m is the total number of mutations in k number of variants each of length l. In case of per base and per gene mutation rate, $t = k \times l$, and $t = k$, respectively. The per base mutation rate is relatively high for E protein followed by M, N, S, whereas mutation rate per gene is high for S and N genes (Table 4).

Table 4: The mutation rate (per base and per gene) in four SARS-CoV-2 structural proteins. For calculating mutation rate, we considered mutation (non-synonymous substitutions) observed in all the variants of respective protein.

Gene	# variant	# mutation	Length (nucleotide)	Mutation rate (per base)	Mutation rate (per gene)
E	23	23	228	0.004386	1
M	50	50	669	0.001495	1
N	254	367	1260	0.001147	1.444882
S	354	691	3822	0.000510	1.951977

A total of five set of distinct simulated sequences (with non-synonymous mutations) are generated. Each set is consisting of equal number of sequences as we have real and distinct mutated sequences (or variant) for each structural protein of SARS-CoV-2 (Table 4).

We compare two sets by the Jaccard index, also known as the Jaccard similarity coefficient [64]. It is used to calculate similarity and diversity of sample sets. Given two sets of real and simulated mutations/substitutions, R_m and S_m, respectively, it can be calculated as follows.

$$J(R_m, S_m) = \frac{|R_m \cap S_m|}{|R_m \cup S_m|} = \frac{|R_m \cap S_m|}{|R_m| + |S_m| - |R_m \cap S_m|}. \tag{1}$$

[3] http://www.biology.arizona.edu/evolution/working/act/mutation/population.html

The value of $J(R_m, S_m)$ lies between 0 and 1, i.e. $0 \leq J(R_m, S_m) \leq 1$. The higher value of $J(R_m, S_m)$ indicates more similarities between the sample sets.

We report in Figure 7 the similarity scores for each protein. It can be observed that the Jaccard similarity for Nucleocapsid (N) and Spike (S) proteins substitutions is approximately 0.6 and 0.5, respectively. However, Envelope (E) and Membrane (M) proteins show low similarity. The low similarity score further reflects that the mutations are non-uniform (as ARGS mode assumes uniform distribution) i.e. probability of mutations is not same in all positions.

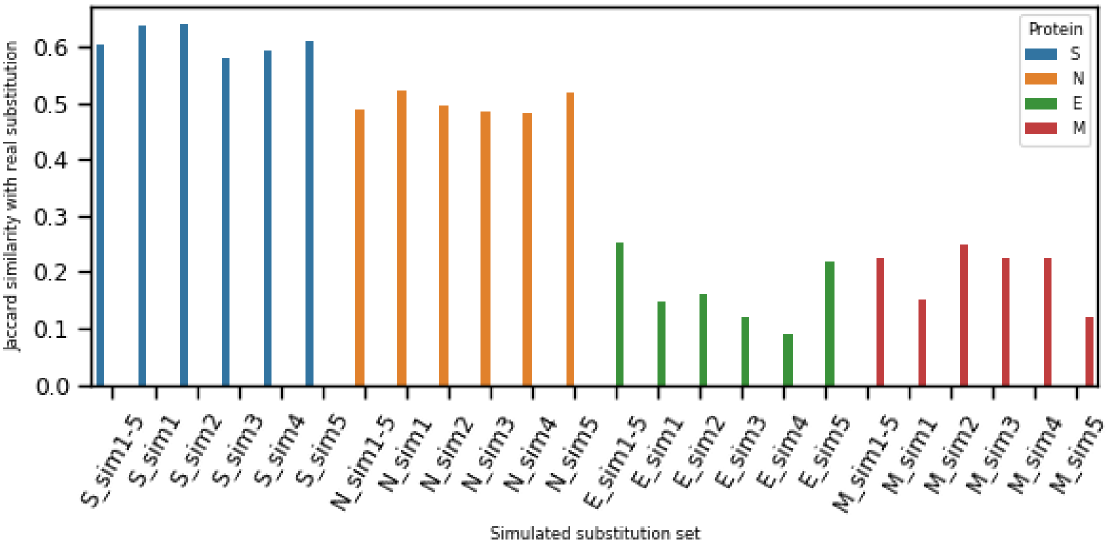


Figure 7: Comparison of real and simulated substitutions. The Jaccard similarity score are shown for six set of compared simulated substitutions (*sim1,sim2,...sim5* and*sim1-5*) with the real substitution set.

4. Conclusion

In this work, we performed a sequence analysis on the structural proteins of SARS-CoV-2 , reported from different parts of the world. All the unique sequences are grouped, and the variations are analyzed based on *Wuhan-Hu-1* SARS-CoV-2 sequence (as reference). We highlighted various commonly and rarely occurring amino acids substitution 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 the above variations specific to any particular country.

Although we observed much more unique variants in S protein than N protein, in terms of substitutions changes, N protein is much vulnerable as it shows the substitution in 40% positions. The majority of substitutions type for all four proteins are observed distinct, and they occur mostly in a particular location. The substitution type $A > V$ is very common for S, N, and M protein. We have highlighted few dominant substitutions that are co-occurred with several substitutions, particularly in S and N proteins. We further noted that the majority of the substitutions lead chemical group change from Aromatic and Aliphatic, Aliphatic and Hydroxyl-containing, and from Hydrophilic to Hydrophobic. A few substitutions in functional domains of S and N proteins are also highlighted found to be deleterious.

The overall study summarizes diversity amongst viruses sequenced early in the pandemic. We believe that the current findings will help better understand the disease pathogenesis of COVID-19 followed by suitable and relatively stable small molecule identification that may bind susceptible structural proteins like Spike (S) protein that is frequently changing.

Author contributions

J.K.D. and S.R. conceived and designed the research. J.K.D. collected, performed and analyzed the data. J.K.D. wrote the original manuscript. S.R. reviewed and edited. All authors read and approved the final manuscript.

Conflict of interest statement

The authors declare no conflict of interest.

Data availability statement

All the relevant data are presented withing the supplementary files.

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] A. Naskalska, A. Dabrowska, A. Szczepanski, A. Milewska, K. P. Jasik, K. Pyrc, Membrane protein of human coronavirus nl63 is responsible for interaction with the adhesion receptor, Journal of virology 93 (2019) e00355–19.

[31] 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.

[32] 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).

[33] 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).

[34] 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.

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

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

[37] 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.

[38] 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.

[39] 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.

[40] 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).

[41] 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).

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

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

[44] 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.

[45] 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).

[46] 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.

[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 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] Y. Choi, A. P. Chan, Provean web server: a tool to predict the functional effect of amino acid substitutions and indels, Bioinformatics 31 (2015) 2745–2747.

[60] M. Kühl, B. Stich, D. Ries, Mutation-simulator: fine-grained simulation of random mutations in any genome, Bioinformatics (2020).

[61] 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.

[62] 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).

[63] J. F. Crow, The high spontaneous mutation rate: is it a health risk?, Proceedings of the National Academy of Sciences 94 (1997) 8380–8386.

[64] P. Jaccard, The distribution of the flora in the alpine zone. 1, New phytologist 11 (1912) 37–50.