International Journal of Biological Macromolecules 266 (2024) 131054

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac

A comprehensive analysis of SARS-CoV-2 missense mutations indicates that all possible amino acid replacements in the viral proteins occurred within the first two-and-a-half years of the pandemic

Nicole Balasco [a,*,1], Gianluca Damaggio [b,c,d,1], Luciana Esposito [e], Vincenza Colonna [b,f], Luigi Vitagliano [e,*]

[a] Institute of Molecular Biology and Pathology, CNR c/o Dep. Chemistry, Sapienza University of Rome, Rome, Italy
[b] Institute of Genetics and Biophysics, CNR, Naples, Italy
[c] Laboratory of Stem Cell Biology and Pharmacology of Neurodegenerative Diseases, Department of Biosciences, University of Milan, Milan, Italy
[d] University of Naples Federico II, Naples, Italy
[e] Institute of Biostructures and Bioimaging, CNR, Naples, Italy
[f] Department of Genetics, Genomics and Informatics, College of Medicine, University of Tennessee Health Science Center, Memphis, TN, United States

ARTICLE INFO

Keywords:
COVID-19
SARS-CoV-2 proteins
Amino acid mutations

ABSTRACT

The surveillance of COVID-19 pandemic has led to the determination of millions of genome sequences of the SARS-CoV-2 virus, with the accumulation of a wealth of information never collected before for an infectious disease. Exploring the information retrieved from the GISAID database reporting at that time >13 million genome sequences, we classified the 141,639 unique missense mutations detected in the first two-and-a-half years (up to October 2022) of the pandemic. Notably, our analysis indicates that 98.2 % of all possible conservative amino acid replacements occurred. Even non-conservative mutations were highly represented (73.9 %). For a significant number of residues (3 %), all possible replacements with the other nineteen amino acids have been observed. These observations strongly indicate that, in this time interval, the virus explored all possible alternatives in terms of missense mutations for all sites of its polypeptide chain and that those that are not observed severely affect SARS-CoV-2 integrity. The implications of the present findings go well beyond the structural biology of SARS-CoV-2 as the huge amount of information here collected and classified may be valuable for the elucidation of the sequence-structure-function relationships in proteins.

1. Introduction

A deep understanding of the mechanisms that viruses employ to optimally adapt to the host organism represents a formidable challenge [1]. Indeed, in contrast to cellular organisms that share a common mechanism to express genomic information and an ensemble of genes that are present in all domains of life [2,3], viruses lack common genes and exploit a vast number of pathways for information transmission [4]. As recently outlined by Rochman et al. [1], the analysis of virus phylodynamics during the epidemics, when the size of the viral population dramatically increases, provides the opportunity to gain important insights into the molecular mechanisms that underlie the adaptive evolution.

In this general scenario, the still-ongoing COVID-19 pandemic, which has been followed by an unprecedented level of attention, has provided a wealth of information that, effectively exploited, is largely increasing our understanding of this intricate biological process. Although the SARS-CoV-2 virus as the pathogen responsible for the disease was rapidly identified (December 2019) and genomically characterized (mid-January 2020, GISAID accession ID: EPI_ISL_402124), all attempts to avoid its transmission failed and, starting approximately from March 2020 [5–8], a worldwide pandemic occurred with several hundred millions of registered infections (https://www.worldometers.info/coronavirus/) which represent a gross underestimation of the actually infected people.

Not surprisingly, the deep medical surveillance of the pandemic has

[*] Corresponding authors.
E-mail addresses: nicole.balasco@cnr.it (N. Balasco), luigi.vitagliano@cnr.it (L. Vitagliano).
[1] These authors contributed equally.

https://doi.org/10.1016/j.ijbiomac.2024.131054
Received 25 January 2024; Received in revised form 27 February 2024; Accepted 8 March 2024
Available online 24 March 2024

been accompanied by extensive characterizations at the genomic and molecular level of the SARS-CoV-2 virus and its progressively emerging variants [9–23]. SARS-CoV-2 belongs to the Baltimore class positive-sense single-stranded RNA virus and is a member of the subgenus Sarbecovirus. Its RNA sequence contains approximately 30,000 bases and encodes for 24 main distinct proteins. Since the publication of the first SARS-CoV-2 genome, the genetic evolution of the virus has been deeply monitored worldwide. This has led to the determination of millions of genome sequences of the virus with an accumulation of a wealth of information never collected before for infectious diseases, thus providing the unique opportunity to monitor the evolution of the mutations during the virus adaptation to humans. As typically observed for RNA viruses [24], SARS-CoV-2 presents a remarkable propensity to mutate with an estimated mutation rate of about 9.8×10^{-4} substitutions *per* site *per* year [13]. Over the years, many studies have been focused on the analyses of SARS-CoV-2 amino acid mutations, with particular attention to specific proteins directly involved in the disease transmission that are potential targets for preventive or therapeutic interventions [9–23]. In this scenario, we have performed a comprehensive analysis of all single point mutations detected in the first year of the pandemic (up to February 2021) [25]. The analysis of the observed/non-observed mutations in the early stages of the pandemic provided insights into the progressive adaptation of the virus to the host that was characterized by an accumulation of hydrophobic residues. Here this approach has been extended to the mutations detected in the first two-and-a-half years (up to October 2022). This novel investigation highlights the emergence of a completely different scenario compared to the previous one. Indeed, the data here presented strongly indicate that, in this novel time interval, the virus explored all possible alternatives in terms of missense mutations for all sites of its polypeptide chains and that those that are not observed likely affect SARS-CoV-2 integrity. The implications of the present findings go well beyond the structural biology of SARS-CoV-2 as the huge amount of information here collected and classified may be valuable for the elucidation of the sequence-structure-function relationships in proteins.

2. Materials and methods

The lists of the amino acid (AA) mutations present in the proteins of the SARS-CoV-2 variants, compared to the sequence of the Wuhan genome (GISAID accession ID: EPI_ISL_402124) used as a reference, were downloaded from the Global Initiative for Sharing All Influenza Data (GISAID) database (https://www.gisaid.org/) [26,27] at four-time points: July 2021 (DataJul21), February 2022 (DataFeb22), May 2022 (DataMay22), and October 2022 (DataOct22). Accordingly, the term WT (wild-type) proteins used in the text refers to the AA sequences derived from the Wuhan genome. Mutations have been collected for the 24 main proteins of the virus: 15 non-structural proteins (NSP1, NSP2, NSP3, NSP4, NSP5, NSP6, NSP7, NSP8, NSP9, NSP10, NSP12, NSP13, NSP14, NSP15, and NSP16), 4 structural proteins (E, M, N, and Spike), and 5 accessory factors (NS3, NS6, NS7a, NS7b, and NS8) [11]. In particular, since the aim of this study is the identification of missense mutations that can be tolerated by the virus and do not compromise its vitality, the frequencies of these amino acid substitutions were not considered. These mutations were manually curated to eliminate insertions/deletions and non-missense mutations. They were then classified following the criteria recently adopted by Balasco et al. [25]. Indeed, AA substitutions were classified as conservative (C) and non-conservative (NC) based on the probability of being accepted during the evolution according to the Point Accepted Mutation substitution score (PAM1 matrix) [28] and on the similarity of the encoding codons [29], i.e. considering if the AA substitution required one or more than one base change in the genetic code to occur. Non-conservative and conservative mutations have PAM values falling in the range 0–12 and >12, respectively [25]. The global list of the mutations detected in the last step of this study (DataOct2022) is reported in Table S1, which also includes the frequency of each mutation considering all genomes.

The secondary structure of proteins has been assigned using the DSSP program [30], which is based on the analysis of backbone dihedral angles and hydrogen bonds. DSSP assigns seven elements, i.e., H: α-helix, G: 3(10) helix, I: p-helix, E: extended strand, B: residue in isolated β-bridge, S: bend, and T: H-bonded turn. Based on these definitions, we assigned residues with either S (structured - H, E, G, and I definition by DSSP) or L (non-repetitive local structure - B, S, T, and blank). Finally, missing residues in the experimental structures were assigned as disordered (D). According to this classification, the number of D, L, and S residues is 939, 4040, and 4727, respectively.

Statistical analyses were performed using R (R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/). Figures of structural models were generated using the PyMOL molecular visualization program. Plots were generated using Xmgrace (https://plasma-gate.weizmann.ac.il/Grace/).

3. Results

3.1. Global trends: evolution of the number of conservative and non-conservative mutations

The inspection of the AA point mutations detected in the SARS-CoV-2 proteins clearly indicates that, although a rapid increase is evident in the first stage of the pandemic, a sort of plateau is reached after two-and-a-half years (October 2022) (Table 1). Indeed, from a total of approximately 39,000 mutations identified after one year (February 2021) [25], a threefold increase (124,680) has been observed after another year (February 2022). On the other hand, the pace of this increase has strongly reduced in the following months (141,639 mutations detected up to October 2022) (Fig. 1a). Indeed, an increase of only 3219 mutations, which represent 2 % of the total, has been observed in the 2022 July-Oct period. Although these trends are expected as the virus explores the effect of new mutations over time, the observation that a sort of plateau has been reached in the time interval here considered may indicate that the mutations not yet observed may be highly deleterious for the structure and/or the function of SARS-CoV-2 proteins. This consideration is corroborated by the observations that emerged from the splitting of the mutations in conservative and non-conservative ones (See Methods and [25]). In this scheme, of the total 380 (19 alternatives for each of the 20 amino acids) possible mutations, 43 are conservative (C, PAM value > 12) while 337 are non-conservative (NC, PAM values 0–12) (Table S2). By assuming that all possible AA replacements occur in SARS-CoV-2 proteins, a total of 184,414 mutations (19 replacements for each of the 9706 residues present in the virus proteins) are expected. Of these, 22,246 and 162,168 are C and NC substitutions, respectively (Table S2). As shown in Fig. 1a, the time evolution of the mutations indicates that a plateau is reached for both C and NC ones. Interestingly, in the database of the mutations found up to October 2022 (Tables S1 and S3), we detect 21,843 conservative mutations which correspond to 98.2 % of the expected ones (Table 1 and Fig. 1b). On the other hand, we detected 119,796 non-conservative mutations (73.9 % of the theoretical ones). In both cases, marginal increases are observed in the last months of our survey. Collectively, these observations suggest that in the two-and-a-half years of the pandemic, the virus has likely explored all possible alternatives for each residue present in its proteins. The data also suggest that the mutations that have not been observed probably negatively affect SARS-CoV-2 integrity. This consideration is corroborated by the observation that there is a significant correlation between the frequencies of the observed NC substitutions with the PAM [28] and BLOSUM62 [31,32] values associated with those changes (Fig. S1). Indeed, on average, remarkably lower detections are observed for mutations with very low PAM (range from 0 to 3) or BLOSUM62 (range from −4 to −2) values.

3.2. Global trends: number of residue substitutions per site

The analysis of the number of observed substitutions for each of the 9706 residues of the virus proteins (Fig. 2a) highlights that most of the sites present a remarkable tendency to mutate. Indeed, the average value of the number of substitutions *per* site is as high as 14.6 with a median value of 15. Moreover, the vast majority of residues (7966; 82.1 %) present >12 substitutions out of the 19 possible ones (variable sites). A low but significant fraction (267; 2.8 %) of residues present all possible 19 changes. Somehow surprisingly, when all these mutation sites are evaluated as a function of the local structure (see Methods for the definition), no clear differences are detected (Fig. 2b). Indeed, the fraction of disordered (D) residues (2.8 %) that exhibit all possible substitutions is similar to those observed for residues located in secondary structure (S) elements (2.6 %) or loop (L) regions (2.9 %). This analogy is also observed when data are globally analyzed since the average number of substitutions for D, L, and S residues is 14.62, 14.63, and 14.56, respectively. This finding suggests that SARS-CoV-2 proteins present a remarkable tolerance to mutations even for well-structured regions.

The inspection of Fig. 2a also indicates that a rather low number of sites (302; 3.1 %) present <10 replacements (conservative sites). Moreover, no site of the entire ensemble of the virus proteins is fully conserved across genomes collected up to October 2022. Notably, all 36 sites that present less than five replacements (hyper-conservative sites) are located at the very termini of the polypeptide chains of the different proteins. In general, as a consequence of the selectivity of the proteases that cleave the virus polypeptide chains, residues located at both the N- and the C-terminus present a limited variability (Fig. 2c and Tables S3–S4). The low variability of these regions has a small but significant impact on the residue mutation rate as a function of the structure/disorder. Indeed, as terminal residues are generally disordered, they tend to lower the substitution level of this class of residues. As a consequence, if the first and last 10 residues of all proteins are omitted from the analysis, a difference in the average number of substitutions for D (15.5), S (14.7), and L (14.9) is observed. This observation indicates that when terminal residues are not considered, in the general framework of a high variability of all amino acids, those belonging to disordered regions present a higher substitution degree.

The analysis of the conserved sites (number of replacements between 5 and 9) still reveals the presence of a remarkable number of amino acid residues that are close to the protein termini. Indeed, out of 266 sites of this class, 153 are located within ten residues from either the N- or the C-terminus. Interestingly, the remaining 113 conservative sites show a prevalence of residues located in structured regions. Indeed, no residues belong to disordered regions, while 73 and 40 are located in secondary structure elements and loop regions, respectively. Significant amounts of them (52) are embodied in helical structures. The analysis of the chemico-physical properties of conserved sites located in helices clearly indicates a prevalence of hydrophobic residues being Leu (18 cases), Phe (11), Val (5), and Pro (5) the most represented ones. This observation suggests that residues embedded in the hydrophobic patches of helices are the most structurally conserved sites. Illustrative examples of this situation are reported in Fig. S2 that shows the clustering of conservative residues in the N-terminal portion of the pentameric membrane channel formed by protein E [33] and at the dimerization interface of the non-structural protein NSP7 [34].

3.3. Global trends: observed and non-observed mutations

A global-scale analysis of the AA substitutions detected in SARS-CoV-2 variants highlights some general trends, despite the heterogeneity of the local environment of the mutation sites. Indeed, the inspection of the replacement matrix reported in Fig. 3 and Table S5, which reports the frequencies of all 380 possible AA substitutions, clearly evidences substitutions endowed with diversified frequencies. This can be rationalized based on the steric and chemico-physical properties of the involved amino acid residues. Most rare replacements are those that involve pairs of residues that are characterized by side chains with different sizes and polarity. In line with this observation, the replacement of aspartic acid and asparagine with tryptophan presents the lowest frequencies. Indeed, D > W and N > W substitutions are observed only in 27.1 and 26.0 %, respectively, of the possible cases (Fig. 3 and Table S5). Most of the other replacements with frequencies lower than 0.4 also involve tryptophan residues (A > W, E > W, H > W, I > W, K > W, P > W, S > W, T > W, V > W, and W > D). Among the other possible replacements, only the substitutions C > E and L > D fall below this threshold.

It is important to note that the replacement matrix is highly asymmetric since the frequency of a generic AA1 > AA2 substitution can be very different from the reverse one (AA2 > AA1). This is particularly evident for the Trp residue that is frequently replaced with several other AAs (horizontal line corresponding to W in Fig. 3) but cannot efficiently replace other residues (vertical column corresponding to W). The reverse is observed for some residues, such as serine, threonine, and valine, which are endowed with small side chains whose replacement is rather difficult although they can easily substitute other AAs. Collectively, these observations indicate that the size of the residues plays a major role in dictating the perturbation of these proteins caused by the mutations. Indeed, small residues may efficiently replace larger ones while the reverse has more negative effects. The role of the size and the polarity in the mutation process is also evident when the frequencies of replacing or including an AA are collectively analyzed. As shown in Fig. S3, some hydrophobic residues (F, I, L, M, and V) and proline are those replaced with the lowest frequencies. This observation indicates the importance of their role in preserving the hydrophobic core of some of the viral proteins. On the other hand, in addition to proline, negatively charged residues (D and E) are those that present the lowest propensity to replace other AAs.

3.4. Mutations in individual SARS-CoV-2 proteins

The 24 SARS-CoV-2 proteins here investigated present a wide range of structures and functions. To evaluate the frequencies and the impact of the mutations detected in the natural variants of the virus, the global

Table 1
Data about the sequences and AA mutations *per* dataset.

Dataset	Sequences	Number of AA mutations			Percentage of observed AA mutations (%)		
		Total	NC	C	Total	NC	C
Mar20[a]	581	404	231	173	0.22	0.14	0.78
Oct20[a]	135,404	25,634	16,830	8804	13.9	10.4	39.6
Feb21[a]	415,516	38,986	25,919	13,067	21.1	16.0	58.7
Jul21	2,216,094	67,381	48,779	18,602	36.5	30.1	83.6
Feb22	8,179,987	124,680	103,148	21,532	67.6	63.6	96.8
Jul22	11,449,212	138,420	116,625	21,795	75.1	71.9	98.0
Oct22	13,141,952	141,639	119,796	21,843	76.8	73.9	98.2

[a] Data from reference [25].

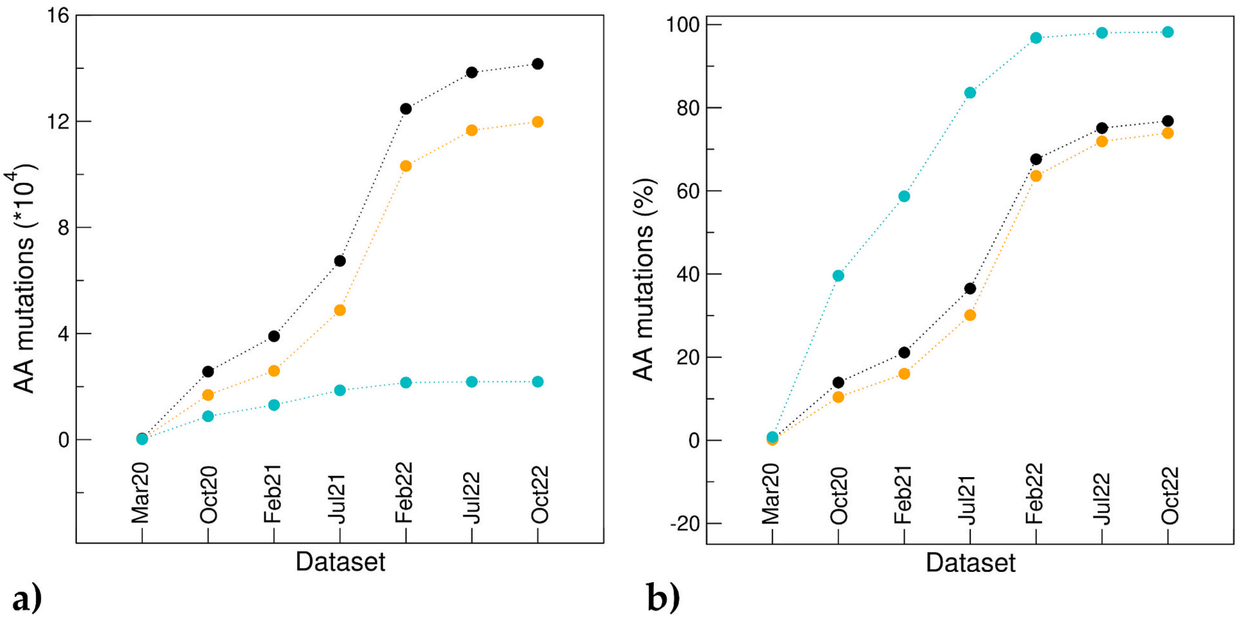

Fig. 1. (a) Number of AA mutations *per* database and (b) percentage of observed mutations compared to all possible ones (19 replacements for each of the 9706 residues present in the virus proteins). Data are shown for all (black), NC (orange), and C (cyan) substitutions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

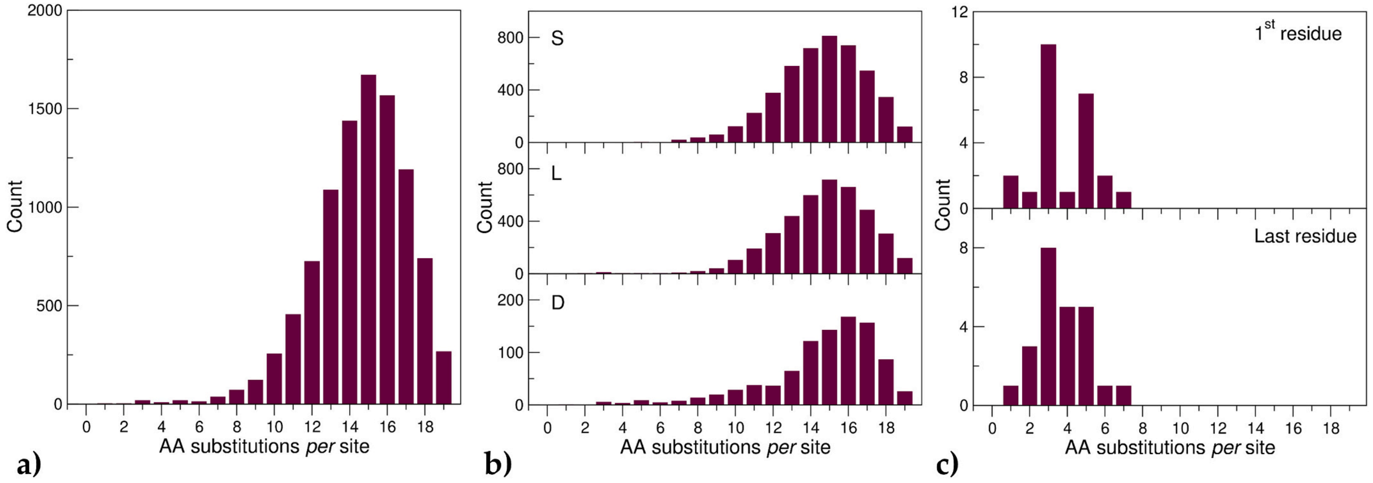

Fig. 2. Distribution of the number of AA substitutions *per* site for (a) all residues of SARS-CoV-2 proteins, (b) residues in secondary structure (S), loop (L) and disordered (D) regions, and (c) the first and last residue of all virus proteins.

ensemble of the substitutions was dissected and analyzed for each individual SARS-CoV-2 protein. As shown in Table 2, a remarkable variability of the percentage of the detected substitutions compared to the possible theoretical ones is observed. Indeed, this value ranges from 54.6 (NSP9) to 84.6 % (Spike). Although differences emerge also for conservative mutations, which are larger than 92 % for all proteins, more striking variations are observed for NC substitutions.

A global view of the number of mutations *per* residue for each SARS-CoV-2 protein is reported in Fig. 4. This picture clearly illustrates the different substitution frequencies of these proteins and their specific regions. The prevalence of the red colour is indicative of a high local variability. SARS-CoV-2 proteins are generally distinct according to their functional role(s) in three different classes: structural, non-structural, and accessory proteins. As highlighted in Table 2, except for the small structural envelope protein E, proteins presenting the lowest degree of substitution are non-structural ones (NSP9, NSP7, and NSP10). Low substitution levels are generally displayed by other non-structural proteins. Among structural proteins, the most conservative ones are proteins E and M, which present significant transmembrane regions. The nucleocapsid protein N and, especially, Spike present remarkable sequence variabilities. Except for NS6 and NS7b, accessory proteins typically present large substitution levels.

These general trends are corroborated by the analysis of protein stretches endowed with the lowest substitution levels. If terminal regions, which are well conserved (see above), are not considered, the occurrence of conservative residues (number of substitutions < 10) that are consecutive in the sequence is a quite rare event. In particular, the occurrence of three consecutive conservative residues is detected only for NSP9 (residues 87–90) and E (residues 11–14 and 18–20), the lowest mutated proteins, and for NSP5 (residues 286–288).

3.5. *Proteins of specific interest*

As delineated above, the SARS-CoV-2 protein that exhibits the lowest

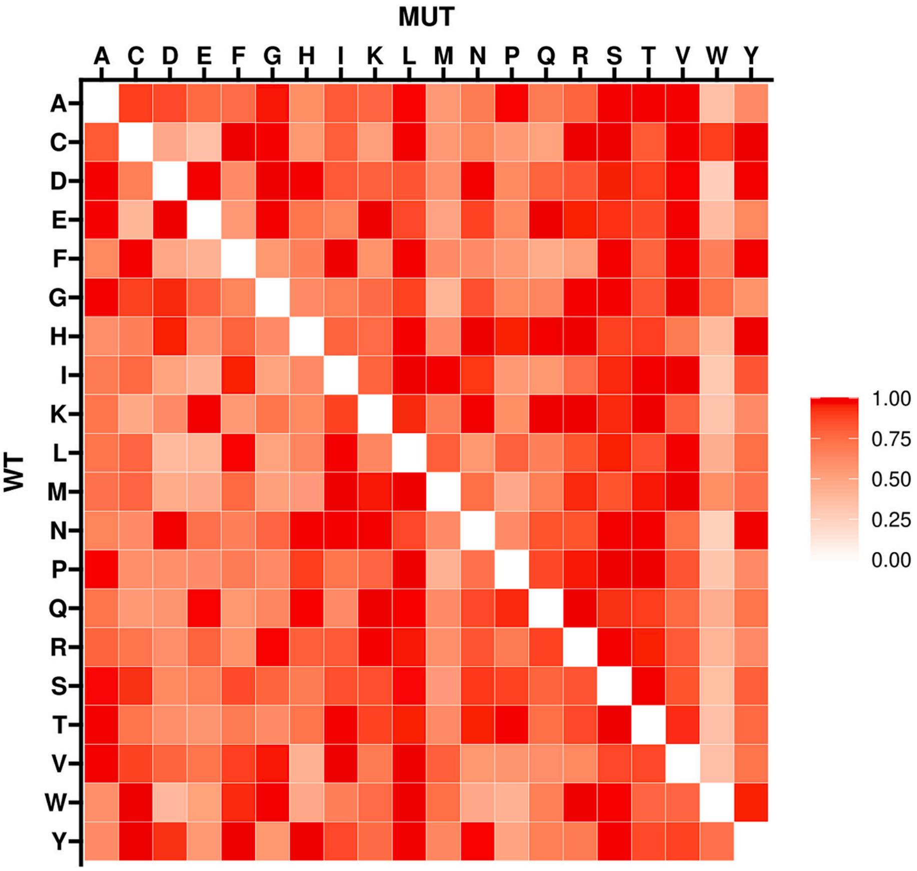

Fig. 3. Global heat map showing the substitution frequency of all pairs of residues. The white colour indicates that the replacement is never observed (frequency = 0) while the red colour indicates substitutions that are always observed (frequency = 1). The numerical values for the frequencies are reported in Table S5. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2

Data about the AA mutations for each individual SARS-CoV-2 protein.

Protein	Length (AA)	Number of AA mutations			Percentage of observed mutations (%)			Average number of AA substitutions *per* site
		Total	NC	C	Total	NC	C	
NSP9	113	1173	936	237	54.6	49.3	95.2	10.38
E	75	860	702	158	60.4	55.8	94.0	11.47
NSP7	83	963	786	177	61.1	56.6	94.1	11.60
NSP10	139	1663	1360	303	63.0	58.5	95.9	11.96
NS6	61	744	613	131	64.2	59.9	97.0	12.15
NSP8	198	2529	2031	498	67.2	62.7	95.2	12.77
NS7b	43	558	488	70	68.3	65.9	92.1	12.98
NSP12	932	12,848	10,788	2060	72.6	69.2	97.3	13.79
NSP13	601	8330	6992	1338	72.9	69.6	97.5	13.86
NSP16	298	4170	3497	673	73.6	70.4	96.8	13.99
NSP1	180	2530	2137	393	74.0	70.8	98.3	14.06
NSP14	527	7424	6267	1157	74.1	70.9	98.1	14.09
NSP15	346	4908	4121	787	74.7	71.4	98.0	14.18
NSP6	290	4122	3518	604	74.8	71.8	98.7	14.21
NSP5	306	4383	3700	683	75.4	72.3	98.4	14.32
M	222	3196	2732	464	75.8	72.9	98.3	14.40
NSP2	638	9309	7865	1444	76.8	73.7	99.1	14.59
NS3	275	4042	3461	581	77.4	74.6	98.8	14.70
N	419	6222	5226	996	78.2	75.2	98.4	14.85
NSP4	500	7475	6365	1110	78.7	76.0	98.8	14.95
NSP3	1945	30,172	25,657	4515	81.6	79.2	98.8	15.51
NS8	121	1892	1642	250	82.3	80.3	98.8	15.64
NS7a	121	1914	1662	252	83.3	81.3	99.2	15.82
Spike	1273	20,212	17,250	2962	83.6	81.4	99.1	15.88

substitution degree *per* site is NSP9. This is a dimeric protein that plays an important role in the overall virulence of the pathogen being implicated in viral replication, and genomic RNA reproduction [35]. The analysis of the number of substitutions *per* residue for NSP9 indicates that regions displaying the lowest values are involved in important structural/functional roles. Indeed, in addition to the N- and C-terminal

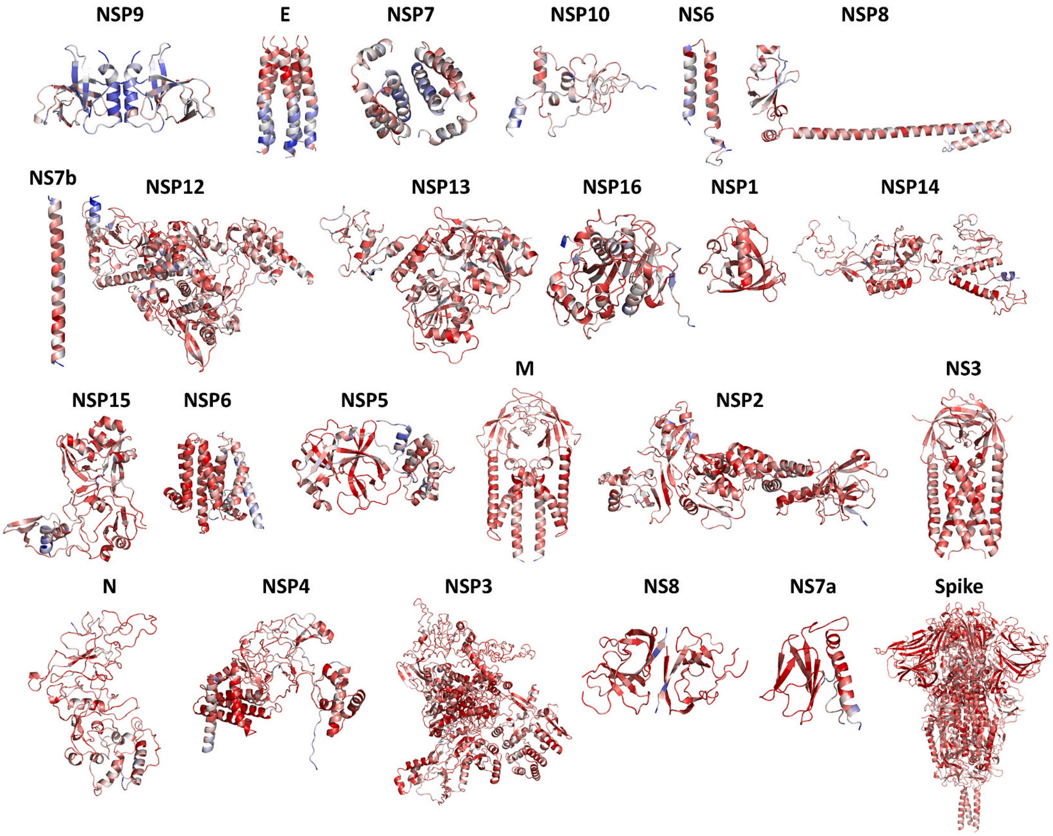

Fig. 4. Cartoon representation of SARS-CoV-2 proteins (from the least to the most mutated ones). Structures are colored (from blue to red) as a function of the number of substitutions *per* residue (from 1 to 19) detected in the DataOct22 dataset (Table S3). NSP9, E, NSP7, M, NS3, NS8 and Spike are reported in their functional oligomeric state. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

residues that, as in most SARS-CoV-2 proteins, are generally well-preserved, low substitution degrees are presented by residues belonging to the central β-structure of the protein core and the helix that is involved in the protein dimerization, which is known to be required for NSP9 functionality (Fig. 5). To achieve a better understanding of the specific observed/non-observed replacements we developed a substitution matrix by using the data reported in Table S3. In this type of matrices, we indicated with black and white boxes the substitutions that were observed or non-observed, respectively, in our dataset of mutations (DataOct22). The inspection of Fig. 5 clearly indicates that, for example,

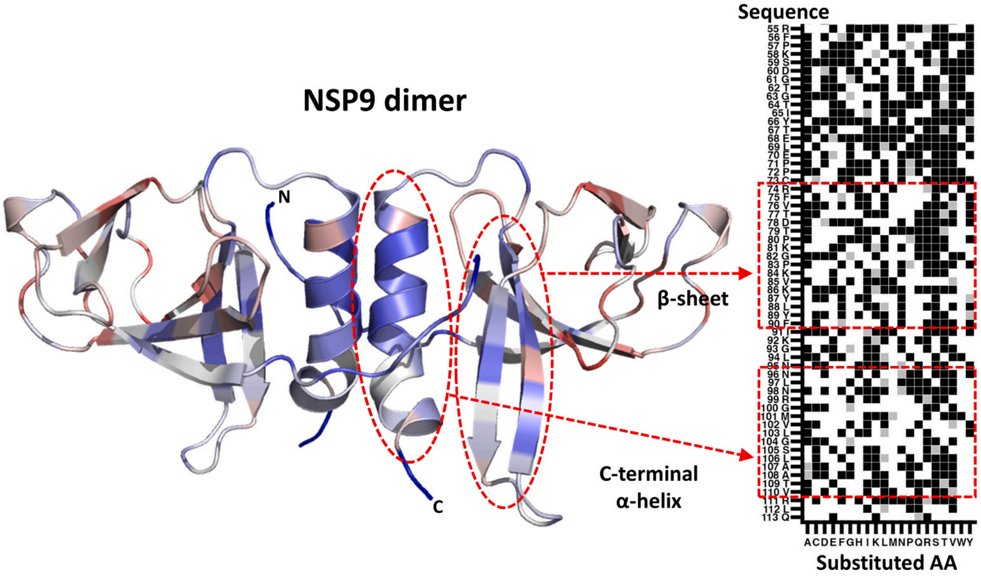

Fig. 5. Cartoon representation of the dimeric non-structural protein NSP9 (PDB ID: 6W9Q). Regions endowed with the lowest number of substitutions are pointed out on the structure (blue regions) and the substitution matrix (red boxes). Boxes in the substitution matrix are colored in white (no substitution observed) or black (substitution observed). The grey box would represent the amino acid replacement with itself. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Trp residue is rarely present in the conserved regions of the β-structure core and of the interface helices while it frequently replaces amino acids in other regions of the protein (Table S3). Similarly, Pro residues are remarkably under-represented in the β-structure, in line with the destabilizing role of Pro for this structural motif. The substitution matrix highlights similar, structurally related, impacts for other residues.

The second least mutated protein of the virus is the envelope protein E. This is an integral membrane protein, the smallest among structural ones, that plays an important role in the virus particle production and maturation. From the structural point of view, individual chains of the E protein self-associate to form a pentameric ion-selective channel (Fig. 6). The inspection of the substitution degree of this protein highlights the occurrence of conservative and variable regions corresponding to the N- (residues 8–20) and the C-terminal (residues 24–39) portions of the channel, respectively (Fig. 6). In particular, there are some residues, especially those with peculiar properties such as Trp, Pro, and Cys, that are specifically rare in the N-terminal region.

In addition to the protein E, the SARS-CoV-2 genome encodes for another predominantly membrane protein i.e. M. This dimeric protein is involved in the virus assembly by interacting with the other structural proteins E, N, and Spike [36]. Despite its specific localization and important structural role, the M protein presents a remarkable variability (Fig. 7). Indeed, on average the residue of this protein presents 14.4 replacements *per* site. Its behavior, in terms of the observed mutations, is much closer to the C- rather than the N-terminal region of the channel formed by the E protein. Indeed, if the terminal residues are not considered, the most conserved sites are Gly192 and Phe193 that present 10 replacements. A relatively conserved region corresponds to the central portion of the N-terminal transmembrane helix (Fig. 7).

The most variable protein of the entire SARS-CoV-2 proteome is the surface Spike glycoprotein. This homotrimeric membrane-anchored protein is responsible for the virus entry [37]. The attachment of the virus to the angiotensin-converting enzyme 2 (ACE2) receptor of the host is mediated by the Spike receptor-binding domain (RBD – residues 319–541) (Fig. 8a). The inspection of the database of mutations here analyzed reveals that, on average, its 1273 residues present nearly 16 out of the 19 possible substitutions (Table S3 and Fig. 8b). Notably, about 10 % (121) of the total residues of the protein exhibit all possible 19 substitutions. Excluding the terminal residues, a single out of nearly 1250 residues presents only 9 replacements (Phe1256) whereas eight additional ones (Phe338, Ala609, Pro1053, Phe1089, Ile1114, Trp1217, Phe1220, and Gly1251) presents 10 replacements. The high variability of the Spike protein is detected also for functionally important regions of the protein. Indeed, the Asn residues of the 23 N-glycosylation sites [38] present a high degree of replacement (Table S6) with an average replacement rate of 15.8 in line with the value observed for the entire protein. Similarly, the transmembrane region (residues 1214–1234) presents an average replacement degree of 15. The RBD region that is deputed to the host receptor binding presents an even higher level of substitutions (average value 16.2).

The Spike protein has been intensively investigated as its mutations have been frequently linked to the insurgence and the diffusion of genetic variant(s) of concern (VoC). Considering the global entropic score, emergence, spread, transmission, and their location in the RBD, Kumar et al. identified several residues of particular interest (Ala222, Asn439, Asn501, Leu452, Tyr453F, Glu484, Lys417, Thr478, Leu981F, Leu212, Asn856, Thr547, Gly496, Asp614 and Tyr369) [23]. As shown in Table S7 and Fig. 8c, these sites also present a remarkable degree of substitutions (range 15–19). This finding indicates that the mutations that emerged in the VoC were the result of a huge number of attempts made by the virus. Notably, of the residues of this subset located in the RBD, mutations that were never observed frequently involve the

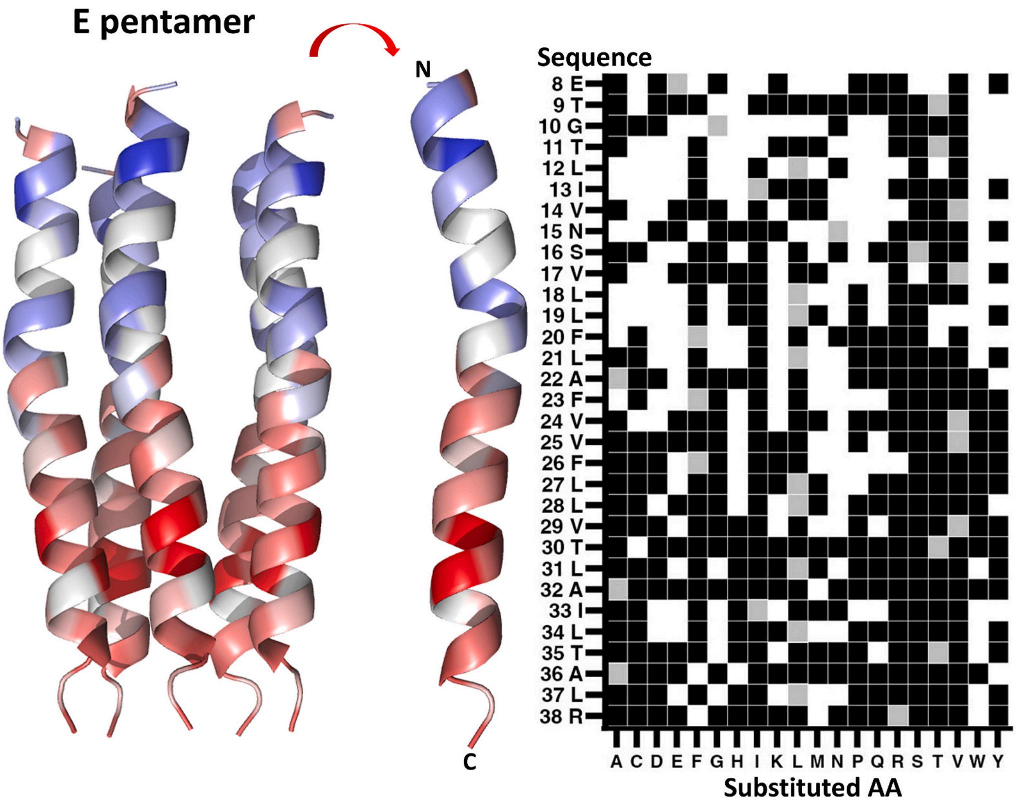

Fig. 6. Cartoon representation of the pentameric structural protein E (PDB ID: 8SUZ). The substitution matrix of the N-terminal transmembrane helix (residues 8–48) shows residues endowed with different degrees of substitution (from the least in blue to the most mutated in red in the structure). Boxes in the substitution matrix are colored in white (no substitution observed) or black (substitution observed). The grey box would represent the amino acid replacement with itself. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

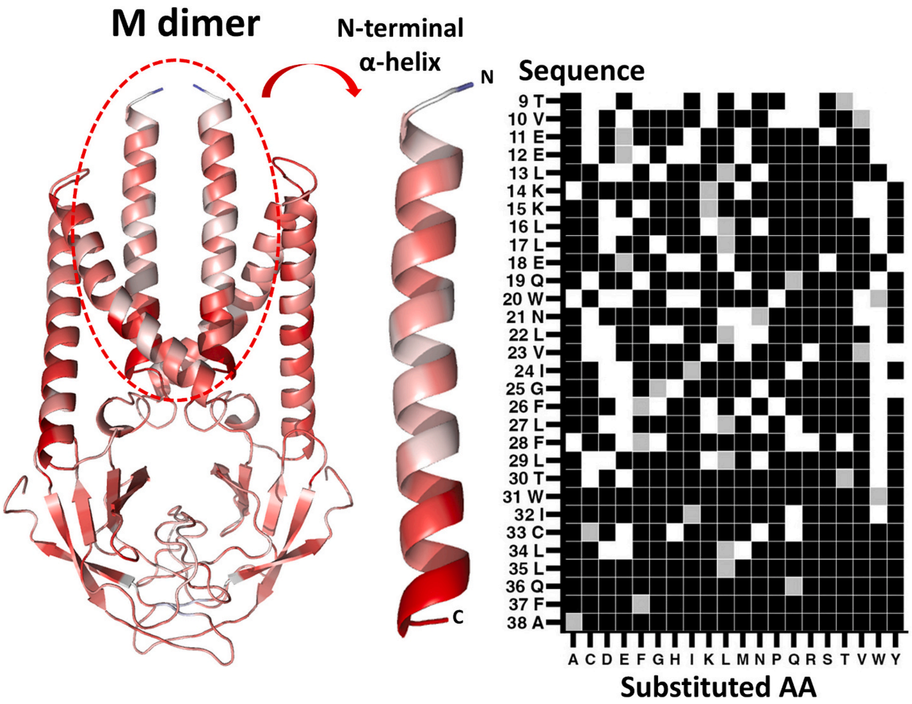

Fig. 7. Cartoon representation of the dimeric structural protein M (PDB ID: 7VGR). The substitution matrix of the N-terminal transmembrane helix (residues 9–38) showing a relatively conserved region in its central portion is shown. Boxes in the substitution matrix are colored in white (no substitution observed) or black (substitution observed). The grey box would represent the amino acid replacement with itself.

replacement of these residues with bulky and hydrophobic side chains such as Trp, Tyr, and Met.

4. Discussion

Since the outbreak of the COVID-19 pandemic, a huge amount of genomic data has been collected on the SARS-CoV-2 virus. This wealth of information constitutes a treasure for improving our knowledge of viral proteins' evolution. Here, we performed a systematic analysis of all single-point mutations, independently of their frequencies, detected in the worldwide campaign of SARS-CoV-2 genome sequencings that are compatible with the virus survival. Notably, the analysis of the ensemble of these amino acid replacements indicates that, after two-and-a-half years of the pandemic, a plateau in the total number of mutations was reached. Indeed, in the period July–October 2022, a minimal increase in new single-point mutations was observed. Furthermore, in October 2022 as many as 98.2 % of all possible conservative mutations were detected (Table 1). Even non-conservative mutations were highly represented in our database (73.9 %). These observations strongly indicate that, in this time interval, the virus explored all possible alternatives in terms of missense mutations for all sites of its polypeptide chain. As a consequence, it can be assumed that those that are not observed severely affect the structural integrity and/or the functionality of SARS-CoV-2 proteins. This consideration is corroborated by the observation that the likelihood of specific amino acid replacement here detected well correlates with validated substitution scales [28,31,32]. The detection of such a huge amount of single-point mutations demonstrates that variants currently present worldwide are the results of a very stringent selection process that has taken place during the pandemic. Indeed, randomly selected sequences of the currently most diffuse (December 2023) SARS-CoV-2 variants (GRA lineages JN.1 and HV.1) contains only 84 (GISAID accession ID: EPI_ISL_18717117) and 75 (GISAID accession ID: EPI_ISL_18717601) single point amino acid mutations compared to the first sequenced genome of the virus (Wuhan genome, GISAID accession ID: EPI_ISL_402124) [26,27].

Assuming that genomic data have been retrieved from vital variants of the virus, the analysis of the 141,639 unique single-point mutations detected up to October 2022 demonstrates that the SARS-CoV-2 proteins are extremely tolerant to single-point mutations. Indeed, a significant percentage of residues of the virus proteins (nearly 3 %) are substituted with all other 19 amino acid alternatives. Furthermore, the vast majority of the sites (>80 %) present >12 substitutions out of the 19 possible ones. On the other hand, a rather limited ensemble of sites presents less than ten substitutions. In general, we observe that the least frequent replacements involve the substitution of residues with small side chains with bulkier ones. In general, hydrophobic residues and those endowed with small side chains are the most tolerated in the replacements (Fig. S3). Notably, the most conservative sites do not correspond to residues playing key structural roles but they are concentrated in the terminal regions of the proteins where the cleavage of newly synthesized polypeptide chains occurs. Among sites playing structural roles, residues involved in interchain interactions display a relative conservation. Rather low substitution levels are indeed exhibited by residues located at the helix-helix interfaces. The relative conservation of these motifs suggests that they represent a delicate aspect of the functionality of the virus variants.

The implications of the present findings go well beyond the structural biology of SARS-CoV-2. Indeed, the elucidation of the sequence-structure-function relationships represents a main issue in protein studies. Although impressive results have been recently achieved in protein structure prediction and design [39–42], the impact that single-point mutations have on protein structure and functionality is difficult to be a priori fore-seen. This is due to the paucity of available data on systematic analyses of the effect that single-point mutations produce on protein stability. Traditional experimental approaches used to collect stability data have rather low throughput. Only recently [43], a domain-wide comprehensive mutagenesis analysis has been conducted on the small protein G (Gβ1), which contains 56 residues. In line with present data, a remarkable tolerance for hydrophobic residues was observed.

Although present data only provide substitution matrices reporting whether the substitution is observed or not, without providing any thermodynamic information, the amount of data collected, which is

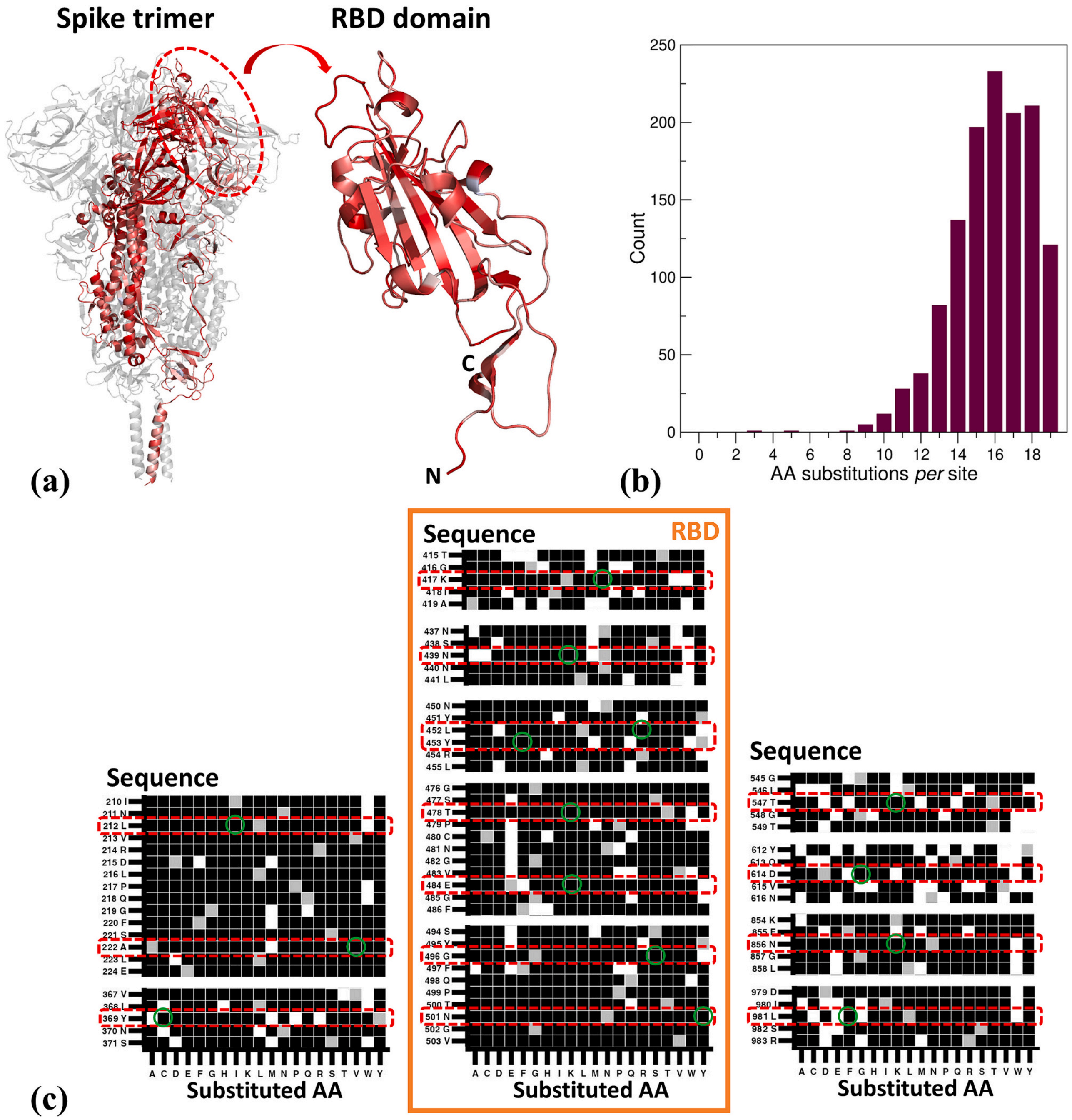

Fig. 8. (a) Cartoon representation of the trimeric structural protein Spike (PDB ID: 6XR8). (b) Distribution of the number of AA substitutions *per* site in Spike. (c) Substitution matrix of the protein regions containing residues of particular interest (pointed out with red boxes) as their mutation (L212I, A222V, Y369C, K417N, N439K, L452R, Y453F, T478K, E484K, G496S, N501Y, T547K, D614G, N856K, and L981F, highlighted with green circle) has been frequently linked to the insurgence and the diffusion of genetic VoC. Boxes in the substitution matrix are colored in white (no substitution observed) or black (substitution observed). The grey box would represent the amino acid replacement with itself. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

related to nearly 10,000 mutation sites detected in 24 proteins characterized by completely different structural features in terms of folds and localizations, may represent a valuable benchmark for algorithms aimed at predicting the effect of specific amino acid substitutions [44]. Although the presence of multiple mutations in the virus may progressively affect the local environment of each mutation site and therefore undermine the use of the structures derived from the Wuhan genome as a reference, it should be noted that virus variants generally present a limited number of mutations (<100, i.e. <1 % of the total residues).

In this context, there is a remarkable interest in predicting the effect (pathological/benign) of missense mutations in human proteins [45]. Although some success has been achieved [46,47], a significant improvement in the predicting algorithms is needed. By assuming that mutations that are not observed in the present dataset undermine the vitality of the virus and therefore resemble pathological human mutations, the data here reported may be fruitfully exploited to test and improve these algorithms.

In conclusion, the analysis of the time evolution of the single-point mutations of SARS-CoV-2 throughout strongly suggests that the virus has virtually explored all possible amino acid substitutions at a single residue level in the first two-and-a-half years of the pandemic. This consideration is corroborated by the observation that a plateau in the number of detected mutations is observed in the 2022 July-Oct trimester and by the finding that more the 98 % of the conservative mutations have been found in the genome sequencing in this period. Under this assumption, the single-point mutations that are not observed (42,775 out of 184,414 possible ones) likely impair the survival of the virus. In addition to the information strictly related to the SARS-CoV-2 virus, this study provides a remarkable collection of single-point replacements that are either tolerated or undermine protein structures that could be used to test the efficacy of algorithms devoted to the prediction of the effect of missense mutations. The present work also represents a solid ground for the analysis of the coevolution of mutation sites in individual SARS-CoV-2 genomes [48]. This topic, whose in-depth analysis is of paramount importance for predicting and understanding protein structures [39,49], can be efficiently investigated using this virus as a model system, also considering that almost all of its proteins have been extensively investigated from the structural point of view.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2024.131054.

CRediT authorship contribution statement

Nicole Balasco: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Gianluca Damaggio:** Writing – review & editing, Software, Formal analysis, Data curation. **Luciana Esposito:** Writing – review & editing, Methodology, Formal analysis. **Vincenza Colonna:** Writing – review & editing, Validation, Data curation. **Luigi Vitagliano:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We gratefully acknowledge the authors and laboratories that originated and submitted the sequences to GISAID. The authors thank Luca De Luca and Massimiliano Mazzucchi for technical assistance.

Funding

This study is financially supported by PNRR MUR – CN00000013 “National Centre for HPC, Big Data and Quantum Computing – Spoke 8”.

References

[1] N.D. Rochman, Y.I. Wolf, E.V. Koonin, Molecular adaptations during viral epidemics, EMBO Rep. 23 (2022) e55393, https://doi.org/10.15252/embr.202255393.

[2] R.L. Tatusov, The COG database: a tool for genome-scale analysis of protein functions and evolution, Nucleic Acids Res. 28 (2000) 33–36, https://doi.org/10.1093/nar/28.1.33.

[3] N.D. Rochman, Y.I. Wolf, E.V. Koonin, Deep phylogeny of cancer drivers and compensatory mutations, Commun. Biol. 3 (2020) 551, https://doi.org/10.1038/s42003-020-01276-7.

[4] J. Iranzo, M. Krupovic, E.V. Koonin, The double-stranded DNA virosphere as a modular hierarchical network of gene sharing, mBio 7 (2016) e00978-16, https://doi.org/10.1128/mBio.00978-16.

[5] G. Alicandro, G. Remuzzi, C. La Vecchia, Italy’s first wave of the COVID-19 pandemic has ended: no excess mortality in May, 2020, Lancet 396 (2020) e27–e28, https://doi.org/10.1016/S0140-6736(20)31865-1.

[6] N. Balasco, V. d’Alessandro, P. Ferrara, G. Smaldone, L. Vitagliano, Analysis of the time evolution of COVID-19 lethality during the first epidemic wave in Italy, Acta Biomed. Atenei Parm. 92 (2021) e2021171, https://doi.org/10.23750/abm.v92i2.11149.

[7] J.A. Ward, E.M. Stone, P. Mui, B. Resnick, Pandemic-related workplace violence and its impact on public health officials, March 2020–January 2021, Am. J. Public Health 112 (2022) 736–746, https://doi.org/10.2105/AJPH.2021.306649.

[8] V. d’Alessandro, N. Balasco, P. Ferrara, L. Vitagliano, The temporal correlation between positive testing and death in Italy: from the first phase to the later evolution of the COVID-19 pandemic: time evolution of COVID-19 weekly lethality rate, Acta Biomed. Atenei Parm. 92 (2022) e2021395, https://doi.org/10.23750/abm.v92i6.12030.

[9] A. Telenti, E.B. Hodcroft, D.L. Robertson, The evolution and biology of SARS-CoV-2 variants, Cold Spring Harb. Perspect. Med. 12 (2022) a041390, https://doi.org/10.1101/cshperspect.a041390.

[10] A.G. Wrobel, D.J. Benton, C. Roustan, A. Borg, S. Hussain, S.R. Martin, P. B. Rosenthal, J.J. Skehel, S.J. Gamblin, Evolution of the SARS-CoV-2 spike protein in the human host, Nat. Commun. 13 (2022) 1178, https://doi.org/10.1038/s41467-022-28768-w.

[11] J.H. Lubin, C. Zardecki, E.M. Dolan, C. Lu, Z. Shen, S. Dutta, J.D. Westbrook, B. P. Hudson, D.S. Goodsell, J.K. Williams, et al., Evolution of the SARS-CoV -2 proteome in three dimensions (3D) during the first 6 months of the COVID -19 pandemic, Proteins Struct. Funct. Bioinforma. 90 (2022) 1054–1080, https://doi.org/10.1002/prot.26250.

[12] R. Arya, P. Tripathi, K. Nayak, J. Ganesh, S.C. Bihani, B. Ghosh, V. Prashar, M. Kumar, Insights into the evolution of mutations in SARS-CoV-2 non-spike proteins, Microb. Pathog. 185 (2023) 106460, https://doi.org/10.1016/j.micpath.2023.106460.

[13] L. Van Dorp, D. Richard, C.C.S. Tan, L.P. Shaw, M. Acman, F. Balloux, No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2, Nat. Commun. 11 (2020) 5986, https://doi.org/10.1038/s41467-020-19818-2.

[14] M. Chiara, D.S. Horner, C. Gissi, G. Pesole, Comparative genomics suggests limited variability and similar evolutionary patterns between major clades of SARS-CoV-2, bioRxiv (2020) 016790, https://doi.org/10.1101/2020.03.30.016790.

[15] M.R. Islam, M.N. Hoque, M.S. Rahman, A.S.M.R.U. Alam, M. Akther, J.A. Puspo, S. Akter, M. Sultana, K.A. Crandall, M.A. Hossain, Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity, Sci. Rep. 10 (2020) 14004, https://doi.org/10.1038/s41598-020-70812-6.

[16] L.J. Klimczak, T.A. Randall, N. Saini, J.-L. Li, D.A. Gordenin, Similarity between mutation spectra in hypermutated genomes of rubella virus and in SARS-CoV-2 genomes accumulated during the COVID-19 pandemic, PLoS One 15 (2020) e0237689, https://doi.org/10.1371/journal.pone.0237689.

[17] I.J. Morais, R.C. Polveiro, G.M. Souza, D.I. Bortolin, F.T. Sassaki, A.T.M. Lima, The global population of SARS-CoV-2 is composed of six major subtypes, Sci. Rep. 10 (2020) 18289, https://doi.org/10.1038/s41598-020-74050-8.

[18] E. Trucchi, P. Gratton, F. Mafessoni, S. Motta, F. Cicconardi, G. Bertorelle, I. D’Annessa, D. Di Marino, Population Dynamics and Structural Effects at Short and Long Range Support the Hypothesis of the Selective Advantage of the G614 SARS-CoV-2 Spike Variant, Mol. Biol. Evol. 38 (2021) 1966–1979, https://doi.org/10.1093/molbev/msaa337.

[19] L. Van Dorp, M. Acman, D. Richard, L.P. Shaw, C.E. Ford, L. Ormond, C.J. Owen, J. Pang, C.C.S. Tan, F.A.T. Boshier, et al., Emergence of genomic diversity and recurrent mutations in SARS-CoV-2, Infect. Genet. Evol. 83 (2020) 104351, https://doi.org/10.1016/j.meegid.2020.104351.

[20] D. Giordano, L. De Masi, M.A. Argenio, A. Facchiano, Structural dissection of viral spike-protein binding of SARS-CoV-2 and SARS-CoV-1 to the human angiotensin-converting enzyme 2 (ACE2) as cellular receptor, Biomedicines 9 (2021) 1038, https://doi.org/10.3390/biomedicines9081038.

[21] N. D’Arminio, D. Giordano, B. Scafuri, A. Facchiano, A. Marabotti, Standardizing macromolecular structure files: further efforts are needed, Trends Biochem. Sci. 48 (2023) 590–596, https://doi.org/10.1016/j.tibs.2023.03.002.

[22] A.M. Carabelli, T.P. Peacock, L.G. Thorne, W.T. Harvey, J. Hughes, COVID-19 Genomics UK Consortium, T.I. De Silva, S.J. Peacock, W.S. Barclay, T.I. De Silva, et al., SARS-CoV-2 variant biology: immune escape, transmission and fitness, Nat. Rev. Microbiol. (2023), https://doi.org/10.1038/s41579-022-00841-7.

[23] R. Kumar, Y. Srivastava, P. Muthuramalingam, S.K. Singh, G. Verma, S. Tiwari, N. Tandel, S.K. Beura, A.R. Panigrahi, S. Maji, et al., Understanding mutations in human SARS-CoV-2 spike glycoprotein: a systematic review & meta-analysis, Viruses 15 (2023) 856, https://doi.org/10.3390/v15040856.

[24] R. Sanjuán, P. Domingo-Calap, Mechanisms of viral mutation, Cell. Mol. Life Sci. 73 (2016) 4433–4448, https://doi.org/10.1007/s00018-016-2299-6.

[25] N. Balasco, G. Damaggio, L. Esposito, F. Villani, R. Berisio, V. Colonna, L. Vitagliano, A global analysis of conservative and non-conservative mutations in SARS-CoV-2 detected in the first year of the COVID-19 world-wide diffusion, Sci. Rep. 11 (2021) 24495, https://doi.org/10.1038/s41598-021-04147-1.

[26] Y. Shu, J. McCauley, GISAID: global initiative on sharing all influenza data – from vision to reality, Eurosurveillance 22 (2017), https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

[27] S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID’s innovative contribution to global health, Global Chall. 1 (2017) 33–46, https://doi.org/10.1002/gch2.1018.

[28] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, A model of evolutionary change in proteins, in: Atlas of Protein Sequence and Structure; Dayhof, M.O. vol. 5, 1978, pp. 345–352.

[29] K.-F. Chan, S. Koukouravas, J.Y. Yeo, D.W.-S. Koh, S.K.-E. Gan, Probability of change in life: amino acid changes in single nucleotide substitutions, Biosystems 193–194 (2020) 104135, https://doi.org/10.1016/j.biosystems.2020.104135.

[30] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637, https://doi.org/10.1002/bip.360221211.

[31] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. 89 (1992) 10915–10919, https://doi.org/10.1073/pnas.89.22.10915.

[32] M.P. Styczynski, K.L. Jensen, I. Rigoutsos, G. Stephanopoulos, BLOSUM62 miscalculations improve search performance, Nat. Biotechnol. 26 (2008) 274–275, https://doi.org/10.1038/nbt0308-274.

[33] J. Medeiros-Silva, A.J. Dregni, N.H. Somberg, P. Duan, M. Hong, Atomic structure of the open SARS-CoV-2 E viroporin, Sci. Adv. 9 (2023) eadi9007, https://doi.org/10.1126/sciadv.adi9007.

[34] M. Biswal, S. Diggs, D. Xu, N. Khudaverdyan, J. Lu, J. Fang, G. Blaha, R. Hai, J. Song, Two conserved oligomer interfaces of NSP7 and NSP8 underpin the dynamic assembly of SARS-CoV-2 RdRP, Nucleic Acids Res. 49 (2021) 5956–5966, https://doi.org/10.1093/nar/gkab370.

[35] D.R. Littler, B.S. Gully, R.N. Colson, J. Rossjohn, Crystal structure of the SARS-CoV-2 non-structural protein 9, Nsp9, iScience *23* (2020) 101258, https://doi.org/10.1016/j.isci.2020.101258.

[36] Z. Zhang, N. Nomura, Y. Muramoto, T. Ekimoto, T. Uemura, K. Liu, M. Yui, N. Kono, J. Aoki, M. Ikeguchi, et al., Structure of SARS-CoV-2 membrane protein essential for virus assembly, Nat. Commun. 13 (2022) 4399, https://doi.org/10.1038/s41467-022-32019-3.

[37] Y. Cai, J. Zhang, T. Xiao, H. Peng, S.M. Sterling, R.M. Walsh, S. Rawson, S. Rits-Volloch, B. Chen, Distinct conformational states of SARS-CoV-2 spike protein, Science 369 (2020) 1586–1592, https://doi.org/10.1126/science.abd4251.

[38] Y. Gong, S. Qin, L. Dai, Z. Tian, The glycosylation in SARS-CoV-2 and its receptor ACE2, Signal Transduct. Target. Ther. 6 (2021) 396, https://doi.org/10.1038/s41392-021-00809-8.

[39] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021) 583–589, https://doi.org/10.1038/s41586-021-03819-2.

[40] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all, Nat. Methods 19 (2022) 679–682, https://doi.org/10.1038/s41592-022-01488-1.

[41] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W.R. Nelson, A. Bridgland, et al., Improved protein structure prediction using potentials from deep learning, Nature 577 (2020) 706–710, https://doi.org/10.1038/s41586-019-1923-7.

[42] P.-S. Huang, S.E. Boyken, D. Baker, The coming of age of de novo protein design, Nature 537 (2016) 320–327, https://doi.org/10.1038/nature19946.

[43] A. Nisthal, C.Y. Wang, M.L. Ary, S.L. Mayo, Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis, Proc. Natl. Acad. Sci. 116 (2019) 16367–16377, https://doi.org/10.1073/pnas.1903888116.

[44] F. Pucci, M. Schwersensky, M. Rooman, Artificial intelligence challenges for predicting the impact of mutations on protein stability, Curr. Opin. Struct. Biol. 72 (2022) 161–168, https://doi.org/10.1016/j.sbi.2021.11.001.

[45] J.A. Marsh, S.A. Teichmann, Predicting pathogenic protein variants, Science 381 (2023) 1284–1285, https://doi.org/10.1126/science.adj8672.

[46] J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė, T. Applebaum, A. Pritzel, L. H. Wong, M. Zielinski, T. Sargeant, et al., Accurate proteome-wide missense variant effect prediction with AlphaMissense, Science 381 (2023) eadg7492, https://doi.org/10.1126/science.adg7492.

[47] A. Ljungdahl, S. Kohani, N.F. Page, E.S. Wells, E.M. Wigdor, S. Dong, S.J. Sanders, AlphaMissense is better correlated with functional assays of missense impact than earlier prediction algorithms, bioRxiv (2023) 562294, https://doi.org/10.1101/2023.10.24.562294.

[48] A.D. Neverov, G. Fedonin, A. Popova, D. Bykova, G. Bazykin, Coordinated evolution at amino acid sites of SARS-CoV-2 spike, eLife 12 (2023) e82516, https://doi.org/10.7554/eLife.82516.

[49] A.G. Green, H. Elhabashy, K.P. Brock, R. Maddamsetti, O. Kohlbacher, D.S. Marks, Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences, Nat. Commun. 12 (2021) 1396, https://doi.org/10.1038/s41467-021-21636-z.