Gene 922 (2024) 148556

Contents lists available at ScienceDirect

Gene

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

Research paper

An entropy-based study on the mutational landscape of SARS-CoV-2 in USA: Comparing different variants and revealing co-mutational behavior of proteins

Daniele Santoni [*]

Institute for System Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Via dei Taurini 19, Rome 00185, Italy

ARTICLE INFO

Edited by: Sk. Sarif Hassan

Keywords:
SARS-CoV-2
Shannon Entropy
Hellinger distance
Co-mutations
Variants
Phylogenetic trees

ABSTRACT

COVID-19 emergency has pushed the international scientific community to use every resource to combat the spread of the virus, to understand its biology and predict its possible evolution in terms of new variants. Since the first SARS-CoV-2 virus nucleotide and amino acid sequences were made available, information theory was used to study how viral information content was changing over time and then trace the evolution of its mutational landscape. In this work we analyzed SARS-CoV-2 sequences collected mainly in the USA in a period from March 2020 until December 2022 and computed mutation profiles of viral proteins over time through an entropy-based approach using Shannon Entropy and Hellinger distance. This representation allows an at-a-glance view of the mutational landscape of viral proteins over time and can provide new insights on the evolution of the virus from different points of view. Non-structural proteins typically showed flat mutation profiles, characterized by a very low Average mutation Entropy, while accessory and structural proteins showed mostly non uniform and high mutation profiles, often coupled with the predominance of variants. Interestingly NSP2 protein, whose function is currently still debated, falls in the same branch of NSP14 and NSP10 in the phylogenetic tree of mutations constructed through correlations of mutation profiles, suggesting a co-evolution of those proteins and a possible functional link with each other. To the best of our knowledge this is the first study based on a massive amount of data (n = 107,939,973) that analyzes from an entropy point of view the mutational landscape of SARS-CoV-2 over time and depicts a mutational temporal profile of each protein of the virus.

1. Introduction

The COVID-19 pandemic has profoundly affected our lives in recent years in every aspect: socially, professionally and obviously from a health point of view. The emergency has pushed the international scientific community to use every resource to combat the spread of the virus, to understand its biology and predict its possible evolution in terms of new variants. Since the first SARS-CoV-2 virus nucleotide and amino acid sequences were made available, information theory was used to study how viral information content was changing over time and then trace the evolution of its mutational landscape. Sequence composition was studied through computational methodologies since late 90s. The first attempt to analyze sequence composition and statistical properties of biological sequences was performed by the pioneering works of Karlin in 1997 (Karlin et al., 1997; Karlin and Mrazek, 1997). In more recent years information theory has been applied in the context of molecular biology to study the information content of biological sequences (Chanda et al., 2020; Vinga, 2014; Vinga and Almeida, 2004; Adami, 2004). Shannon entropy has been shown to be of particular interest in evaluating mutation rate of single genomic or amino acid positions as well as of specific loci (Vopson and Robson, 2021; Gregori et al., 2016; Rhee et al., 2008). Mutual information has been used to study co-mutations (Pensar et al., 2019) and Kulback-Liebler like distances to study how mutational distributions are far from each other (Vergni et al., 2022).

Several studies in literature have focused on mutational landscape of SARS-CoV2 from an information theory point of view. Ghanchi and colleagues studied available SARS-CoV-2 sequences collected in Pakistan between March and October 2020 to investigate their genomic diversity and compare site-specific mutations and entropies among strains isolated before and after June 2020 (Ghanchi et al., 2021). Ashraf and colleagues followed up this study by investigating associations

[*] Corresponding author.
E-mail address: daniele.santoni@iasi.cnr.it.

https://doi.org/10.1016/j.gene.2024.148556
Received 21 February 2024; Received in revised form 8 May 2024; Accepted 9 May 2024
Available online 14 May 2024

between mutation rates and entropy in Pakistan comparing three time intervals 2020, 2021 and 2022 (Ashraf et al., 2023). Mullick and colleagues in 2021 deployed Shannon entropy to identify positions in the Spike proteins of SARS-CoV-2 that are most susceptible to mutations and built a model based on ProtBERT to predict potential mutation hotspots (Mullick et al., 2021). Namazi and colleagues in 2020 applied complexity and information theory to investigate the variations of SARS-CoV-2 genome showing that the fractal dimension and Shannon entropy of genome walk change significantly between different USA states (Namazi et al., 2020). In 2024 Formentin and colleagues proposed a study based on Shannon entropy of SARS-CoV-2 sequences as well as the relative entropy and the mutual information between the reference sequence and the mutated ones suggesting new optimal entropic properties of the mutation process (Formentin et al., 2024).

In 2022 an entropy-based study on mutational trajectories of SARS-CoV-2 protein sequences collected in India has been proposed (Santoni et al., 2022). In particular, in this paper the average entropy profiles over time were computed for all the viral proteins showing a clear different behaviour between structural and non-structural proteins. In the present work we followed the approach proposed in Santoni et al. (2022) analyzing a significantly much larger data set which covers a period from March 2020 to December 2022, mainly considering sequences collected in USA. The peculiarity of the proposed approach consists in collapsing all the information related to amino acid mutations over a given time interval in two values the Average Shannon Entropy and the Average Hellinger distance between two consecutive months providing a global view of the mutational evolution of viral proteins over time. There are several works focusing on the evolution of the virus studying single or set of mutations or focusing on measures of similarity among strains (Rogozin et al., 2024; Markov et al., 2023; Magazine et al., 2022) but they lack a quantitative global evaluation of mutational rate at a protein level over time that we performed through the application of information theory. To the best of our knowledge this is the first study based on a massive amount of data that analyzes from an entropy point of view the mutational landscape of SARS-CoV-2 over time (in terms of average Shannon entropy over all the positions) and depicts a mutational temporal profile of each protein of the virus. Previous works, such as those mentioned above, either focus on amino acid specific positions or analyze a relatively limited amount of data. These studies were usually performed on a limited interval time, when not so many data were available, and moreover some of them were limited to countries where the number of sequences was not that large. The current work was conducted on a dataset approximately more than a thousand time larger than the previous one (Santoni et al., 2022). Furthermore we extended the analysis by also studying the relationship between virus variants on one hand and the average mutation entropy and the Hellinger distance on the other hand. Finally we studied the correlations between mutational profiles of proteins hypothesizing that interacting proteins show similar (correlated) mutational profiles over time.

2. Materials and methods

2.1. Preparation of sequence dataset

Initially, all available protein sequences (n = 387,171,639) of SARS-CoV-2 spanning March 2020 to December 2022 were retrieved (January 2023) from GISAID (https://www.gisaid.org/) (Khare et al., 2021). These sequences were firstly divided into 27 datasets, each one corresponding to a certain viral protein and then were further subdivided depending on the country from which they were collected. The present study was focused mainly on sequences collected from USA as their number is largely the highest compared to the number of sequences of all other countries. We are confident that such a large data size can ensure statistical reliability and consistency of results. We resolved to focus on sequences collected in a single country because the timing and spreading of the virus are not geographically uniform and variants appear months later in some countries and months before in some others, so it would have been unfair and ineffective to analyze together sets of sequences collected in different regions. All sequences from USA were selected and parsed using *ad hoc* perl scripts to discard those containing non standard amino acid, including the symbol X (associated with the non sequenced amino acid), obtaining a total of 108,448,489 sequences.

Furthermore the length distribution was also analysed for each protein. All sequences whose length frequency was smaller than 0.001 were removed in order to exclude from the study unreliable and partial sequences. The final number of sequences for all the 27 considered proteins is 107,939,973. For each protein the sequences were aligned by using kalign3 (Lassmann, 2020) and then the multiple alignment files were parsed through *ad hoc* perl scripts to compute amino acid position frequencies. The same procedure was applied to Spike sequences collected in UK, Germany and Denmark.

In Table 1 the total number of sequences collected in USA for each protein is reported along with the minimum and maximum length, the average number of sequences per month and the lowest and highest number of sequences among all the considered months.

As can be observed in Table 1 the number of sequences considered in the study is truly remarkable (107,939,973) guaranteeing solidity and reliability to the analysis. Most proteins (20 out of 27) show a fixed length (trivially minimal and maximal lengths are the same). On average the number of sequences per month is higher than one hundred thousand (excluding Spike 80,598) ranging from a minimum of around seven thousands up to around four hundreds thousands.

2.2. Entropy-based approach

Let A be the set of symbols or alphabet made of the twenty canonical amino acids plus the symbol "-" indicating a deletion. Let $S = s_1, s_2, \ldots, s_m$ be a set of sequences (of a given protein), where m is the total number of

Table 1
Statistics on sequences collected in USA from March 2020 to December 2022 and considered in this study. For each protein the total number of sequences is reported along with the minimum and maximum length, the average number of sequences per month and the lowest and highest number of sequences among all the considered months.

Protein	Tot Seq	Min Len	Max Len	Av Seq month	Min Seq month	Max Seq month
E	4,241,188	75	75	124,741	11,806	403,414
M	3,748,816	222	222	110,259	10,889	297,124
N	3,937,933	416	419	115,822	10,285	389,842
NS3	4,071,404	275	275	119,747	10,415	399,476
NS6	4,311,133	61	61	126,798	11,226	448,812
NS7a	3,942,485	121	121	115,955	9,154	426,240
NS7b	4,034,747	43	43	118,669	9,271	406,812
NS8	3,807,377	105	121	111,982	11,294	392,354
NS9b	4,199,927	94	97	123,527	11,199	415,633
NS9c	4,073,636	73	73	119,813	7,618	444,827
NSP1	4,030,937	175	180	118,557	11,293	429,004
NSP2	3,891,272	638	638	114,449	9,831	407,652
NSP3	4,071,404	1,944	1,945	119,747	10,415	399,476
NSP4	3,820,546	500	500	112,369	10,294	389,246
NSP5	4,190,185	306	306	123,241	11,385	419,457
NSP6	4,007,023	287	290	117,854	10,683	398,616
NSP7	4,391,459	83	83	129,161	11,814	451,999
NSP8	4,272,114	198	198	125,650	11,202	445,346
NSP9	4,361,033	113	113	128,266	11,720	453,342
NSP10	4,258,397	139	139	125,247	10,917	445,078
NSP11	4,350,328	13	13	127,951	11,421	450,313
NSP12	3,873,878	932	932	113,938	10,253	380,608
NSP13	4,019,087	601	601	118,208	10,367	400,987
NSP14	3,559,628	527	527	104,695	7,853	380,200
NSP15	3,900,545	346	346	114,722	9,590	403,579
NSP16	3,833,149	298	298	112,740	9,391	427,078
Spike	2,740,342	1267	1273	80,598	8,661	189,850

sequences. Let X be a matrix of dimension $n \times m$ derived from a multi alignment (where n is the length of aligned sequences). Each element $x(i, j) \in X$ is an amino acid or a deletion occurring in the alignment at position i of j^{th} sequence (for $i = 1,2,..n$, and $j = 1,2,..m$). For every aligned position i, the distribution frequency $p_i(a)$ for $a \in A$ is defined as the ratio between the number of occurrences of a at position i and the total number of sequences m. The mutation Entropy (E) of a position i is computed as Shannon entropy of the frequency function $p_i(a)$:

$$E(i) = -\sum_{a \in A} p_i(a) \ \ log_2(p_i(a)) \tag{1}$$

The Average mutation Entropy (AE) is trivially the Mutation Entropy (E) averaged over all the positions i of the given multi alignment:

$$AE = \frac{\sum_{i=1..n} E(i)}{n} \tag{2}$$

The Average mutation Entropy was computed separately on sequence sets related to each month starting from March 2020 till December 2022. In order to evaluate how residues are differently distributed between different sample sets (different months) for a given position we used the Hellinger distance.

Given two sets of sequences A and B (in our case they will correspond to sequences associated with two different months) the relative frequencies of a given amino acid (or deletion "-") for a given position i in the multi alignment are defined as $p_i^A(a)$ and $p_i^B(a)$ for A and B respectively. It is worth noting that the multi alignment was performed on all sequences together (before separating them into months) so the length of each aligned sequence is the same. The Hellinger distance between sequence sets A and B related to position $i, H(i)_{A,B}$ is defined as follows:

$$H(i)_{A,B} = \frac{1}{\sqrt{(2)}}\sqrt{\sum_{a \in A}\left(\sqrt{p_i^A(a)} \ - \ \sqrt{p_i^B(a)}\right)^2} \tag{3}$$

Similar to the average entropy, the Average Hellinger distance between the two sets A and B over all the positions of the given protein is defined as:

$$AH_{A,B} = \frac{\sum_{i=1..n} H(i)_{A,B}}{n} \tag{4}$$

Both Average Shannon Entropy and Average Hellinger distance have been computed through designed *ad hoc* perl scripts. Both the scripts are available online at https://www.iasi.cnr.it/dsantoni/SARS-CoV2_USA_entropy/ along with a Readme document providing further explanations about the information theory methods and with sample files in order to run the scripts on sample data.

2.3. Pairwise distance correlation matrix and Phylogenetic trees

For each pair of proteins p_i, p_j for $1 \leqslant i, j \leqslant 27$ the two Average mutation Entropy profiles were compared and the Pearson correlation $C(i,j)$ value was computed. A pairwise distance Correlation Matrix CM was built where each component of the matrix was derived from $C(i,j)$ through the formula

$$CM(i,j) = \sqrt{2 * (1 - C(i,j))} \tag{5}$$

.

The function CM maps Pearson correlation values in the range [-1,1] to the reverse range [0,2], where two profiles whose Pearson correlation is equal to 1 will provide a CM equal to 0 and on the other hand a negative correlation -1 will provide a CM equal to 2. The function CM is a distance since all the three properties characterizing a distance are satisfied, trivially $CD(i,i) = 0$ and $CM(i,j) = CM(j,i)$, for the triangular inequality see (Mantegna, 1999) as a reference. The pairwise distance matrix CM was used to build a phylogenetic tree through UPGMA algorithm (Sokal and Michener, 1958), implemented in Phylip package (http://cmgm.stanford.edu/phylip/). The Phylogenetic Tree (as reported in Fig. 7) was displayed through the online tool Tree Of Life (iTOL) v5 (Letunic and Bork, 2021).

3. Results

The mutation rate of SARS-CoV-2 over time (with a monthly temporal interval) was evaluated by considering the Average mutation Entropy AE and Average Hellinger distance AH between consecutive months. We first analyzed Spike protein profiles related to sequences collected in four difference countries for comparison. We then analyzed all protein profiles derived from sequences collected in USA (all of them are available as Supplementary Material S1) and compared profiles of different functional protein classes providing some global statistics. The profiles were also analyzed to investigate the relationships between variants and AE. Finally we computed pairwise distances between proteins via the Pearson correlation values of their AE profiles and built a mutation phylogenetic tree of proteins.

3.1. Average mutation Entropy and Average Hellinger profiles of SARS-CoV-2 proteins

Shannon Entropy and Hellinger distance provide complementary information about global protein mutation landscape. AE measures how much the given protein is exploring the configurational space while AH measures the change of mutational "direction" between two time intervals. To some extent we could say they account for *quantity* and *quality* of mutations respectively.

Fig. 1 reports Average mutation Entropies and Average Hellinger distances for Spike sequences collected in four different countries (the countries where the highest numbers of sequences were collected): USA (panel A), UK (panel B), Germany (panel C) and Denmark (panel D). In each panel (as in the panels of the following figures) the black dots represent AE (y axis) as a function of the time interval - month (x axis) while the asterisks represent AH (y axis) between two consecutive months (x axis, midpoint between two consecutive months). The three color bands represent the period in which variants Alpha (gray), Delta (pink) and Omicron (cyan) are predominant respectively. We considered a variant as predominant (with respect to a set of sequences) if all the mutations officially associated with that variant (source CDC [1]) show a frequency higher than 50%. In all four panels the typical AE V slope can be observed when profiles are superimposed on the variants. Entropy tends to increase when a new variant appears and reaches its peak when the existing variant and the new one coexist. When the new variant becomes predominant a minimum entropy value is observed (V bottom). Then again AE starts to increase as the virus tries to explore new configurations until a new variant appears competing with the existing one.

As expected the shapes of the four profiles are comparable even if some differences can be observed. The typical V shape is evident for all the variants with the exception of Alpha variant in panel A (USA) where a higher and ascending AE profile is observed. It is worth noting the AH peak in panel B (UK) between December 21 and January 22. Even though AE is comparable between the two months a very different mutational scenario is observed due to the extremely high AH value, testifying to a change in the mutational trajectory associated with the new variant. The same peak is observed in panel D (Denmark) and also in panel A (USA) and C (Germany) even if AH is lower than AE.

To further investigate this issue and to provide a practical example of the significance of the coupled analysis provided by AH and AE we focused on the mutational scenarios of sequences collected in UK in

[1] https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classificaVirusestions.html

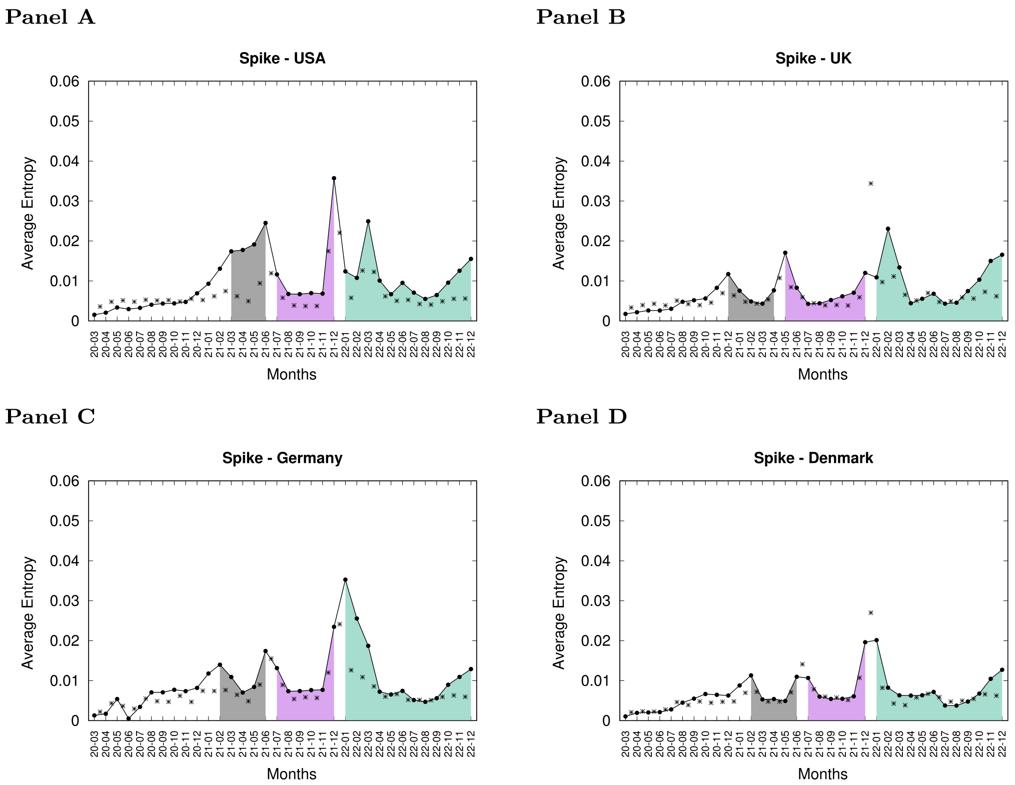

Fig. 1. Average mutation Entropy and Average Hellinger distance profiles of the Spike protein in different countries. Time periods in which a particular variant is predominant are highlighted with a colored band, Alpha in gray, Delta in pink and Omicron in cyan.

December 2021 and January 2022. The number of occurrences of each amino acid (20 standard amino acids) or Deletion for each position of the protein in the multiple alignemnts are reported in Supplementary Material STAB1 (B sheet for December 21 and C sheet for January 22) along with the Hellinger distance between the two distributions and the Shannon entropies of the two months for each position (A sheet). Table 2 provides detailed information on amino acid positions showing the highest Hellinger distances (third column), the highest Entropies for

Table 2
Most significant amino acid positions for sequences collected in UK in December 21 and January 22. The top rows of the table display the amino acid positions showing the five highest Hellinger distances (Hellinger distances are highlighted in bold). The center rows of the table show the five highest entropies for December 2021 (entropy values for December 2021 are highlighted in bold). The bottom rows of the table show the five highest entropies for January 2022 (entropy values for January 2022 are highlighted in bold). Amino acid positions are indicated according to the multialignmment (first column) and according to the reference Spike sequence YP_009724390.1 (second column). The third column represents the Hellinger distance, followed by the entropies for December 2021 (fourth column) and January 2022 (eighth column). Additionally, the three highest amino acid frequencies (when higher than 0.01) for December 2021 are reported in columns 5–7, and for January 2022 in columns 9–11.

					Positions showing the highest Hellinger distance					
Pos ma	**Pos ref**	**H**	**E 21–12**	**1AA**	**2AA**	**3AA**	**E 22–01**	**1AA**	**2AA**	**3AA**
557	505	**0.87**	0.13	Y 0.98	H 0.02		0.09	H 0.99	Y 0.01	
550	498	**0.87**	0.13	Q 0.98	R 0.02		0.09	R 0.98	Q 0.01	
553	501	**0.87**	0.13	N 0.98	Y 0.02		0.09	Y 0.98	N 0.01	
502	452	**0.87**	0.14	R 0.98	L 0.02		0.09	L 0.98	R 0.01	
413	371	**0.87**	0.14	S 0.98	L 0.02		0.35	L 0.94	F 0.04	S 0.01
					Positions showing the highest Entropy for December 21					
Pos ma	**Pos ref**	**H**	**E 21–12**	**1AA**	**2AA**	**3AA**	**E 22–01**	**1AA**	**2AA**	**3AA**
157	145	0.80	**0.99**	Y 0.70	H 0.28	DEL 0.02	0.33	DEL 0.94	Y 0.05	
256	222	0.34	**0.87**	A 0.72	V 0.028		0.05	A 0.99		
99	95	0.16	**0.69**	I 0.82	T 0.18		0.27	I 0.95	T 0.04	
1327	1264	0.18	**0.42**	V 0.92	L 0.08		0.06	V 0.99	L 0.01	
39	36	0.14	**0.27**	V 0.95	F 0.04		0.01	V 0.99		
					Positions showing the highest Entropy for January 22					
Pos ma	**Pos ref**	**H**	**E 21–12**	**1AA**	**2AA**	**3AA**	**E 22–01**	**1AA**	**2AA**	**3AA**
388	346	0.40	0.03	R 0.99			**0.92**	R 0.66	K 0.33	
757	701	0.27	0.07	A 0.99			**0.74**	A 0.79	V 0.21	
413	371	0.87	0.14	S 0.98	L 0.02		**0.35**	L 0.94	F 0.04	S 0.01
21	19	0.87	0.15	R 0.98	T 0.02		**0.35**	T 0.94	I 0.04	R 0.01
157	145	0.80	0.99	Y 0.70	H 0.28	DEL 0.02	**0.33**	DEL 0.94	Y 0.05	

December 2021 (fourth colum) and for January 2022 (eighth column) along with the three highest amino acid frequencies (when higher than 0.01). As can be observed, the positions with the highest Hellinger distances (top rows of the table) exhibit completely different amino acid frequency distributions indicating a transition from a variant to another. For position 557 the Y - Tyrosine - is dominant with 0.98 in December 21 while H - Histidine -becomes dominant with 0.99 in January 22. Similar behaviour can be observed for position 550 (Q - Glutamine - 0.98 in December 21 and R - Arginine - 0.98 in January 22), 553 (N - Asparagine - 0.98 in December 21 and Y - Tyrosine - 0.98 in January 22), 502 (R - Arginine - 0.98 in December 21 and L - Leucine - 0.98 in January 22) and 413 (S - Serine - 0.98 in December 21 and L - Leucine - 0.94 in January 22). Position with the highest entropies for both the months typically show the same dominant amino acid albeit with different frequencies.

In Fig. 2–4 *AE* - *AH* profiles are reported for proteins of different functional classes: non-structural, structural and accessory proteins respectively. Four representatives for each class are shown (see Supplemental Material S1 for the complete list of profiles). As can be observed non-structural proteins typically show a flat and low profile (see S1) when compared to proteins of structural and accessory classes, with only a few exceptions (NSP1 and NSP6 panel B Fig. 2). Some Non-Structural proteins have a very low mutation rate and only a few amino acids are different from the original reference sequence, for example NSP10 and NSP12 reported in Panel C and D of Fig. 2 respectively. It is reasonable to expect this scenario since it can be hypothesized that structural and accessory proteins are more exposed to selective pressures resulting in broader exploration of the configurational space.

In general *AE* peaks within the period of a variant are associated with low *AH* while peaks occurring in periods when the variant is changing are associated with high *AH* (see for example Panel A Fig. 1 Spike protein USA - Alpha variant - gray - March/June 21 or Panel C in Fig. 4 protein NS7b protein - Omicron variant - cyan - June/October 22).

Fig. 5 shows the average and standard deviation of *AE* over all the months of the period considered for each protein. Once again as observed above the non-structural proteins show lower values for both average and standard deviation, while the accessory proteins in particular NS8, NS9b and NS9c and structural proteins in particular N and Spike show the highest values.

In Fig. 6 the mean *AE* for all the proteins is shown for every single variant period. As can be observed, *AE*s typically show higher values for Alpha variant specifically N, NS3, NS8, NS9c, NSP6 and Spike. NS9c profile reported in Fig. 4 panel D is particularly significant; *AE* values reach a peak of almost 0.04 in the period ranging from March until June 21 where the Alpha variant is predominant. Interestingly, some proteins such as NS7a and NS7b show their highest values for Delta variant while some others such as NSP1, NS9b and NS6 show their highest values for the Omicron variant.

3.2. Pairwise distances between profiles and Average mutation Entropy phylogenetic tree

Pairwise distances between Average mutation Entropy protein profiles were derived from Pearson correlation values and a phylogenetic protein tree was built using UPGMA (see Material and Methods Section for details). The phylogenetic tree, as reported in Fig. 7, shows how proteins tend to co-mutate or co-evolve. The closer the proteins are in the tree (when they occur in the same branch) the higher the Pearson correlation value of their profiles. It is reasonable to hypothesize that two proteins that physically or functionally interact with each other are likely to co-evolve or share on average a common mutation pattern over time, whereby a high correlation value should be observed between their entropy profiles. The phylogenetic tree can be a valuable tool to infer physical or functional relationships between protein or cluster of proteins. It can be observed that Spike and Membrane proteins fall in the same branch of the tree and three out of four structural proteins Spike, Membrane and Envelope fall very close to each other, providing consistency to our approach as the three structural proteins are known to interact with each other (Kumar et al., 2023).

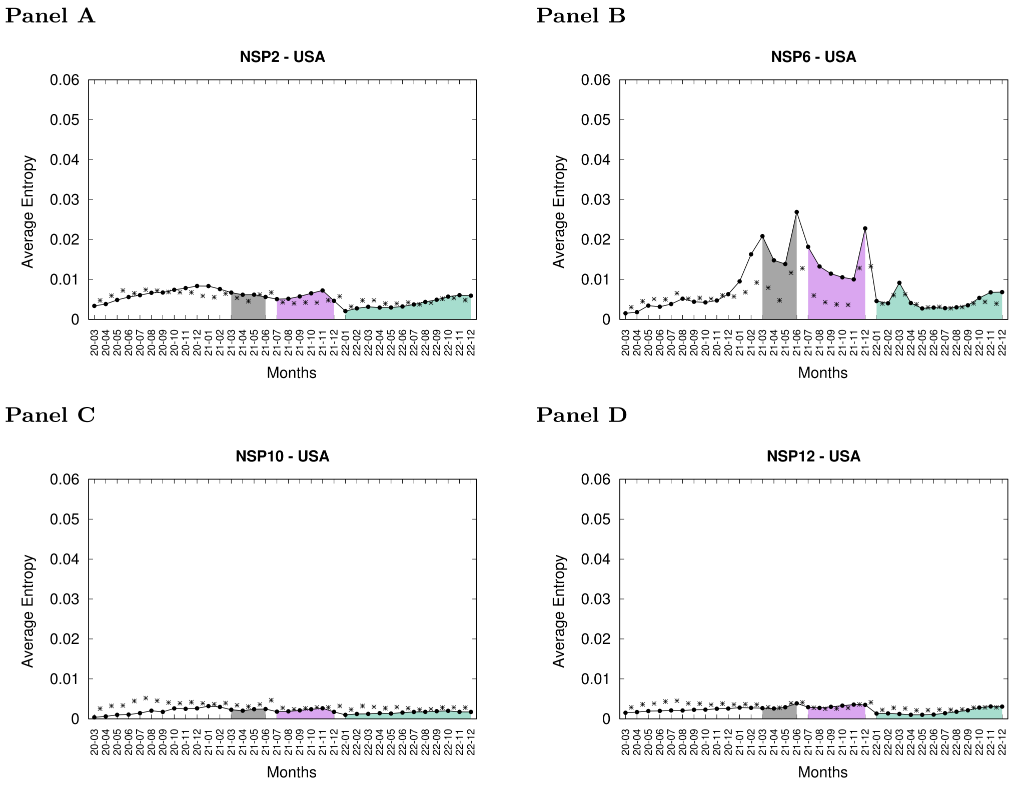

Fig. 2. Average mutation Entropy and Average Hellinger distance profiles of four representative non-structural Proteins. Time periods in which a particular variant is predominant are highlighted with a colored band, Alpha in gray, Delta in pink and Omicron in cyan.

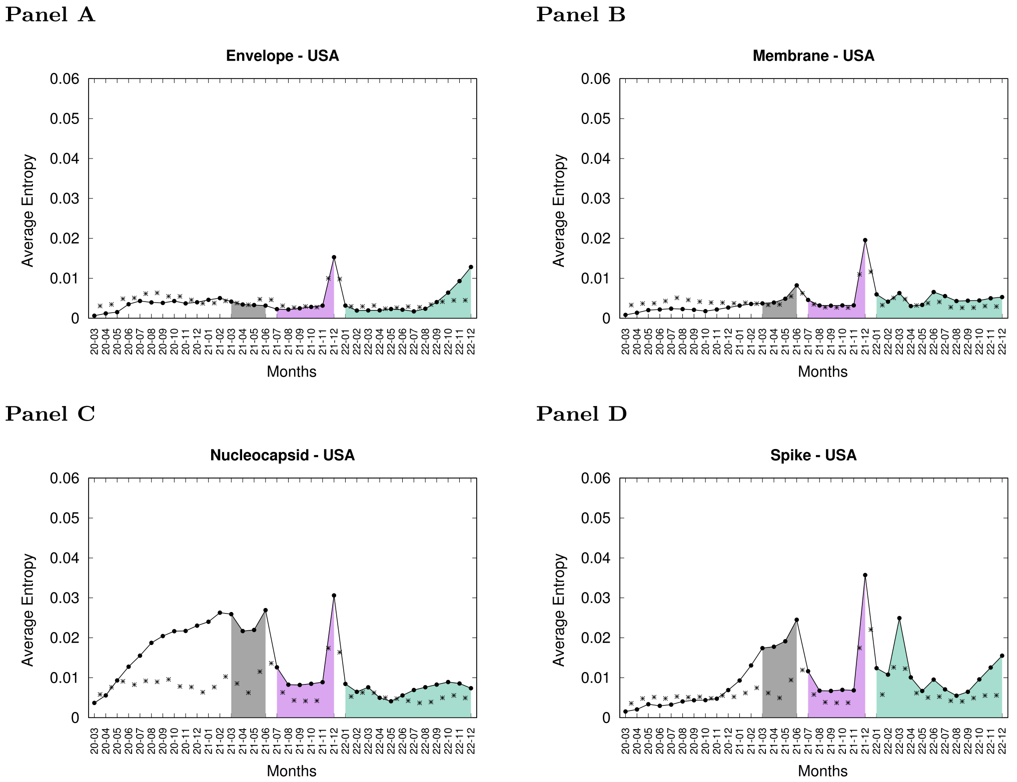

Fig. 3. Average mutation Entropy and Average Hellinger distance profiles of four representative structural Proteins. Time periods in which a particular variant is predominant are highlighted with a colored band, Alpha in gray, Delta in pink and Omicron in cyan.

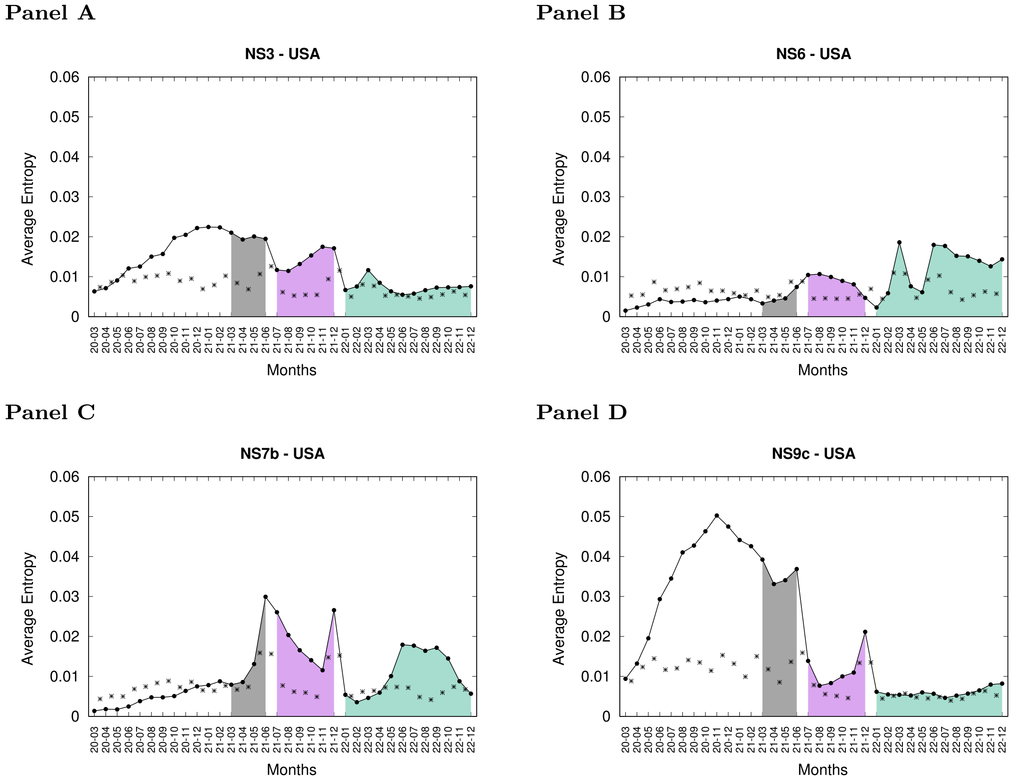

Fig. 4. Average mutation Entropy and Average Hellinger distance profiles of four representative accessory Proteins. Time periods in which a particular variant is predominant are highlighted with a colored band, Alpha in gray, Delta in pink and Omicron in cyan.

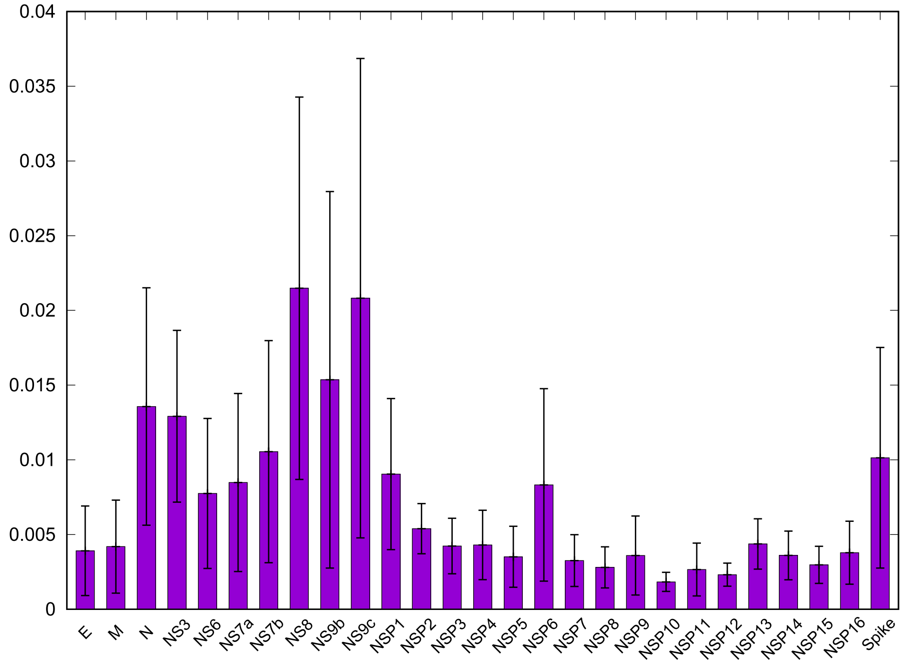

Fig. 5. Average and Standard Deviation of protein mutational entropy over the months in the period March 2020 - December 2022.

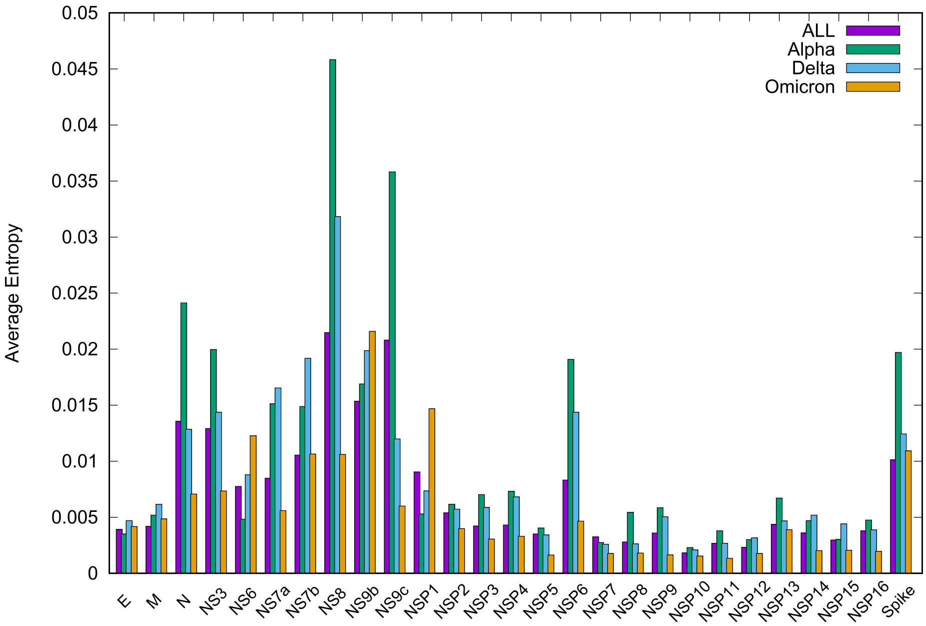

Fig. 6. Mean *AE* of proteins for all the considered months (purple), for months in which Alpha (green), Delta (cyan) or Omicron (orange) were predominant.

Interestingly the NSP2 protein, whose function is currently still debated (Angeletti et al., 2019; Davies et al., 2020) and remains substantially unknown, falls in the same branch as NSP14 and NSP10 (which are known to form a complex (Baddock et al., 2022)). This finding would deserve further investigation to suggest a possible role for NSP2.

4. Conclusion

The present work focuses on the analysis of SARS-CoV-2 mutation profiles over time through an entropy-based approach. Average mutation Entropy *AE* and Average Hellinger distance *AH* (between two consecutive months) profiles are shown over time for each viral protein as reported in Fig. 1, Fig. 2, Fig. 3 and Fig. 4 (see Supplemental Material S1 for a complete view of protein profiles). This representation allows an at-a-glance view of the mutational landscape of viral proteins over time in a period from March 2020 until December 2022. The analysis of mutational profiles can provide insights on the evolution of the virus from different perspectives. In this view we can claim that the present work has two major aims or in other words that the significance of the present work can be read from two different points of view. Firstly it paves the way for the analysis of the mutational landscape of an organisms over time via an entropy-based approach. This novel approach is based on coupling Shannon Entropy and Hellinger distance analysis. Combined together they are able to provide complementary information resulting in a global view on the mutation trajectory of the considered

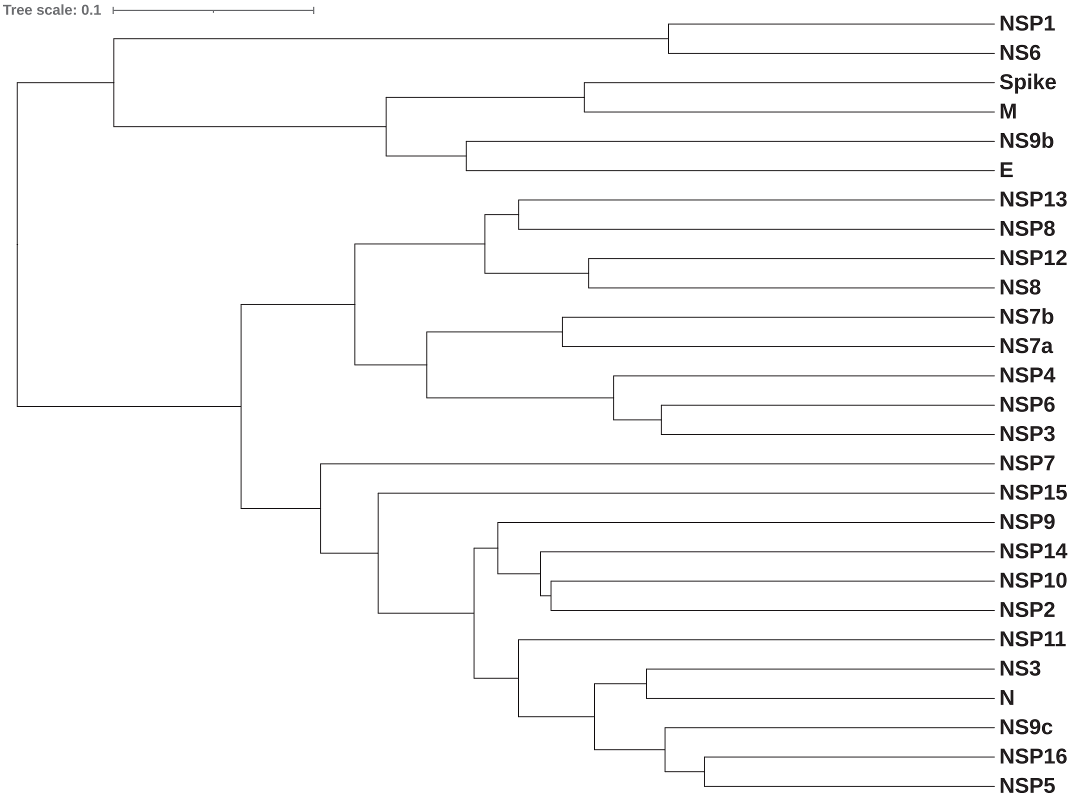

Fig. 7. Phylogenetic tree.

organism. To the best of our knowledge this kind of approach is something that was lacking or at least was under explored. Secondly this analysis has been applied to SARS-CoV2 protein sequences providing useful insights on the virus biology. Comparison of protein profiles coming from different functional classes reveals different behaviours. Non-structural proteins show flat profiles characterized by a very low Average mutation Entropy, with only a few exceptions. On the contrary accessory and structural proteins mostly show (in particular N, NS3, NS8, NS9b, NS9c and Spike) non uniform and high *AE* and *AH* profiles, often coupled with the predominance of variants. This observed different behaviour between non-structural on one hand and accessory and structural proteins on the other hand is something expected to some extent, since structural and accessory proteins are more exposed to selective pressures resulting in a broader exploration of the configurational space while non-structural proteins are more conserved. Average mutation Entropy profiles can also provide a valuable tool to investigate how proteins are linked to each other by hypothesizing that functionally or physically interacting proteins co-mutate over time and therefore they should share similar profiles. In this view, interestingly, the NSP2 protein, whose function is currently still debated, falls in the same branch as NSP14 and NSP10 in the mutation phylogenetic tree of Fig. 7. It is worth noting that results in this work were obtained by analyzing a massive amount of data (n = 107,939,973) reinforcing the significance and providing effectiveness to our insights. However we believe that a broader and comparative analysis, which was not feasible within the scope of this study due to the extensive amount of data already processed, encompassing more countries, could be an interesting research issue to confirm and generalize the observed behaviors.

Ethics approval and consent to participate

The ethical approval or individual consent was not applicable.

Funding

This work has been partially supported by CNR-IASI Project BIOSYS3 - Optimization, models and algorithms for Bioinformatics and System Science - DIT.AD021.128.

Author contributions

Daniele Santoni: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Supervision; Validation; Writing - original draft; Writing - review & editing.

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 are available at GISAID database as reported in the manuscript in the Section Material and Methods

Acknowledgment

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. Daniele Santoni is member of the Gruppo Nazionale Calcolo Scientifico - Istituto Nazionale di Alta Matematica (GNCS-INdAM).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.gene.2024.148556.

References

Adami, C., 2004. Information theory in molecular biology. Phys. Life Rev. 1 (1), 3–22.

Angeletti, S., Benvenuto, D., Bianchi, M., Giovanetti, M., Pascarella, S., Ciccozzi, M., 2019. COVID-2019: the role of the nsp2 and nsp3 in its pathogenesis. J. Med. Virol. 92, 584–588.

Ashraf, J., Bukhari, S.A.R.S., Kanji, A., et al., 2023. Substitution spectra of SARS-CoV-2 genome from Pakistan reveals insights into the evolution of variants across the pandemic. Sci. Rep. 13, 20955.

Baddock, H.T., Brolih, S., Yosaatmadja, Y., Ratnaweera, M., Bielinski, M., Swift, L.P., Cruz-Migoni, A., Fan, H., Keown, J.R., Walker, A.P., Morris, G.M., Grimes, J.M., Fodor, E., Schofield, C.J., Gileadi, O., McHugh, P.J., 2022. Characterization of the SARS-CoV-2 ExoN (nsp14ExoN-nsp10) complex: implications for its role in viral genome stability and inhibitor identification. Nucl. Acids Res. 50 (3), 1484–1500.

Chanda, P., Costa, E., Hu, J., Sukumar, S., Van Hemert, J., Walia, R., 2020. Information theory in computational biology: where we stand today. Entropy 22, 627.

Davies, J.P., Almasy, K.M., McDonald, E.F., Plate, L., 2020. Comparative multiplexed interactomics of SARS-CoV-2 and homologous coronavirus nonstructural proteins identifies unique and shared host-cell dependencies. ACS Infect. Dis. 6, 3174–3189.

Formentin, M., Chignola, R., Favretti, M., 2024. Optimal entropic properties of SARS-CoV-2 RNA sequences. R Soc Open Sci. 11 (1), 231369.

Ghanchi, N.K., Nasir, A., Masood, K.I., Abidi, S.H., Mahmood, S.F., Kanji, A., Razzaj, S., Khan, W., Shahid, S., Yameen, M., Raza, A., Ashraf, J., Ansar, Z., Dharejo, B., Islam, N., Hasan, Z., Hasan, R., 2021. Higher entropy in SARS-CoV-2 genomes from the first COVID-19 wave in Pakistan. PLOS ONE. 16 (8), e0256451.

Gregori, J., Perales, C., Rodriguez-Frias, F., Esteban, J.I., Quer, J., Domingo, E., 2016. Viral quasispecies complexity measures. Virology. 493, 227–237.

Karlin, S., Mrazek, J., Campbell, A.M., 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. 179 (12), 3899–3913.

Karlin, S., Mrazek, J., 1997. Compositional differences within and between eukaryotic genomes. Proc. Natl. Acad. Sci. USA 94, 10227–10232.

Khare, S., Gurry, C., Freitas, L., Schultz, M.B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R.T., Yeo, W., Curation Team GC and Maurer-Stroh S., 2021. GISAID's Role in Pandemic Response. China CDC Wkly. 3(49): 1049-1051.

Kumar, P., Kumar, A., Garg, N., Giri, R., 2023. An insight into SARS-CoV-2 membrane protein interaction with spike, envelope, and nucleocapsid proteins. J. Biomol. Struct. Dyn. 41 (3), 1062–1071.

Lassmann, T., 2020. Kalign 3: multiple sequence alignment of large datasets. Bioinformatics. 36(6), 1928–1929.

Letunic, I., Bork, P., 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucl. Acids Res. 49, W293–W296.

Magazine, N., Zhang, T., Wu, Y., McGee, M.C., Veggiani, G., Huang, W., 2022. Mutations and Evolution of the SARS-CoV-2 Spike Protein. Viruses. 14 (3), 640.

Mantegna, R.N., 1999. Hierarchical structure in financial markets. Eur. Phys. J. B. 11, 193–197.

Markov, P.V., Ghafari, M., Beer, M., Lythgoe, K., Simmonds, P., Stilianakis, N.I., Katzourakis, A., 2023. The evolution of SARS-CoV-2. Nat. Rev. Microbiol. 21, 361–379.

Mullick, B., Magar, R., Jhunjhunwala, A., Barati, Farimani A., 2021. Understanding mutation hotspots for the SARS-CoV-2 spike protein using Shannon Entropy and K-means clustering. Comput. Biol. Med. 138, 104915.

Namazi, H., Krejcar, O., Subasi, A., 2020. Complexity and information-based analysis of the variations of the SARS-COV-2 genome in the United States of America. Fractals. 28 (7), 2150023.

Pensar, J., Puranen, S., Arnold, B., MacAlasdair, N., Kuronen, J., Tonkin-Hill, G., Pesonen, M., Xu, Y., Sipola, A., Sánchez-Busó, L., Lees, J.A., Chewapreecha, C., Bentley, S.D., Harris, S.R., Parkhill, J., Croucher, N.J., Corander, J., 2019. Genome-wide epistasis and co-selection study using mutual information. Nucl. Acids Res. 47 (18), e112.

Rhee, S.Y., Liu, T.F., Kiuchi, M., Zioni, R., Gifford, R.J., Holmes, S.P., Shafer, R.W., 2008. Natural variation of HIV-1 group M integrase: implications for a new class of antiretroviral inhibitors. Retrovirology. 7, 5–74.

Rogozin, I.B., Saura, A., Poliakov, E., Bykova, A., Roche-Lima, A., Pavlov, Y.I., Yurchenko, V., 2024. Properties and Mechanisms of Deletions, Insertions, and Substitutions in the Evolutionary History of SARS-CoV-2. Int. J. Mol. Sci. 25, 3696.

Santoni, D., Ghosh, N., Saha, I., 2022. An entropy-based study on mutational trajectory of SARS-CoV-2 in India. Infect Genet Evol. 97, 105154.

Vergni, D., Santoni, D., Bouba, Y., Lemme, S., Fabeni, L., Carioti, L., Bertoli, A., Gennari, W., Forbici, F., Perno, C.F., Gagliardini, R., Ceccherini-Silberstein, F., Santoro, M.M., on behalf of the HIV drug-resistance group, 2022. Evaluation of HIV-1 integrase variability by combining computational and probabilistic approaches. Infect Genet. Evol. 101: 105294.

Vinga, S., Almeida, J.S., 2004. Rényi continuous entropy of DNA sequences. J Theor Biol. 231 (3), 377–388.

Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating systematic relationships. Univ. Kans Sci. Bull. 38, 1409–1438.

Vinga, S., 2014. Information theory applications for biological sequence analysis. Brief Bioinform. 15 (3), 376–389.

Vopson, M.M., Robson, S.C., 2021. A new method to study genome mutations using the information entropy. Phys. A: Stat. 584, 126383.