Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Infection, Genetics and Evolution 97 (2022) 105154

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

An entropy-based study on mutational trajectory of SARS-CoV-2 in India

Daniele Santoni [a,*], Nimisha Ghosh [b,c], Indrajit Saha [d]

[a] Institute for System Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Via dei Taurini 19, Rome 00185, Italy
[b] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland
[c] Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India
[d] Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

ARTICLE INFO

Keywords:
COVID-19
Entropy
Hellinger distance
Mutation
SARS-CoV-2

ABSTRACT

The pandemic of COVID-19 has been haunting us for almost the past two years. Although, the vaccination drive is in full swing throughout the world, different mutations of the SARS-CoV-2 virus are making it very difficult to put an end to the pandemic. The second wave in India, one of the worst sufferers of this pandemic, can be mainly attributed to the Delta variant i.e. B.1.617.2. Thus, it is very important to analyse and understand the mutational trajectory of SARS-CoV-2 through the study of the 26 virus proteins. In this regard, more than 17,000 protein sequences of Indian SARS-CoV-2 genomes are analysed using entropy-based approach in order to find the monthly mutational trajectory. Furthermore, Hellinger distance is also used to show the difference of the mutation events between the consecutive months for each of the 26 SARS-CoV-2 protein. The results show that the mutation rates and the mutation events of the viral proteins though changing in the initial months, start stabilizing later on for mainly the four structural proteins while the non-structural proteins mostly exhibit a more constant trend. As a consequence, it can be inferred that the evolution of the new mutative configurations will eventually reduce.

1. Introduction

Almost two years but COVID-19, the disease caused by SARS-CoV-2 is still disrupting our daily lives. Most of the cities around the globe including India have gone through various stages of lockdown to contain the spread of the virus. While the development and dissemination of vaccines have brought rays of hope, the circulation of the different variants of SARS-CoV-2 is still a cause of worry. As of now, the major variants of concern as declared by W.H·O are Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1) and Delta (B.1.617.2). Among these, the Delta variant was mainly responsible for the catastrophic second wave in India. Thus, mutations of SARS-CoV-2 need to be studied in order to understand its evolution. In this regard, Martin et al. (2021) studied Alpha, Beta and Gamma lineages to understand the evolutionary pattern of the virus. Dorp et al. (2020) curated 7666 SARS-CoV-2 genomes to analyse the genomic diversity of SARS-CoV-2, thereby identifying 198 filtered recurrent mutations. Yuan et al. (2020) performed global analysis of 11,183 sequences to reveal the genetic diversity of SARS-CoV-2. Phylogenetic analysis was also carried out by Bai et al. (2020) with 16,373 SARS-CoV-2 genomes to reveal the evolution and molecular characteristics of SARS-CoV-2 while in Saha et al. (2021) the authors analysed more than 10,000 global SARS-CoV-2 genomes which identified 7209, 11,700, 119 and 53 unique mutation points as substitutions, deletions, insertions and SNPs. In Ghosh et al. (2021) performed phylogenetic analysis on more than 18,000 sequences to identify signature SNPs. Furthermore, in Saha et al. (2020a, 2020b), Saha et al. performed analysis of 566 Indian SARS-CoV-2 genomes to find the major mutation points in such genomes. On the other hand, entropy has been proven to be a very potent tool for the analysis of epidemics (Lucia et al., 2020). In Santos et al. (2021), Santos et al. proposed EntroPhylo which is an entropy based tool to select phylogenetic informative genetic regions and primer design. The tool considers the entropy value of each site and consequently the selected region is used for primer design. For evaluation purpose, EntroPhylo was used on the sequences of bovine

* Corresponding author.
E-mail address: daniele.santoni@iasi.cnr.it (D. Santoni).

https://doi.org/10.1016/j.meegid.2021.105154
Received 1 October 2021; Received in revised form 17 November 2021; Accepted 17 November 2021
Available online 19 November 2021

Table 1
Number of sequences SARS-CoV-2 proteins for each month.

SARS-CoV-2 Proteins	Number of sequences per month																
	Mar 20	Apr 20	May 20	Jun 20	Jul 20	Aug 20	Sep 20	Oct 20	Nov 20	Dec 20	Jan 21	Feb 21	Mar 21	Apr 21	May 21	Jun 21	Jul 21
STRUCTURAL																	
Envelope	184	435	971	1062	681	631	628	380	451	981	500	980	1904	3051	2401	1289	631
Membrane	189	441	977	1062	683	632	626	380	451	979	500	980	1857	3006	2407	1293	632
Nucleocapsid	188	440	969	1054	679	629	629	380	449	976	495	971	1875	3032	2386	1279	620
Spike	188	437	954	1025	669	616	602	353	410	893	476	893	1735	2600	2048	1124	601
Open Reading Frames (ORFs)																	
ORF3a	189	441	963	1049	633	560	589	334	376	832	354	783	1802	3014	2399	1289	631
ORF6	189	441	976	1061	681	632	629	380	452	981	499	979	1905	3046	2401	1292	630
ORF7a	186	440	968	1057	678	622	627	377	449	979	493	962	1881	3020	2330	1245	623
ORF 7b	189	441	975	1055	681	626	627	379	448	969	488	942	1860	3041	2395	1288	628
ORF8	159	432	951	1033	671	621	609	367	429	912	460	819	1266	2640	2352	1264	622
ORF10	189	441	976	1057	680	632	629	377	450	968	497	973	1869	3051	2404	1292	625
Non-Structural Proteins (NSPs)																	
NSP1	189	441	970	1061	680	629	628	379	452	979	499	977	1903	3040	2398	1285	630
NSP2	188	434	966	1060	683	630	629	380	450	977	499	966	1897	3036	2384	1276	627
NSP3	185	429	961	1044	680	629	621	370	437	942	494	937	1777	2872	2334	1249	623
NSP4	187	440	975	1062	683	631	629	380	451	983	498	976	1900	3045	2400	1290	632
NSP5	189	440	977	1061	682	632	628	380	452	982	499	979	1906	3052	2408	1293	632
NSP6	189	439	977	1061	682	632	629	380	452	981	499	951	1722	2938	2392	1288	631
NSP7	189	441	977	1062	683	632	629	380	451	983	500	980	1905	3054	2406	1293	632
NSP8	189	441	977	1062	683	632	629	380	452	982	499	980	1901	3053	2408	1293	632
NSP9	189	440	977	1062	683	632	629	380	452	983	500	980	1905	3054	2407	1293	632
NSP10	189	441	977	1062	683	632	629	379	452	981	500	980	1904	3053	2400	1293	632
NSP11	189	441	977	1062	683	632	629	377	452	981	498	980	1907	3054	2408	1293	632
NSP16	187	440	967	1053	675	605	613	371	450	969	496	972	1903	3041	2402	1292	632
OTHER NSPs																	
EndoRNAse	189	441	974	1060	683	632	628	380	451	979	499	979	1907	3042	2398	1291	632
Exon	188	437	960	1035	658	599	627	376	450	975	489	964	1837	2946	2271	1240	567
Helicase	189	440	974	1062	683	632	629	379	452	983	500	979	1906	3050	2401	1290	631
RdRp	189	437	971	1057	681	632	629	376	449	963	495	974	1898	3049	2405	1290	629

papillomavirus L1 gene. Vopson et al. (Vopson and Robson, 2021) have studied genetic mutations by considering information entropy of genome and have tested their method on the reference sequence of SARS-CoV-2. Site specific entropy analysis was carried out by Ghanchi et al. (2021) on 90 SARS-CoV-2 genomes to investigate phylogeny, genetic variation and mutation rates of the SARS-CoV-2 strains in Pakistan. They concluded that the higher entropy and diversity that was observed in the early days of pandemic as compared with later strains suggest the increasing stability of the genomes in the subsequent waves of COVID-19.

Motivated by the literature, in this work, we have analysed the 26 virus proteins of SARS-CoV-2 using entropy-based approach to find the mutational trajectory of SARS-CoV-2. In this regard, more than 17,000 protein sequences of Indian SARS-CoV-2 genomes from the months of March 2020 to July 2021 are considered for alignment using MAFFT (Katoh et al., 2002). Subsequently, it can be observed that till the months of February–March 2021 there is an increase in average entropy of each of the structural proteins and then they start declining, thereby indicating that the entropy of each such SARS-CoV-2 protein seem to have reached a sort of stability in India. On the other hand, the non-structural proteins mostly exhibit a more constant trend of mutational entropy. Moreover, Hellinger distance is also used in this work to show the difference of the mutation events between each consecutive month for each of the 26 virus proteins which further substantiates our claim that the evolution of the new mutative configurations will eventually reduce. It is to be noted that, in the past year, there has been a lot of work pertaining to the evolution of SARS-CoV-2. But mostly those works involve phylogenetic analysis to identify the mutation points in the virus like (Saha et al., 2021). However, to the best of the authors' knowledge, mutational trajectory considering entropy for Indian sequences is a novel topic which has not been addressed as yet in the literature.

2. Materials and methods

2.1. Preparation of sequence dataset

Initially, 17,271 available genomic sequences of SARS-CoV-2 from India spanning March 2020 to July 2021 are retrieved from GISAID.[1] These sequences are aligned with respect to the reference sequence (NC 045512.2)[2] using MAFFT (Katoh et al., 2002). Thereafter, the aligned sequences are translated into the protein sequences. In this study, 26 virus proteins are considered and corresponding sequence datasets are built accordingly from the aforementioned 17,271 protein sequences. Each dataset is cleaned by removing those sequences with 1) more than three amino acid consecutive deletions, 2) stop codons inside the Open Reading Frame and 3) number of consecutive deletions not a multiple of 3. After these refinements, the number of sequences for each month is as reported in Table 1. Each sequence is related to a given month starting from March 2020 till July 2021.

2.2. Entropy-based approach

Let A be a set of symbols 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 (in our case, we have considered the 26 different virus proteins as reported above), where m is the total number of sequences. Since there are no insertions in the considered sequences with respect to the reference, a matrix M of dimension $n \times m$ (where n is the length of reference and m is the number of sequences) can be derived from the performed alignment where each element $m(i,j) \in A$ is an

[1] https://www.gisaid.org/
[2] https://www.ncbi.nlm.nih.gov/nuccore/1798174254

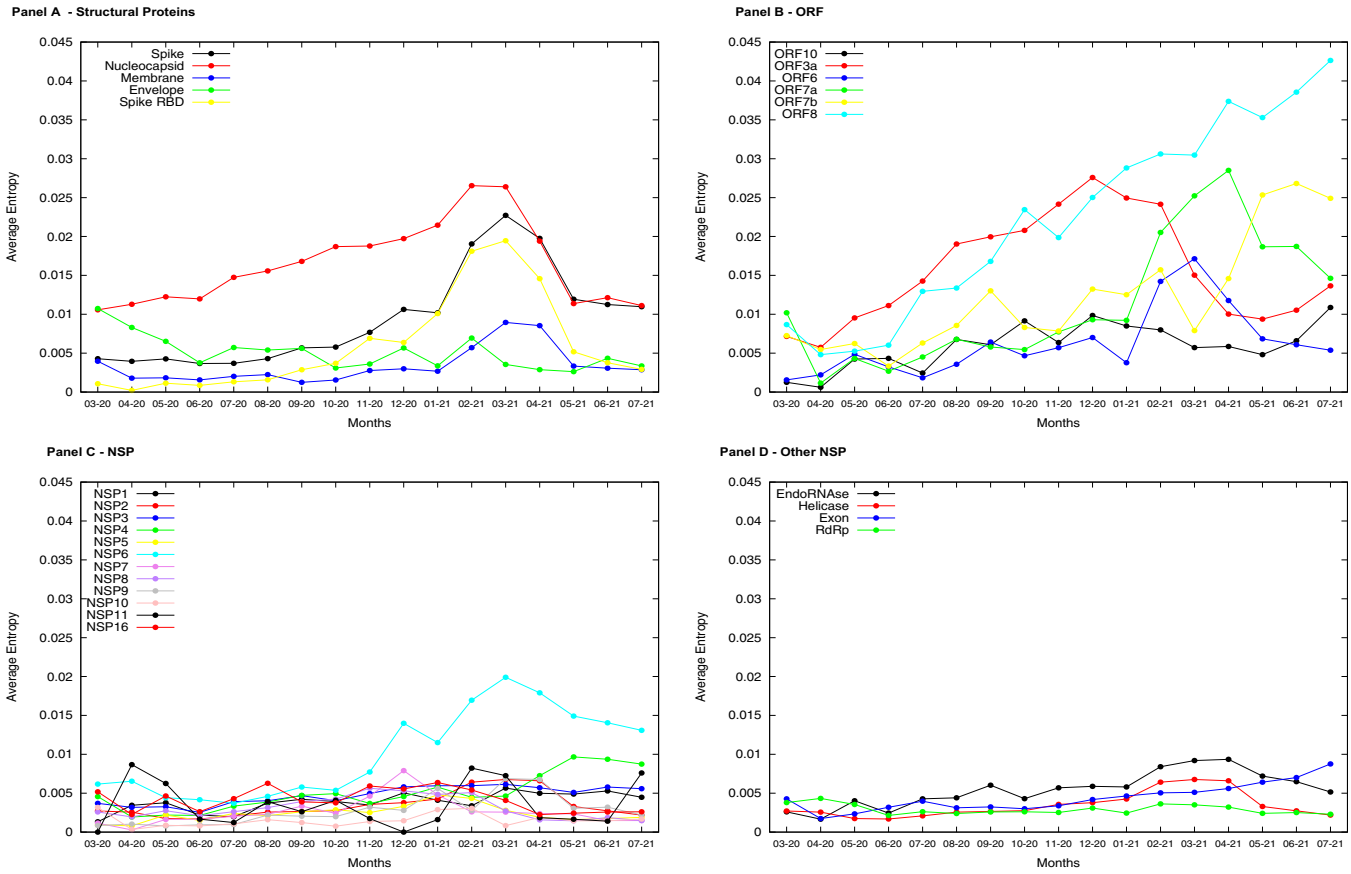

Fig. 1. Average entropy, AVE, over all position sites of amino acid distribution is shown for each month starting from March 2020 till July 2021. Panel A is related to structural proteins (yellow plot is related to average entropy restricted to Receptor Binding Domain of Spike, AVE (RBD)), Panel B to ORF, Panel C to NSP and Panel D to other NSP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

amino acid or a deletion occurring in the alignment in position i of the jth sequence. 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 in position i and the total number of sequences m. Entropy of a position i is computed through the canonical Shannon entropy as:

$$E\left(i\right) = \sum_{a \in A} p_i\left(a\right) log_2\left(p_i\left(a\right)\right)$$

The average entropy AVE over all the positions of a given protein is defined as:

$$AVE = \frac{\sum_{i=1..n} E(i)}{n}$$

When the average entropy is computed on a given subset of sites S, it is indicated as $AVE(S)$:

$$AVE(S) = \frac{\sum_{i \in S} E(i)}{|S|}$$

where $|S|$ is the number of sites in S.

Average entropy is separately computed on sequence sets related to each month starting from March 2020 till July 2021. Another instrument commonly used in information theory to compare two probability distributions, viz. the Hellinger distance is implemented to evaluate how residues are differently distributed between different sample sets for a given position.

Given two sequence sets A and B (in our case they will correspond to sequence associated to two different months) the relative frequencies of a given amino acid (or "-" deletion) for a given position i are defined as $p_i^A(a)$ and $p_i^B(a)$ for A and B respectively. The Hellinger distance between sequence sets A and B related to position i, $H(i)_{A,\,B}$ is defined as follows:

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

As for the entropy, as described above, the average Hellinger distance between the two sets A and B over all the positions of the given protein is defined as:

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

3. Results

The mutation rate of SARS-CoV-2 is evaluated in time (with a temporal interval of a month) by considering average entropy AVE and average Hellinger distance AVH between consecutive months. Figs. 1 and 2 report entropy and Hellinger analysis respectively. In the four panels of both figures, data related to four classes of proteins are reported: structural proteins (panel A), Open Reading Frames (ORF) (panel B), Non-Structural Proteins (NSP) (panel C) and other NSPs (panel D). In panel A of Fig. 1, the analysis is also focused on the Receptor Binding Domain (RBD) of Spike protein restricting the average entropy on positions from 318 to 540 (yellow plot). The entropy analysis clearly shows a significant different behavior between structural, ORF and NSP proteins. In panels A (structural) and B (ORF), there is an

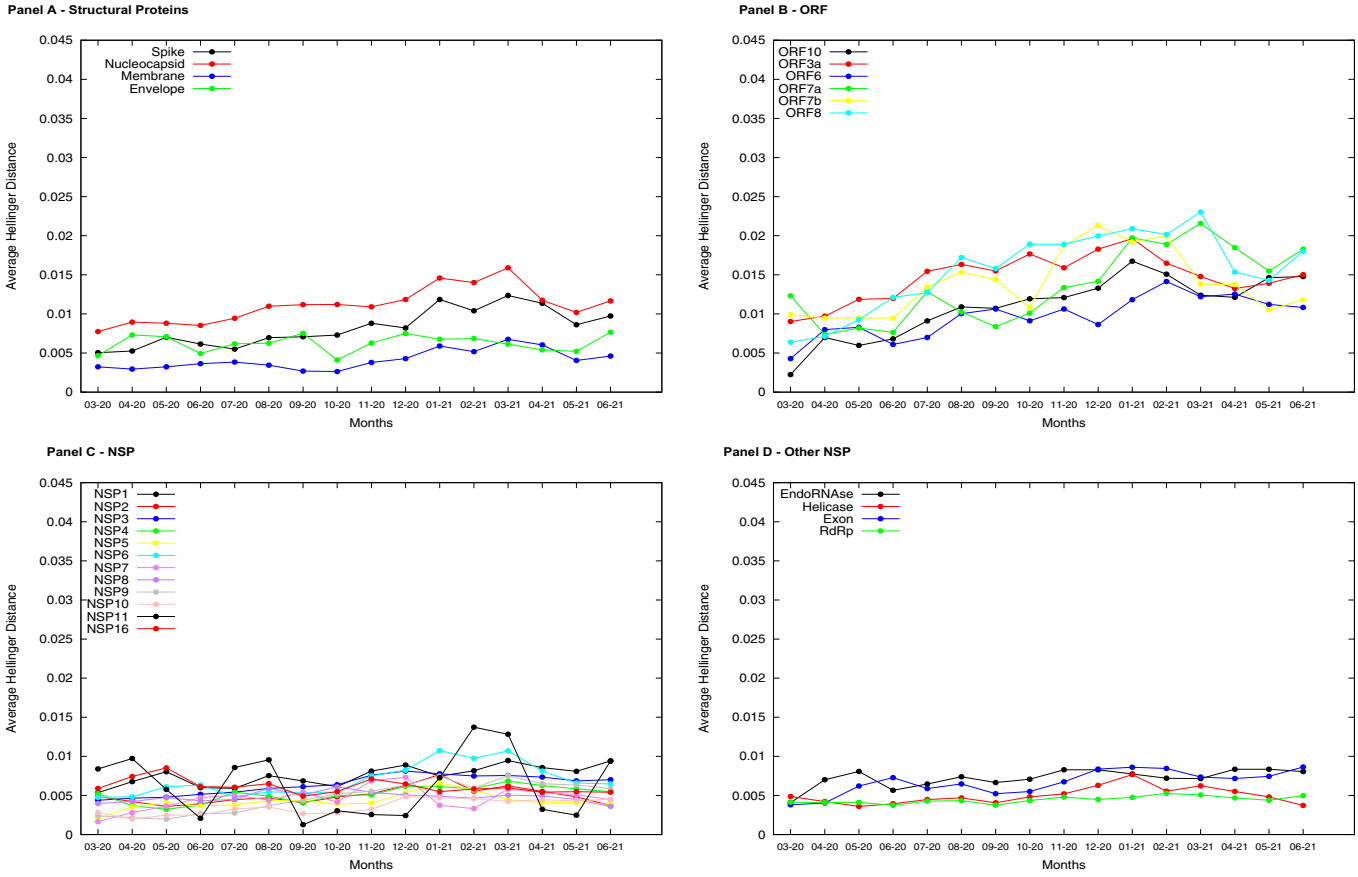

Fig. 2. Average Hellinger distance, AVH, over all position sites, between amino acid distribution of two consecutive months is shown starting from March 2020 till July 2021. Panel A is related to structural proteins, Panel B to ORF, Panel C to NSP and Panel D to other NSP.

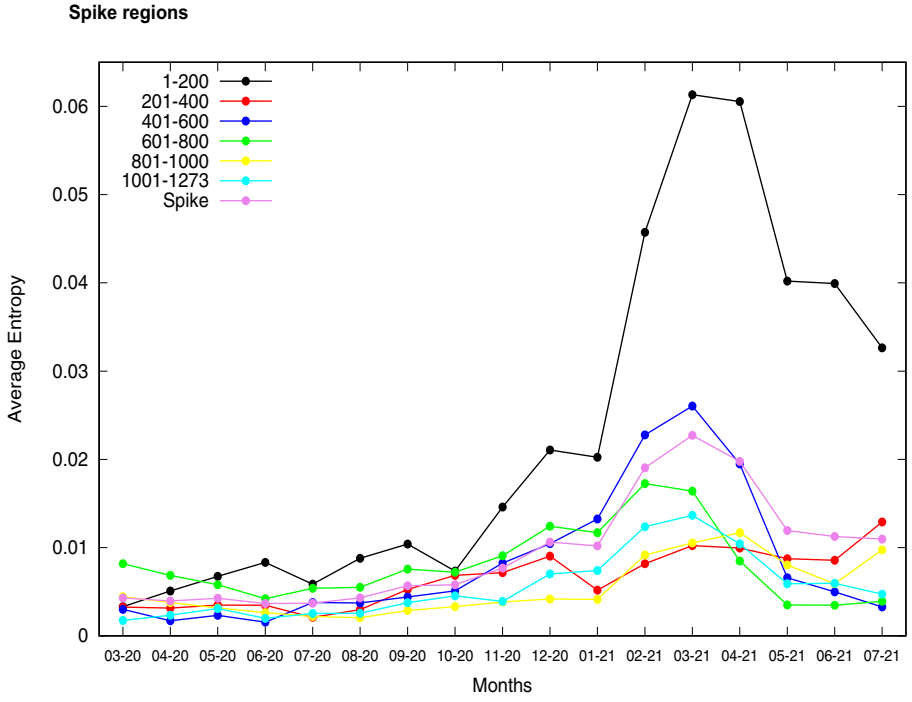

Fig. 3. Fragment-wise entropy of Spike Protein for Delta variant of SARS-CoV-2.

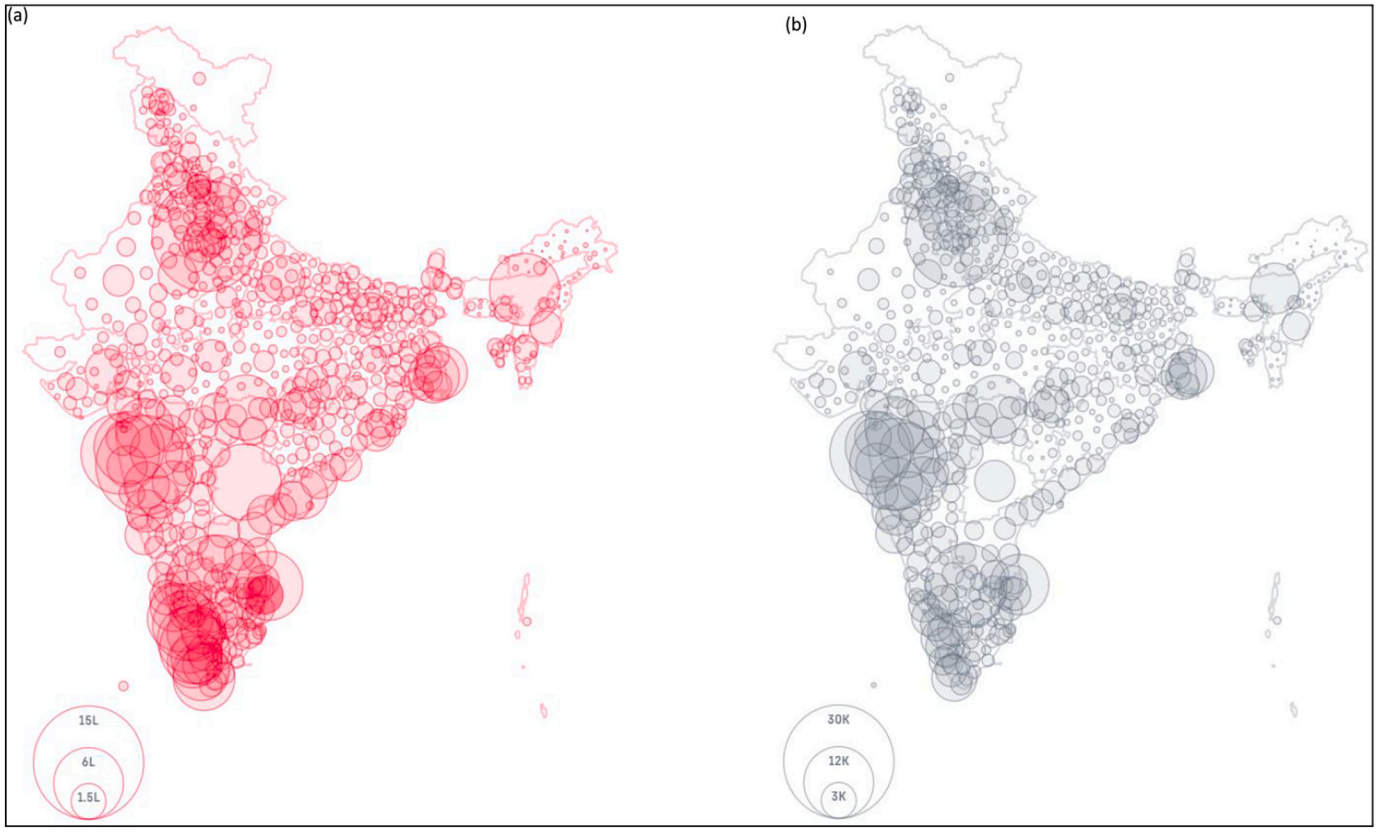

Fig. 4. Illustration of (a) Confirmed and (b) Deceased cases of India to show the effects of SARS-CoV-2 in the different regions of the country.

overall increase of mutation entropy from March 2020 till February–April 2021, where structural proteins typically report the highest entropy values. On the other hand, panels C (NSP) and D (other NSP) exhibit a constant and significantly lower entropy values. Spike (panel A black plot) and Nucleocapsid proteins (panel A red plot) show a more evident and constant increase of entropy till March 2021 while Membrane (panel A blue plot) shows a smaller peak in the same period. After initial higher values (March–April 2020 probably due to deletions around position 38), Envelope (panel A green plot) protein shows low entropy values till July 2021. RBD in Spike follows the same trend as the whole Spike protein, even if with smaller entropy values till February 2021 and then values rapidly decrease till July 2021. From March 2021 till July 2021 entropy values of Spike, Nucleocapsid and Membrane decrease due to the delta variant becoming predominant and limiting the exploration of new mutations that were initially less homogeneously distributed. Plots of Panel B (ORF) show a less homogeneous behavior, in particular ORF8 is the only protein with a constant entropy increase along the entire observed interval time. ORF7a shows a trend that is very similar to structural proteins with a clear entropy peak in April 2021. NSP proteins (panels C and D) show low entropy values and there is no particularly significant increase excepting NSP6 which shows comparatively higher values from November 2020. Obtained results clearly highlight different mutational scenarios between structural and non-structural proteins. Structural proteins typically show high and increasing entropy values in time till March 2021 as the result of the escape strategy from immune system pressure while non-structural proteins, being less under immune system pressure, show lower and constant entropy values.

Hellinger analysis provides a complementary information about how the overall mutation rate, estimated through entropy, is differentially distributed in terms of residue distribution between consecutive months. The entropy accounts for how much the virus is mutating while the Hellinger distance shows the difference of the mutation events between two months. In general, Hellinger distance between two months could be high even if the corresponding entropies assume the same value. Considering this view, a high Hellinger distance value coupled with low difference between entropy values can be interpreted as a change of direction of the mutational trajectory. Hellinger distance values, observed in Fig. 2, are coherent with entropy values observed in Fig. 1, typically showing regular behavior. This is particularly evident for non-structural proteins (panels C and D) showing flat Hellinger distance profiles.

The overall coherence between the two measures (Figs. 1 and 2) means that the mutational scenario, in particular for structural proteins is following a quite regular trend evolving more rapidly till February to March 2021 and then stabilizing the mutational events after those months due to the predominance of Delta variant. Furthermore, Fig. 3 depicts the fragment wise entropy of Spike protein for Delta variant. As can be seen from the figure, the highest contribution to entropy is due to the 1–200 amino acid region. It is to be noted that mutations for Delta variant like T19R, V70F, T95I, G142D, E156-, F157- and R158G occur in this region of Spike protein. The pink line in the figure refers to the reference sequence of the entire Spike protein. Moreover, Fig. 4(a) and (b) show the plot of confirmed and deceased cases in India till 31st October 2021. For example, western part of India has a very high number of confirmed and deceased cases which can be attributed to the Delta variant which was mostly responsible for the catastrophic 2nd wave in India. These two figure are considered from https://www.covid19india.org/.

4. Conclusion

The present work focuses on analysing the average mutation rate of SARS-CoV-2 by considering the 26 virus proteins in India from March 2020 to July 2021 through an entropy-based approach. It is clearly shown that, concerning structural proteins, there is an overall increase

of mutation rate from March 2020 till March 2021 and then it starts to decrease, thereby indicating that the mutation scenario is reaching a sort of stability so that the average entropy decreases. This can be explained by the emergence of Delta variants becoming predominant, so that evolution of the new mutative configurations have reduced. The same trend is observed among the four structural proteins, even if with different scales and minor differences, for both entropy and Hellinger analysis. This reinforces the significance of results providing effectiveness to our insights. On the contrary, non-structural proteins show an overall constant or slightly increasing trend with low entropy values. This behavior can be read as the result of the immune system pressure, acting primarily on structural proteins, pushing the virus to preferentially mutate those proteins, while the effect of selective pressure on non-structural proteins has a minor impact.

In conclusion, the proposed work constitutes a novel approach to the study of mutational trajectories in India and can be applied to different countries and biological datasets to shed light and to provide a different point of view of virus evolution.

Ethics approval and consent to participate

The ethical approval or individual consent was not applicable.

Availability of data and materials

The aligned protein sequences of SARS-CoV-2 genomes are available at "http://www.nitttrkol.ac.in/indrajit/projects/COVID-Mutation Trajectory-India/".

Consent for publication

Not applicable.

Funding

This work has been partially supported by CRG short term research grant on COVID-19 (CVD/2020/000991) from Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India. However, it does not provide any publication fees.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgment

We would like to thank Mr. Suman Nandi for providing the data for the experiments. We would also like to thank all those who have contributed sequences to GISAID and NCBI databases.

References

Bai, Y., Jiang, D., Lon, J.R., et al., 2020. Comprehensive evolution and molecular characteristics of a large number of SARS-CoV-2 genomes reveal its epidemic trends. Int. J. Infect. Dis. 100, 164–173. https://doi.org/10.1016/j.ijid.2020.08.066.

Dorp, L.V., Acman, M., Richard, D., et al., 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. 83, 104351 https://doi.org/10.1016/j.meegid.2020.104351.

Ghanchi, N.K., Nasir, A., Masood, K.I., et al., 2021. Higher entropy observed in SARS-CoV-2 genomes from the first COVID-19 wave in Pakistan. PLoS One 16 (8), e0256451. https://doi.org/10.1371/journal.pone.0256451.

Ghosh, N., Saha, I., Nandi, S., Sharma, N., 2021. Characterisation of sars-cov-2 clades based on signature snps unveils continuous evolution. Methods. https://doi.org/10.1016/j.ymeth.2021.09.005.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res. 30 (14), 3059–3066. https://doi.org/10.1093/nar/gkf436.

Lucia, U., Deisboeck, T.S., Grisolia, G., 2020. Entropy-based pandemics forecasting. Front. Phys. 8, 274. https://doi.org/10.3389/fphy.2020.00274.

Martin, D.P., Weaver, S., Tegally, H., et al., 2021. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. Cell. https://doi.org/10.1016/j.cell.2021.09.003.

Saha, I., Ghosh, N., Maity, D., et al., 2020a. Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP. Infect. Genet. Evol. 85, 104457 https://doi.org/10.1016/j.meegid.2020.104457.

Saha, I., Ghosh, N., Maity, D., et al., 2020b. Inferring the genetic variability in Indian SARS-CoV-2 genomes using consensus of multiple sequence alignment techniques. Infect. Genet. Evol. 85, 104522 https://doi.org/10.1016/j.meegid.2020.104522.

Saha, I., Ghosh, N., Pradhan, A., et al., 2021. Whole genome analysis of more than 10 000 SARS-CoV-2 virus unveils global genetic diversity and target region of NSP6. Brief. Bioinform. 22 (2), 1106–1121. https://doi.org/10.1093/bib/bbab025.

Santos, F.L.S.G., de Sá Leitão Paiva Júnior, S., de Freitas, A.C., et al., 2021. EntroPhylo: an entropy-based tool to select phylogenetic informative regions and primer design. Infect. Genet. Evol. 92, 104857 https://doi.org/10.1016/j.meegid.2021.104857.

Vopson, M.M., Robson, S.C., 2021. A new method to study genome mutations using the information entropy. Phys. A: Stat. Mech. Appl. 584, 126383 https://doi.org/10.1016/j.physa.2021.126383.

Yuan, F., Wang, L., Fang, Y., et al., 2020. Global SNP analysis of 11,183 SARS-CoV-2 strains reveals high genetic diversity. Transbound. Emerg. Dis. 11 https://doi.org/10.1111/tbed.13931.