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.

Computers in Biology and Medicine 141 (2022) 105024

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Feature-extraction and analysis based on spatial distribution of amino acids for SARS-CoV-2 Protein sequences

Ranjeet Kumar Rout [a], Sk Sarif Hassan [b], Sabha Sheikh [a], Saiyed Umer [c], Kshira Sagar Sahoo [d], Amir H. Gandomi [e,*]

[a] Department of Computer Science & Engineering, National Institute of Technology Srinagar, Hazratbal, Jammu and Kashmir, India
[b] Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram, Paschim Medinipur, 721140, India
[c] Department of Computer Science and Engineering, Aliah University, Kolkata, India
[d] Department of Computer Science and Engineering, SRM University, Amaravati, AP, 522240, India
[e] Faculty of Engineering and Information Technology, University of Technology Sydney, NSW, Australia

ARTICLE INFO

Keywords:
Shannon entropy
Hurst exponent
Amino acid
Frequency distribution
SARS-CoV-2

ABSTRACT

Background and objective: The world is currently facing a global emergency due to COVID-19, which requires immediate strategies to strengthen healthcare facilities and prevent further deaths. To achieve effective remedies and solutions, research on different aspects, including the genomic and proteomic level characterizations of SARS-CoV-2, are critical. In this work, the spatial representation/composition and distribution frequency of 20 amino acids across the primary protein sequences of SARS-CoV-2 were examined according to different parameters.

Method: To identify the spatial distribution of amino acids over the primary protein sequences of SARS-CoV-2, the Hurst exponent and Shannon entropy were applied as parameters to fetch the autocorrelation and amount of information over the spatial representations. The frequency distribution of each amino acid over the protein sequences was also evaluated. In the case of a one-dimensional sequence, the Hurst exponent (HE) was utilized due to its linear relationship with the fractal dimension (D), i.e. $D + HE = 2$, to characterize fractality. Moreover, binary Shannon entropy was considered to measure the uncertainty in a binary sequence then further applied to calculate amino acid conservation in the primary protein sequences.

Results and conclusion: Fourteen (14) SARS-CoV protein sequences were evaluated and compared with 105 SARS-CoV-2 proteins. The simulation results demonstrate the differences in the collected information about the amino acid spatial distribution in the SARS-CoV-2 and SARS-CoV proteins, enabling researchers to distinguish between the two types of CoV. The spatial arrangement of amino acids also reveals similarities and dissimilarities among the important structural proteins, E, M, N and S, which is pivotal to establish an evolutionary tree with other CoV strains.

1. Introduction

The novel coronavirus (COVID-19) has rapidly become a major global emergency that has and continues to affect all lives around the globe [1–3]. Presently, this disease, a pandemic as announced by the WHO, is a major health concern [4,5]. Currently, the largest genome (of size approximately 30 kb) for RNA viruses is known as severe acute respiratory syndrome **coronavirus** 2 (SARS-CoV-2) [6,7]. Coronaviruses (CoVs) are classified into three different classes, including α-CoV, β-CoV, and γ-CoV, based on genetic and antigenic criteria [8,9]. The SARS-CoV-2 is classified as β-CoV [10] and has received widespread research attention across the world [11–13]. Every day, new genome sequences, as well as primary protein sequences of SARS-CoV-2, are being added to databases, such as the NCBI virus database [14,15] As of this writing, no antiviral drugs with proven efficacy nor vaccines for CoV2 prevention have been reported [16,17], while researchers have yet to attain a complete understanding of the molecular biology of SARS-CoV-2 infection [18,19]As a result, COVID-19 cases increase and

* Corresponding author.
E-mail addresses: ranjeetkumarrout@nitsri.net (R.K. Rout), sksarifhassan@pinglacollege.ac.in (S.S. Hassan), sabha99sheikh@gmail.com (S. Sheikh), saiyedumer@gmail.com (S. Umer), kshirasagar12@gmail.com (K.S. Sahoo), gandomi@uts.edu.au (A.H. Gandomi).

https://doi.org/10.1016/j.compbiomed.2021.105024
Received 22 June 2021; Received in revised form 15 October 2021; Accepted 4 November 2021
Available online 10 November 2021

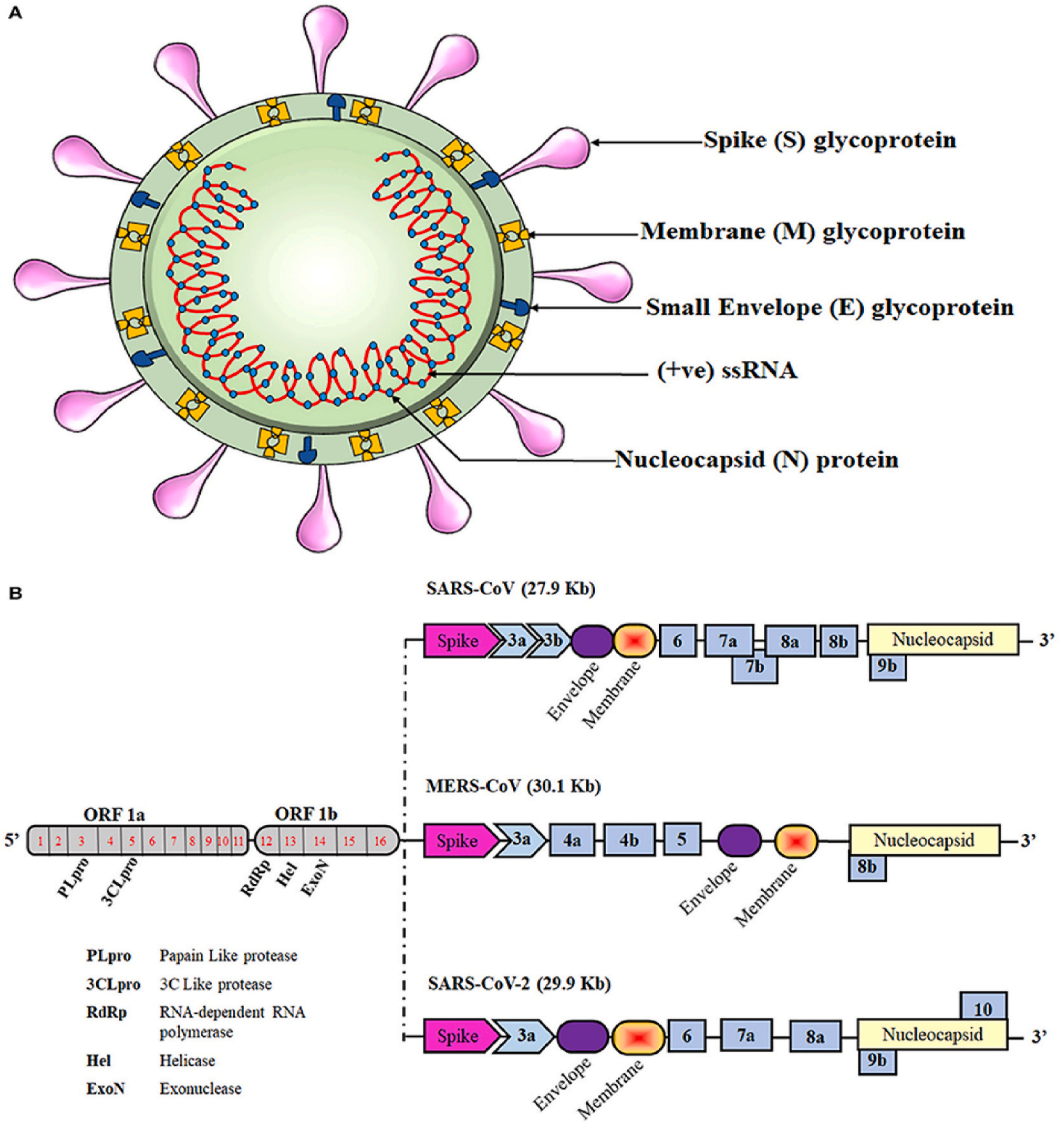

Fig. 1. Schematic representation of the coronavirus structure and genomic comparison of coronaviruses. **(A)** Representation of coronavirus showing different Components of the particle, which is 100–160 nm in diameter. The single-stranded RNA (ssRNA) genome, covered with the envelope and membrane proteins, gains Access into the host cell and hijacks the replication machinery. **(B)** The ssRNA of SARS-cov-2 is about 30 kb and has similarities with the genomes of SARS-CoV and MERS-CoV. Translation of this ssRNA results in the formation of two polyproteins, namely pp1a and pp1ab that are further sliced to generate numerous non-structural Proteins (NSPA). The remaining ORFS encode for various structural and accessory proteins that help in the assembly of the viral particle and evading immune response. This figure is taken from [36].

have reached a global pandemic level, thus urgently requiring in-depth knowledge, infection mechanism, and other aspects of the virus-like forecasting its progression [18,20]. Although various protein-protein interactions (PPIs) of the virus and host are known, its viral infection mechanism is not fully understood [21,22]Therefore, identifying interactions between the SARS-CoV-2 virus proteins and host proteins will largely help to understand this mechanism and further develop treatments and vaccines [23]. As a first step, it is critical to gain clarity of SARS-CoV-2 proteins and PPIs between the virus and host proteins [24]. It is known that the protein fold depends on the number, spatial arrangement, and topological connectivity of secondary structure elements (SSEs) [25], yet the spatial arrangement of secondary structure elements (SSEs) is not well-understood [26]. Because the geometric three-dimensional structure of a protein depends on the spatial arrangement of the SSEs [27,28], both the spatial distribution and presence/absence of different amino acids over a primary protein sequence of SARS-CoV-2 are significant. It is also pertinent to mention that the spatial arrangement uncovers the rules that govern the folding of polypeptide chains, and the primary sequence of a protein reveals the molecular events in evolution [29,30]. Specifically, the alternation and spatial arrangement of amino acids over the primary sequence appear to affect the function and conformability of the protein, respectively [31–33].

In the present study, the spatial composition of 20 amino acids across the primary proteins of SARS-CoV-2 was examined according to the Hurst exponent and Shannon entropy. A frequency analysis of the amino acids was also conducted and further compared to a similar analysis for 89 genomes of SARS-CoV-2 [34]. The usability of Shanon entropy and Hurst exponent for analysis of protein sequences is reported in [29] which is to find out correlation among all these sequences.

1.1. Database and specifications

As of March 24, 2020, there are 944 known primary protein sequences of SARS-CoV-2 in the NCBI Virus Database (*https : //www.ncbi.nlm.nih.gov/labs/virus/vssi/*) [35]. Out of these

Table 1
Lengths of the 105 primary protein sequences.

Seq	Length	Seq	Length	Seq	Length	Seq	Length	Seq	Length	Seq	Length
N99	13	N9	275	N6	638	N13	7091	N33	7096	N53	7096
N80	38	N10	275	N100	932	N44	7095	N34	7096	N54	7096
N81	43	N11	275	N70	1272	N14	7096	N35	7096	N55	7096
N68	61	N101	290	N69	1273	N16	7096	N37	7096	N56	7096
N96	75	N105	298	N71	1273	N17	7096	N38	7096	N57	7096
N97	75	N102	306	N72	1273	N18	7096	N39	7096	N59	7096
N103	83	N104	346	N73	1273	N19	7096	N40	7096	N60	7096
N98	113	N88	419	N74	1273	N20	7096	N41	7096	N61	7096
N82	121	N89	419	N75	1273	N21	7096	N42	7096	N62	7096
N83	121	N90	419	N76	1273	N22	7096	N43	7096	N63	7096
N84	121	N91	419	N77	1273	N23	7096	N45	7096	N64	7096
N85	121	N92	419	N78	1273	N24	7096	N46	7096	N65	7096
N86	121	N93	419	N79	1273	N25	7096	N47	7096	N66	7096
N87	121	N94	419	N4	1945	N27	7096	N48	7096	N67	7096
N2	139	N95	419	N32	4405	N28	7096	N49	7096	N26	7097
N15	180	N7	500	N36	4405	N29	7096	N50	7096		
N3	198	N1	527	N58	4405	N30	7096	N51	7096		
N8	222	N5	601	N12	7088	N31	7096	N52	7096		

Table 2
HE of 105 B_ (1_j) for j = 1, 2…105 corresponding to amino acid A_1 (A).

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N80	0.509	3	N18	0.584	7	N42	0.584	7	N59	0.586	7	N1	0.603	2	N73	0.67	1
N4	0.531	3	N19	0.584	7	N45	0.584	7	N65	0.586	7	N5	0.604	2	N75	0.67	1
N103	0.562	6	N21	0.584	7	N46	0.584	7	N29	0.586	7	N6	0.605	2	N76	0.67	1
N87	0.574	7	N23	0.584	7	N47	0.584	7	N88	0.594	2	N100	0.635	5	N77	0.67	1
N105	0.578	7	N24	0.584	7	N49	0.584	7	N89	0.594	2	N104	0.635	5	N78	0.67	1
N20	0.58	7	N25	0.584	7	N51	0.584	7	N90	0.594	2	N3	0.641	5	N79	0.67	1
N7	0.581	7	N27	0.584	7	N52	0.584	7	N91	0.594	2	N102	0.642	5	N101	0.676	1
N81	0.582	7	N28	0.584	7	N53	0.584	7	N92	0.594	2	N15	0.647	5	N98	0.697	8
N48	0.582	7	N30	0.584	7	N54	0.584	7	N93	0.594	2	N82	0.649	5	N96	0.709	10
N50	0.582	7	N31	0.584	7	N55	0.584	7	N94	0.594	2	N83	0.649	5	N97	0.709	10
N61	0.582	7	N33	0.584	7	N56	0.584	7	N95	0.594	2	N84	0.649	5	N2	0.714	9
N43	0.582	7	N34	0.584	7	N57	0.584	7	N64	0.584	7	N85	0.649	5	N99	0.718	9
N12	0.583	7	N35	0.584	7	N60	0.584	7	N66	0.584	7	N86	0.649	5	N9	0.733	4
N13	0.584	7	N37	0.584	7	N62	0.584	7	N67	0.584	7	N74	0.666	1	N10	0.733	4
N44	0.584	7	N38	0.584	7	N63	0.584	7	N32	0.595	2	N70	0.67	1	N11	0.733	4
N14	0.584	7	N39	0.584	7	N26	0.584	7	N36	0.595	2	N69	0.67	1			
N16	0.584	7	N40	0.584	7	N8	0.585	7	N58	0.597	2	N71	0.67	1			
N17	0.584	7	N41	0.584	7	N22	0.586	7	N68	0.599	2	N72	0.67	1			

Table 3
HE of 105 B_(2_j) for j = 1,2, …105 corresponding to the amino acid A_2 (C).

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N68	*	2	N7	0.567	6	N79	0.6	1	N33	0.6	1	N57	0.6	1	N32	0.6	1
N88	*	2	N15	0.576	6	N70	0.6	1	N34	0.6	1	N59	0.6	1	N36	0.6	1
N89	*	2	N8	0.578	6	N13	0.6	1	N35	0.6	1	N60	0.6	1	N58	0.6	1
N90	*	2	N87	0.583	7	N44	0.6	1	N37	0.6	1	N61	0.6	1	N102	0.6	1
N91	*	2	N98	0.59	7	N3	0.6	1	N38	0.6	1	N62	0.6	1	N4	0.6	8
N92	*	2	N104	0.59	7	N14	0.6	1	N43	0.6	1	N63	0.6	1	N2	0.6	8
N93	*	2	N81	0.594	7	N16	0.6	1	N45	0.6	1	N64	0.6	1	N1	0.7	8
N94	*	2	N80	0.613	1	N17	0.6	1	N46	0.6	1	N65	0.6	1	N6	0.7	8
N95	*	2	N72	0.615	1	N18	0.6	1	N47	0.6	1	N66	0.6	1	N9	0.7	5
N99	*	2	N12	0.617	1	N19	0.6	1	N48	0.6	1	N67	0.6	1	N10	0.7	5
N100	0.5	3	N69	0.617	1	N20	0.6	1	N49	0.6	1	N22	0.6	1	N11	0.7	5
N105	0.5	3	N71	0.617	1	N21	0.6	1	N50	0.6	1	N25	0.6	1	N5	0.7	10
N103	0.5	3	N73	0.617	1	N23	0.6	1	N51	0.6	1	N31	0.6	1	N101	0.7	9
N82	0.5	3	N74	0.617	1	N24	0.6	1	N52	0.6	1	N39	0.6	1	N96	0.7	4
N83	0.5	3	N75	0.617	1	N27	0.6	1	N53	0.6	1	N40	0.6	1	N97	0.7	4
N84	0.5	3	N76	0.617	1	N28	0.6	1	N54	0.6	1	N41	0.6	1			
N85	0.5	3	N77	0.617	1	N29	0.6	1	N55	0.6	1	N42	0.6	1			
N86	0.5	3	N78	0.617	1	N30	0.6	1	N56	0.6	1	N26	0.6	1			

sequences, only 105 sequences are distinct, although these sequence data have been taken from wide ranges of geographic locations over the world. The complete list of 105 distinct sequences, which are denoted *N*1, *N*2, …, *N*105, with their corresponding accessions is provided at the end of the article in Appendix C. These 105 distinct protein sequences were considered in this study. The SARS-CoV and MERS-CoV, the SARS-CoV-2 genome comprises of 12 open reading frames (ORFs) in number. Genes encoding structural proteins such as spike (S), membrane

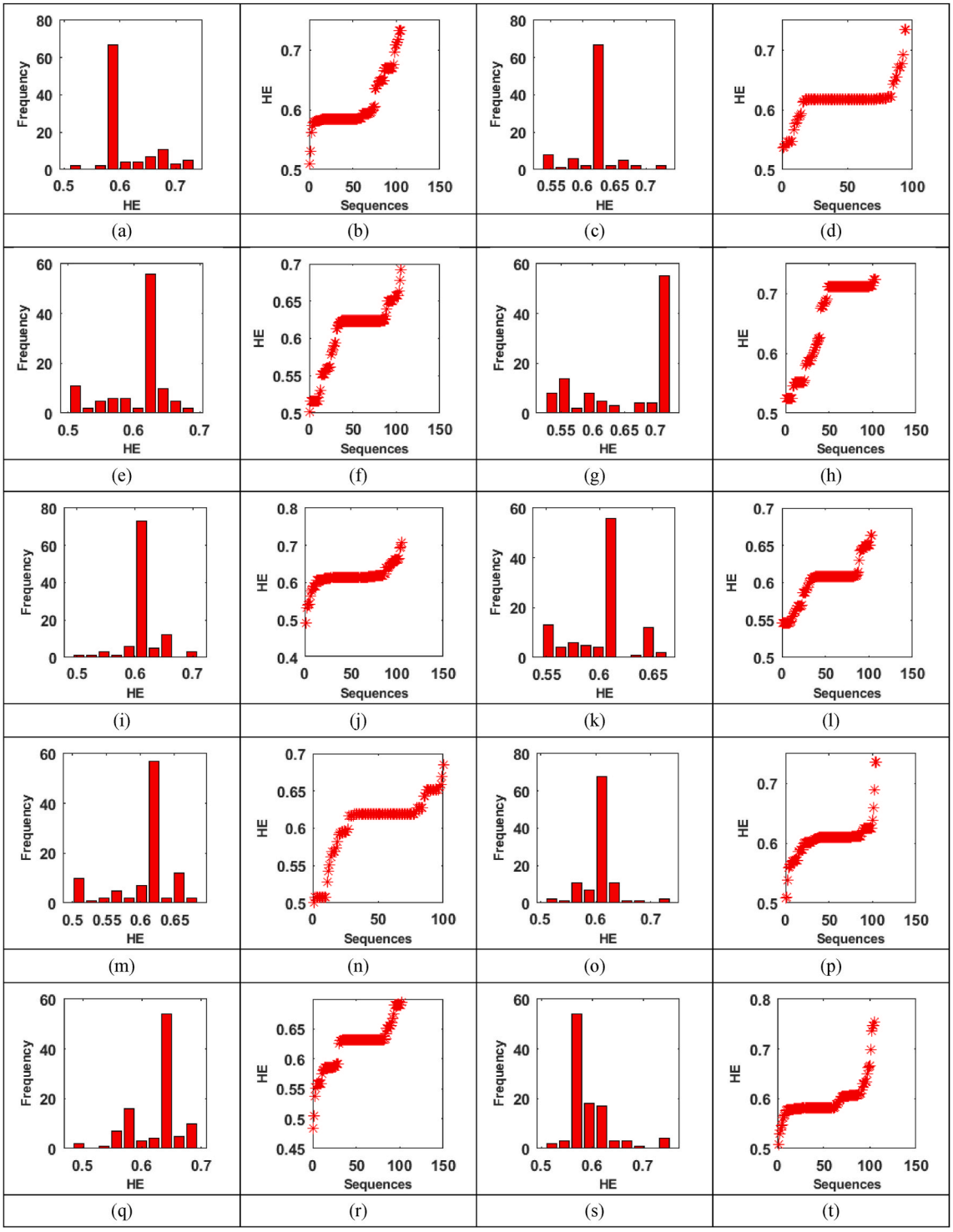

Fig. 2. Shows the Plot of HEs and Histogram all the binary sequences, (a) and (b) for the amino acid $A_1(A)$, (c) and (d) for the amino acid $A_2(C)$, (e) and (f) for the amino acid $A_3(F)$, (g) and (h) for the amino acid $A_4(G)$, (i) and (j) for the amino acid $A_5(H)$, (k) and (l) for the amino acid $A_6(I)$, (m) and (n) for the amino acid $A_7(L)$, (o) and (p) for the amino acid $A_8(M)$, (q) and (r) for the amino acid $A_9(N)$, (s) and (t) for the amino acid $A_{10}(P)$.

(M), envelope (E), and nucleocapsid (N), are present in the remaining one-third of its genome spanning from the 5′ to the 3′ terminal, along with several genes encoding non-structural proteins (NSPs) and accessory proteins scattered in between is shown in Fig. 1 [36].

The 20 amino acids are distinguished below:

- **Essential amino acids:** H, I, K, L, M, F, T, W, and V
- **Conditionally essential:** R, C, Q, G, P, and Y
- **Non-essential:** A, D, N, E, and S

The replication of a virus depends on the availability of amino acids [37]. Because amino acids are required for protein synthesis, they play a crucial role in virus-related infections [38]. The absence of essential amino acids may result in empty virus particles that are free of viral nucleic acids [39]. Arginine (R) is a conditionally essential amino acid that is vital for virus replication and progression of virus infection.

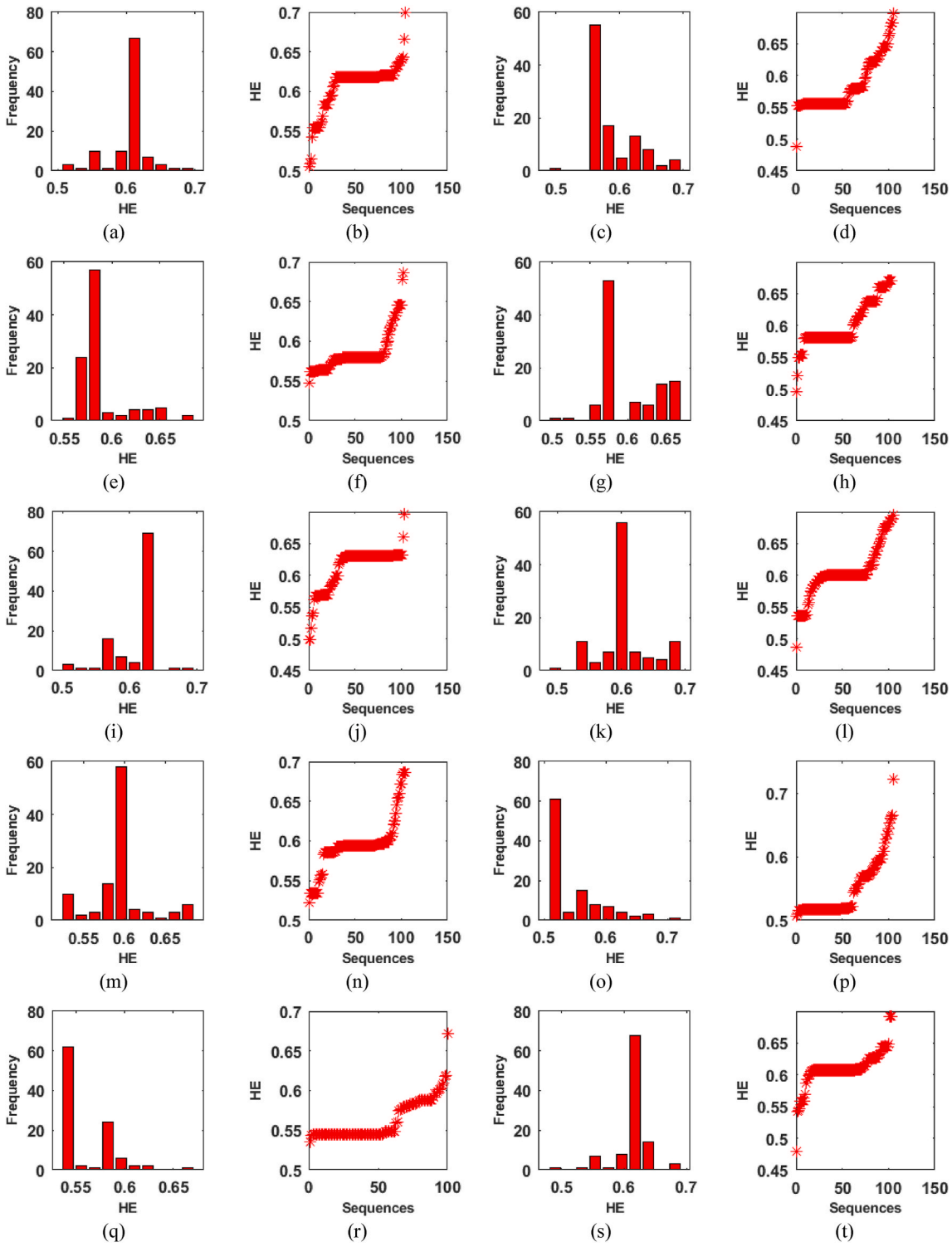

Fig. 3. Shows the Plot of HEs and Histogram all the binary sequences, (a) and (b) for the amino acid $A_{11}(Q)$, (c) and (d) for the amino acid $A_{12}(S)$, (e) and (f) for the amino acid $A_{13}(T)$, (g) and (h) for the amino acid $A_{14}(V)$, (i) and (j) for the amino acid $A_{15}(W)$, (k) and (l) for the amino acid $A_{16}(Y)$, (m) and (n) for the amino acid $A_{17}(D)$, (o) and (p) for the amino acid $A_{18}(E)$, (q) and (r) for the amino acid $A_{19}(K)$, (s) and (t) for the amino acid $A_{20}(R)$.

Carbon is the basic backbone of amino acids, which is attached to a carboxyl group (-COOH), amino group, (-NH_2), hydrogen, and another group of atoms (R) [40]. The R group gives the amino acid its unique characteristics and distinguishes its interaction with other amino acids. Based on the structural and general chemical characteristics, R groups are classified as:

- **Aliphatic:** G, A, V, L, I
- **Hydroxyl:** S, C, T, M
- **Cyclic:** P
- **Aromatic:** F, Y, W
- **Basic:** H, K, R
- **Acidic:** D, Q, Z, N

Table 4
Absence of amino acids on various SARS-CoV-2 proteins.

Amino Acids: Absent	Types	Sequences
C	Hydroxyl, Conditionally Essential	N68, N88, N89, N90, N95, N99
G	Aliphatic, Conditionally Essential	N68, N81
H	Basic, Essential	N3, N80, N97, N98, N99
I	Aliphatic, Essential	N99
M	Hydroxyl, Essential	N99
P	Cyclic, Conditionally Essential	N81, N99, N103
Q	Acidic, Conditionally Essential	N96, N97
T	Hydroxyl, Essential	N99
W	Aromatic, Essential	N80, N87, N96, N97, N99
Y	Aromatic, Conditionally Essential	N99, N103
E	Aromatic, Non Essential	N80, N99
K	Basic, Essential	N80, N81, N99
R	Basic, Conditionally Essential	N81, N99

Herein, we represent the studied amino acids as $A_1, A_2, A_3, \ldots, A_{20}$ corresponding to A, C, F, G, H, I, L, M, N, P, Q, S, T, V, W, Y, D, E, K, and R respectively. Each primary protein sequence was decomposed into 20 different binary sequences of 0 and 1, according to the following rule: Given a primary protein sequence of SARS-CoV-2 for every amino acid $A_i \in \{A, C, F, G, H, I, L, M, N, P, Q, S, T, V, W, Y, D, E, K, R\}$, where i = 1 to 20, put one wherever A_i is present and elsewhere put zero.

Consequently, for every given primary protein sequence *Nj* for all sequences $j = 1, 2, \ldots 105$, there are 20 binary sequences B_{ij} corresponding to the 20 different amino acids A_i, $i = 1, 2, \ldots 20$. The length of these complete 105 primary protein sequences widely varies from 13 to 7097. One complete SARS-CoV-2 protein sequence, N99, has the smallest length of 13, and one protein sequence, N26, has the largest length of 7097. There are 6, 3, 8, 10, 3, and 48 sequences of lengths 121, 275, 419, 1273, 4405, and 7096 respectively, and the other sequences have unique length ranges. Then, all 105 sequences were grouped into six groups, excluding the individual sequences of different unique lengths. The complete list of 105 proteins with their corresponding lengths is given in Table 1 and Accession ID with details of 944 number of sequences are provided in **Appendix C**.

2. Proposed methods

To characterize the amino acid spatial distribution over the primary protein sequences of SARS-CoV-2, the Hurst exponent and Shannon entropy were applied as parameters, and the amino acid density/frequency analysis was performed. Unsupervised machine learning was mostly utilized for analysis of gene and genome sequences and also used for intra-protein analysis. Markov Clustering and Affinity Propagation procedures were compared directly to the method described in [41,42] and K-means clustering techniques in [43]. K-means algorithm is better

Table 5
Correlation matrix of HEs.

	Q	S	T	V	W	Y	D	E	K	R
A	0.280	−0.342	0.271	0.667	0.599	0.306	−0.513	−0.711	−0.607	−0.625
C	−0.434	0.067	0.385	−0.239	−0.101	0.657	0.062	0.223	0.308	0.246
F	0.538	0.061	−0.273	0.051	0.265	−0.104	0.107	0.032	0.230	0.122
G	−0.376	0.407	−0.126	−0.453	−0.439	0.130	0.598	0.780	0.660	0.702
H	0.282	−0.201	−0.134	−0.095	0.112	0.052	−0.241	−0.140	0.025	0.006
I	0.027	−0.374	−0.142	−0.278	−0.292	0.218	−0.066	0.155	0.279	0.339
L	0.103	0.064	0.491	0.355	0.400	0.546	0.038	−0.193	−0.200	−0.107
M	−0.096	0.034	−0.053	−0.333	−0.204	0.443	0.300	0.281	0.389	0.504
N	0.548	0.102	0.082	0.806	0.636	0.116	−0.165	−0.509	−0.613	−0.452
P	0.163	0.385	0.262	0.376	0.240	−0.091	0.103	−0.097	−0.296	−0.088

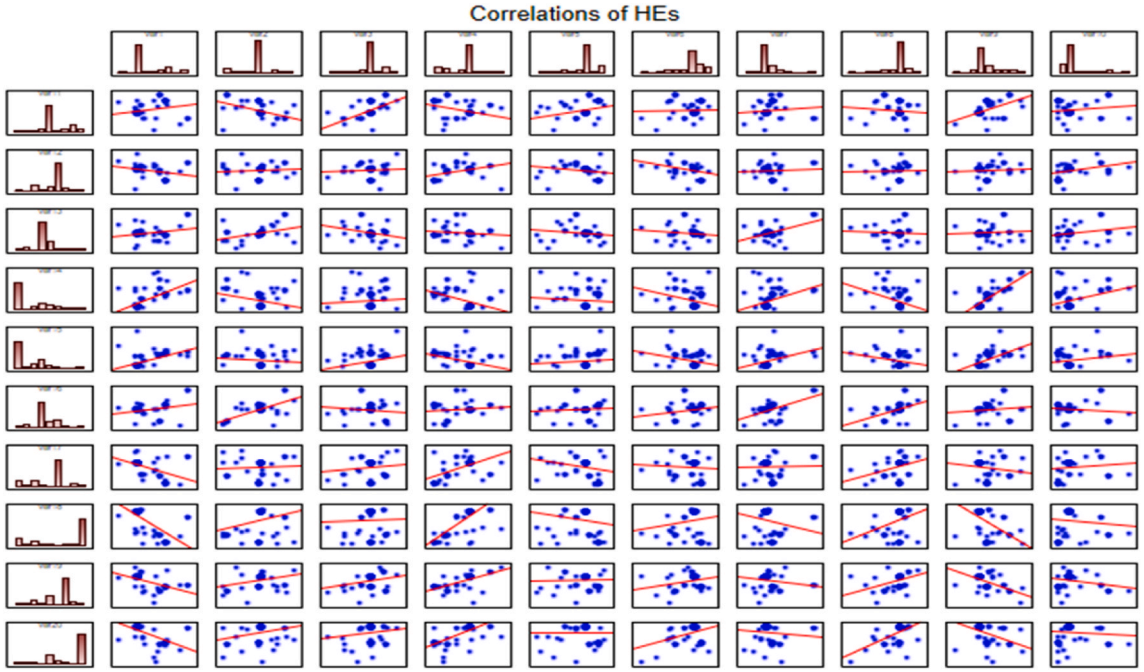

Fig. 4. The correlation plot of HEs of the distribution of amino acids M and Y.

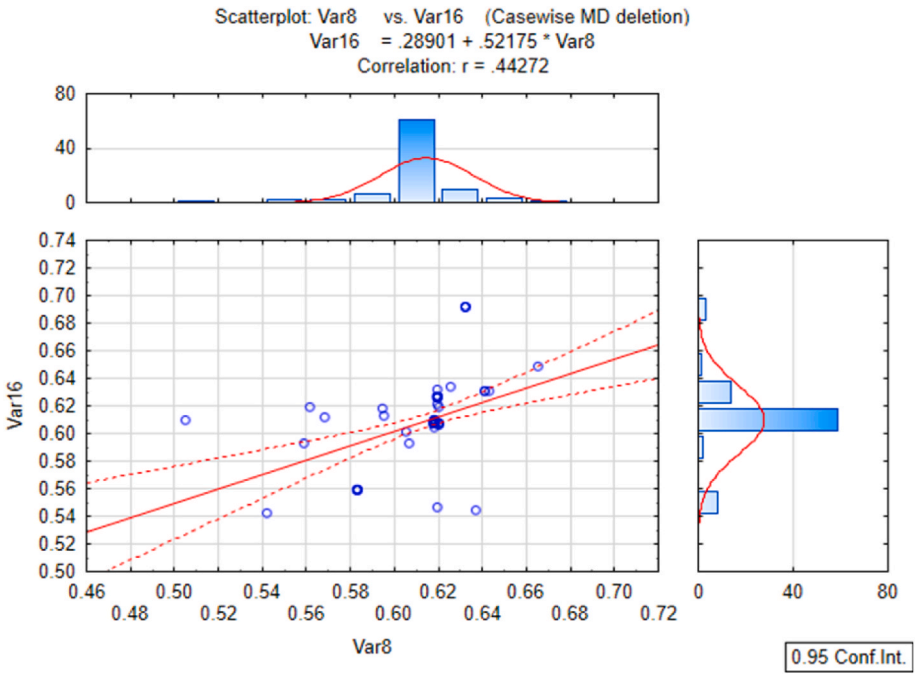

Fig. 5. The correlation plot of HEs of amino acids M and L+.

for analyzing inter and intra class analysis of protein sequences [44]. A recent application of minimum variance cluster analysis for hierarchical agglomerative clustering technique was performed well and discussed in [45] and also identified groups of molecular systems to enhance insight into peptide dynamics. K-mean clustering algorithm is used to develop homogeneous subclasses inside the data. These data points in each cluster are as analogous as possible according to a widely used distance measure viz. Euclidean distance. Based on the performance and applicability one of the most commonly used simple clustering techniques is the K-means clustering [42,46]. In this paper, k-mean clustering algorithm has been used to generate 10 clusters for respective amino acids with the 105 SARS-CoV-2 datasets. The implementation of the spatial feature extraction has been performed using MATLAB-2016a version, on Microsoft 2010 OS. The statistical analysis of these spatial features is also analyzed with the help of STATISTICA 10.0 software in the upcoming sections. The following section briefly describes these methods with reference to similar works [47–49].

2.1. Hurst exponent of binary sequences

The HE lies in the interval $(0,1)$, where HE is strictly less than 0.5 for rough anti-correlated sequences and lies in the ranges 0.5-1 for positively correlated sequences. If HE $= 0.5$, then the sequence depicts its randomness with white noise [50–52]. The HE of a binary sequence s_n is defined as given in Equ. 1 where n is the length of the sequence:

$$\left(\frac{n}{2}\right)^{HE} = \frac{X(n)}{Y(n)} \tag{1}$$

where

$$Y(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - m)}$$

and $X(n) = maxT(i,n) - minT(i,n)$, where

$$T(i) = \sum_{j=1}^{n}(\mathrm{s_i} - \mathrm{t})$$

and

$$t = \sqrt{\frac{1}{n}\sum_{i=1}^{n} s_i}$$

The autocorrelation of the binary representations of each amino acid over the SARS-CoV-2 protein sequences was obtained by measuring the Hurst exponent.

2.2. Shannon entropy

There are two kinds of Shannon entropy that were considered in this present study.

- **Binary Shannon entropy:** The entropy of a Bernoulli process is measured with probability p of the two outcomes $(0/1)$, which is defined in equation (2):

$$SE = -\sum_{i=1}^{2} \mathrm{p_i}\log_2(\mathrm{p_i}) \tag{2}$$

where frequency probabilities of 1's and 0's are respectively $p_1 = \frac{k}{2^l}$ and $p_2 = \frac{l-k}{2^l}$; l is the length of the binary sequence; and k is the number of 1's in the binary sequence of length l [53]. The binary Shannon entropy is a measure of the uncertainty in a binary sequence. When probability $p = 0$, the event is certain to never occur; so there is no uncertainty, and entropy is 0. When probability $p = 1$, the result is certain; thus entropy must be 0. When $p = 0.5$, the uncertainty is at a maximum and consequently, the SE is 1.

- **Amino acid conservation Shannon entropy:** Protein Post Translational Modification (PTM) is an important biological mechanism for expanding the genetic code [54,55]. To find the conservation of amino acids in primary protein sequences, Shannon entropy is deployed. For a given protein sequence, the SE is calculated as follows:

$$SE = -\sum_{i=1}^{20} \mathrm{p_{A_i}}\log_2\left(\mathrm{p_{A_i}}\right) \tag{3}$$

where p_{A_i} represents the occurrence frequency of amino acid A_i in the sequence.

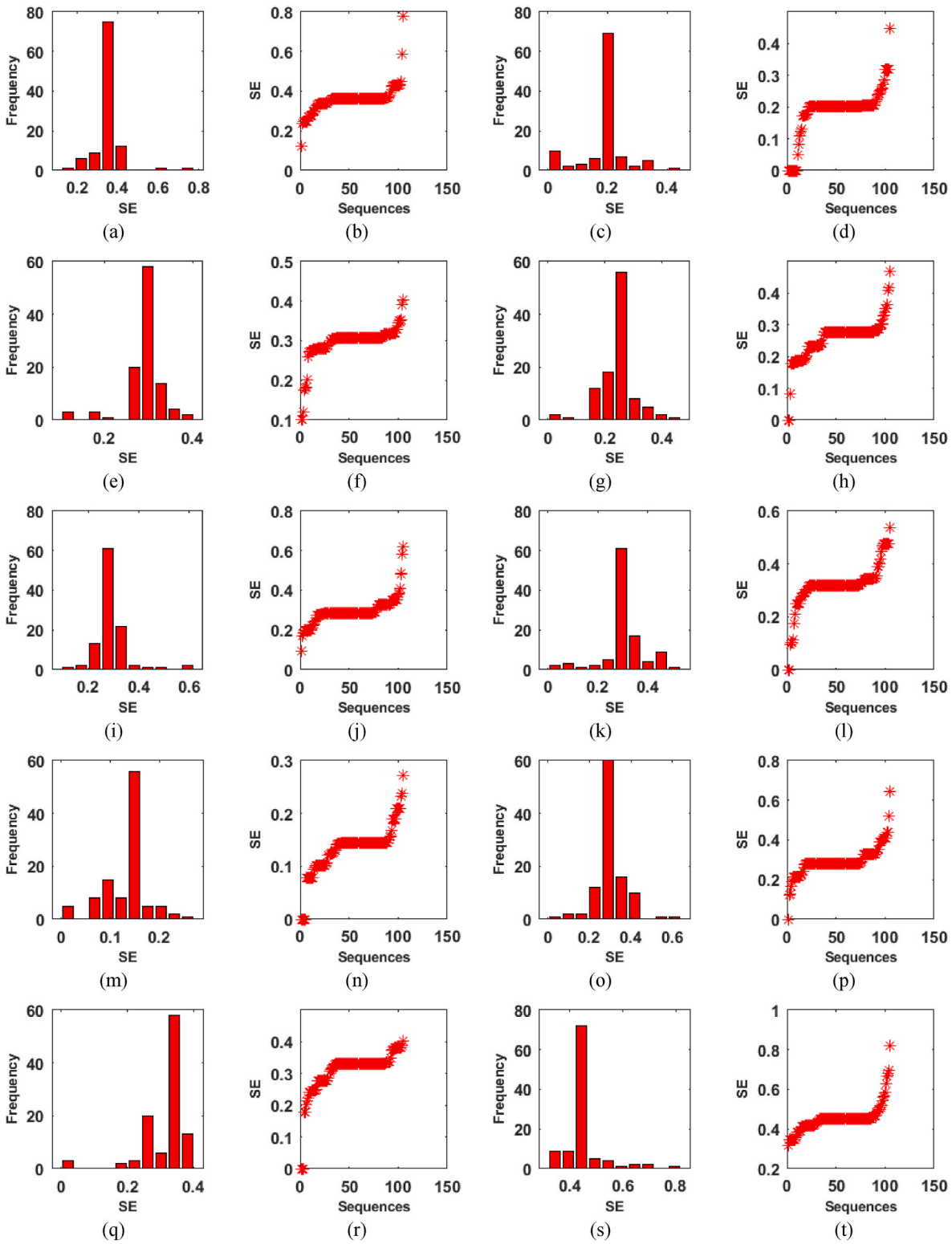

Fig. 6. Shows the Plot of SEs and Histogram all the binary sequences, (a) and (b) for the amino acid $A_1(A)$, (c) and (d) for the amino acid $A_2(C)$, (e) and (f) for the amino acid $A_3(F)$, (g) and (h) for the amino acid $A_4(G)$, (i) and (j) for the amino acid $A_5(H)$, (k) and (l) for the amino acid $A_6(I)$, (m) and (n) for the amino acid $A_7(L)$, (o) and (p) for the amino acid $A_8(M)$, (q) and (r) for the amino acid $A_9(N)$, (s) and (t) for the amino acid $A_{10}(P)$.

2.3. Amino acid density

Over the primary protein sequences of SARS-CoV-2, we aimed to explore the amino acid frequency distributions and corresponding statistical descriptions [11,56]. The density of the amino acids over a primary protein sequence can also be found using the following formula:

$$D(A_i) = \frac{F(A_i)}{L(P)} \times 100\% \tag{4}$$

where A_i is an amino acid present in the primary protein sequence P;

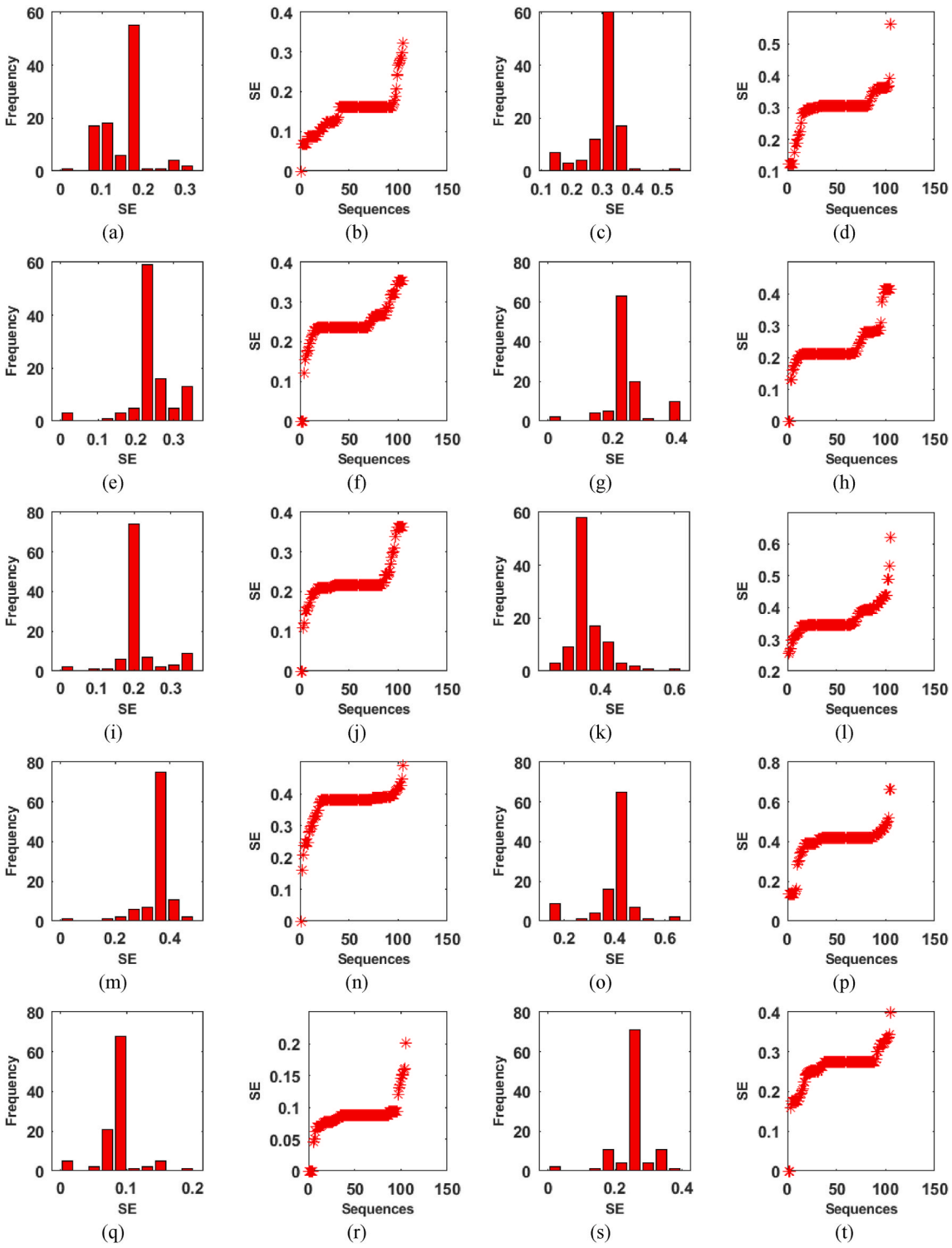

Fig. 7. Shows the Plot of SEs and Histogram all the binary sequences, (a) and (b) for the amino acid $A_{11}(Q)$, (c) and (d) for the amino acid $A_{12}(S)$, (e) and (f) for the amino acid $A_{13}(T)$, (g) and (h) for the amino acid $A_{14}(V)$, (i) and (j) for the amino acid $A_{15}(W)$, (k) and (l) for the amino acid $A_{16}(Y)$, (m) and (n) for the amino acid $A_{17}(D)$, (o) and (p) for the amino acid $A_{18}(E)$, (q) and (r) for the amino acid $A_{19}(K)$, (s) and (t) for the amino acid $A_{20}(R)$.

$L(P)$ is the length of sequence P; and $F(A_i)$ is the frequency of amino acid A_i in sequence P. This amino acid density would clarify the richness of essential amino acids in contrast to others.

3. Results and discussion

Herein, the positive/negative trend of the spatial distribution of the 20 amino acids over the SARS-CoV-2 protein sequences based on the Hurst exponent and Shannon entropy is reported. As mentioned earlier, the Hurst exponent implies the fractality (organized non-linearity) of the spatial representations. Also, the amount of uncertainty in the presence/absence of amino acids over the protein sequences was determined through Shannon entropy measurements, which provide conservation information about the amino acids. Based on the frequency distributions

Table 6
Correlation matrix of SEs of present amino acids over the protein sequences.

r (SE)	Q	S	T	V	W	Y	D	E	K	R
A	0.321	0.290	−0.019	−0.367	−0.143	−0.491	0.192	−0.481	0.073	0.126
C	−0.566	−0.402	0.020	0.621	−0.152	0.530	−0.238	0.237	−0.211	−0.467
F	−0.300	0.037	−0.552	0.267	−0.252	0.181	−0.253	−0.261	−0.840	−0.539
G	0.494	0.007	0.351	−0.454	0.059	−0.230	0.265	−0.212	0.396	0.523
H	−0.279	−0.427	−0.112	0.223	0.363	0.359	0.172	0.565	−0.019	−0.284
I	−0.225	−0.223	−0.108	0.093	0.341	0.436	−0.191	0.309	−0.245	−0.292
L	−0.606	−0.086	−0.234	0.355	0.132	0.016	−0.516	0.184	−0.424	−0.356
M	−0.244	−0.455	0.103	−0.001	0.345	0.022	0.055	0.074	0.098	−0.117
N	−0.039	0.010	0.220	−0.021	−0.227	−0.089	−0.024	−0.424	−0.032	0.116
P	0.411	−0.053	0.472	−0.352	−0.051	0.245	0.097	−0.069	0.451	0.646

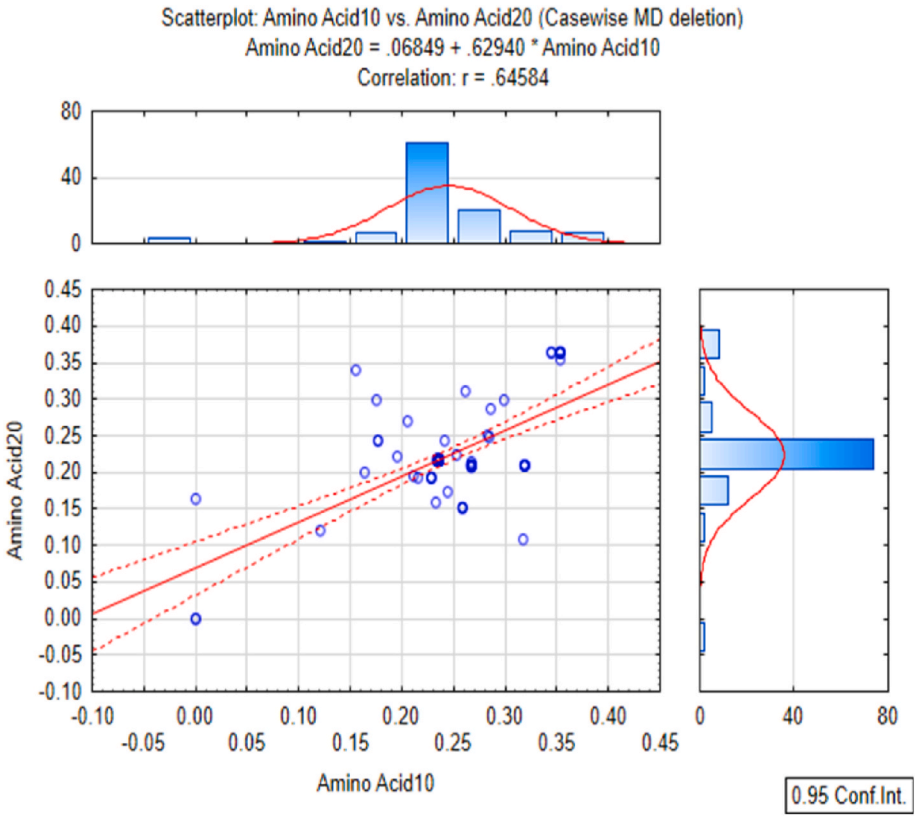

Fig. 8. Correlation plot of SEs of amino acids R and P.

of all amino acids over the SARS-CoV-2 protein sequences, 14 SARS-CoV protein sequences were subsequently compared with 105 SARS-CoV-2 proteins.

3.1. Hurst exponent results

For the amino acid $A_n (n = 1, 2, \ldots, 20)$, the Hurst exponent (HE) was determined for the 105 binary sequences B_{ij}, where i = 1,2 …,20 and $j = 1, 2, \ldots, 105$. Based on the HEs of the binary sequences of all 105 primary protein sequences of SARS-CoV-2, ten clusters (C) are formed for amino acids A_1, A_2, A_3, A_4, A_5, A_6, and A_7; eight clusters for A_{12}, A_{18}, A_{19}, and A_{20}; six clusters for A_{16} and A_{17}; and five clusters for A_8, A_9, A_{10}, A_{11}, A_{13}, A_{14}, and A_{15}. Tables 2 and 3 present the results for Amino Acids A1 and A2, respectively, while the corresponding tables for all other amino acids are given in Appendix A. The HE plot for the binary sequences and the corresponding histogram for all amino acids is shown in Figs. 2 and 3 respectively. It was anticipated that the HE of the binary representations for the ordering of amino acids A_n over all the primary protein sequences reveals the autocorrelation among the amino acids.

The HE of the binary representation of the amino acids forming ten clusters ranges from 0.493 to 0.754 with a standard deviation between 0.0296 and 0.136. For amino acid A_1, cluster 3 consists of two sequences, N4 and N80. For amino acid A_2, clusters 3 and 6 contain 8 and 3 sequences respectively. Both the amino acids A_1 and A_2 have an HE of approximately 0.5, which depicts the random walk/Brownian motion-like character of the ordering of the amino acids over the corresponding protein sequences. For amino acid A_1, 103 primary protein sequences excluding (N4 and N80) and almost all 105 SARS-CoV-2 protein sequences for amino acid A_2 are trending (persistent) sequences. For amino acid A_1, clusters 4, 9 and 10 consist of seven binary representations with an HE of approximately 0.7 and for amino acid A_2, cluster 4 contains two binary representations with an HE of approximately 0.734, which indicates positive autocorrelation (more persistent). The largest cluster i.e cluster 8 contains 65 sequences for the amino acid A_3, cluster 5 contains 71 protein sequences for amino acid A_6, and cluster 8 has 54 protein sequences for amino acid A_5, which all have an HE approximately equal to 0.61 and are positively autocorrelated/persistent. All binary spatial distributions of the 105 proteins for amino acid A_4 have

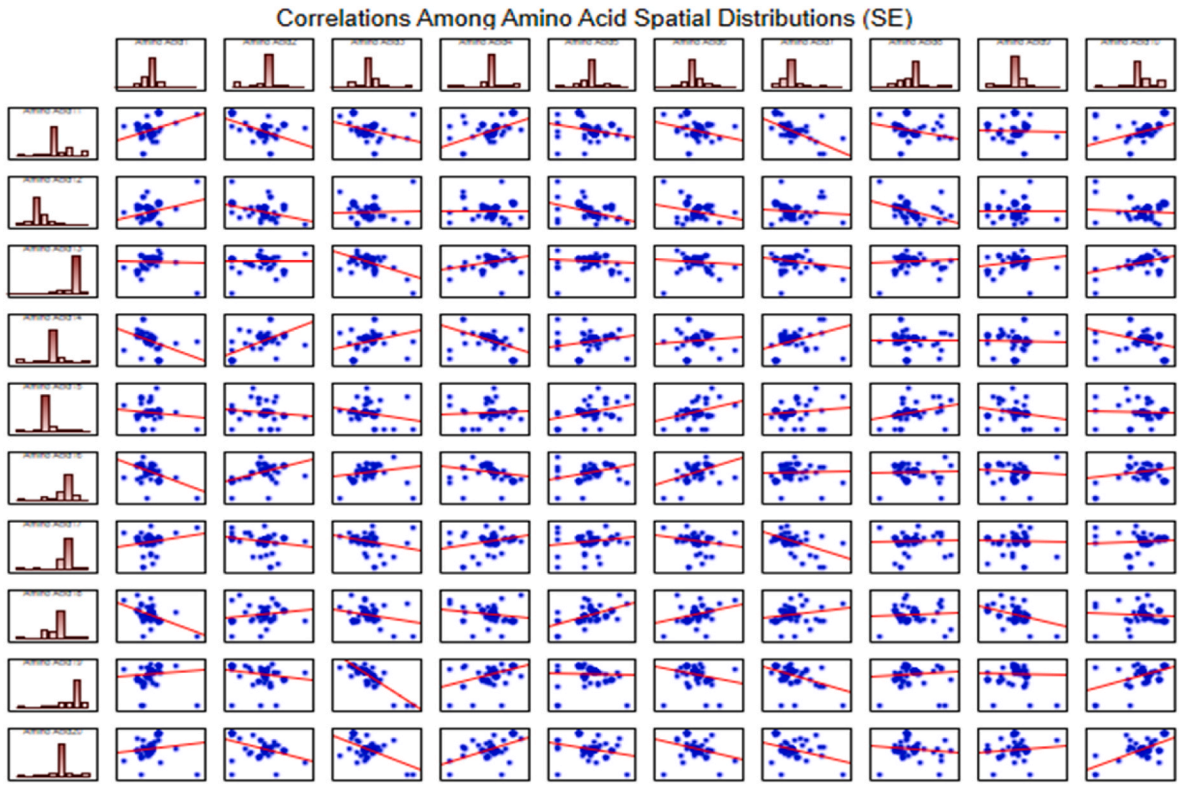

Fig. 9. Correlation plot of SE of the distribution of the amino acids distinct pairwise.

Table 7
Amino acid conservation shannon entropy.

Seq	SE	C	Seq	SE	C	Seq	SE	C	Seq	SE	C	Seq	SE	C	Seq	SE	C
N99	**0.7**	**4**	**N87**	**0.936**	**1**	**N78**	**0.962**	**8**	**N13**	**0.97**	**2**	**N50**	**0.97**	**2**	**N21**	**0.97**	**2**
N81	**0.815**	**6**	**N8**	**0.939**	**3**	**N75**	**0.962**	**8**	**N23**	**0.97**	**2**	**N51**	**0.97**	**2**	**N44**	**0.97**	**2**
N97	**0.846**	**6**	**N101**	**0.942**	**3**	**N74**	**0.962**	**8**	**N37**	**0.97**	**2**	**N25**	**0.97**	**2**	**N24**	**0.97**	**2**
N96	**0.862**	**5**	**N2**	**0.953**	**7**	**N77**	**0.962**	**8**	**N49**	**0.97**	**2**	**N26**	**0.97**	**2**	**N33**	**0.97**	**2**
N103	**0.874**	**5**	**N104**	**0.953**	**7**	**N73**	**0.962**	**8**	**N64**	**0.97**	**2**	**N45**	**0.97**	**2**	**N28**	**0.97**	**2**
N80	**0.879**	**5**	**N9**	**0.955**	**7**	**N72**	**0.962**	**8**	**N66**	**0.97**	**2**	**N46**	**0.97**	**2**	**N27**	**0.97**	**2**
N68	**0.892**	**5**	**N7**	**0.955**	**7**	**N71**	**0.963**	**8**	**N60**	**0.97**	**2**	**N14**	**0.97**	**2**	**N52**	**0.97**	**2**
N15	**0.921**	**9**	**N82**	**0.956**	**7**	**N5**	**0.963**	**8**	**N12**	**0.97**	**2**	**N31**	**0.97**	**2**	**N47**	**0.97**	**2**
N3	**0.925**	**9**	**N6**	**0.956**	**7**	**N76**	**0.963**	**8**	**N65**	**0.97**	**2**	**N39**	**0.97**	**2**	**N62**	**0.97**	**2**
N91	**0.928**	**9**	**N11**	**0.957**	**7**	**N58**	**0.965**	**8**	**N56**	**0.97**	**2**	**N57**	**0.97**	**2**	**N34**	**0.97**	**2**
N94	**0.928**	**9**	**N10**	**0.958**	**7**	**N36**	**0.965**	**8**	**N41**	**0.97**	**2**	**N16**	**0.97**	**2**	**N22**	**0.97**	**2**
N90	**0.928**	**9**	**N84**	**0.958**	**7**	**N32**	**0.965**	**8**	**N55**	**0.97**	**2**	**N29**	**0.97**	**2**	**N67**	**0.97**	**2**
N88	**0.928**	**9**	**N85**	**0.958**	**7**	**N105**	**0.965**	**8**	**N30**	**0.97**	**2**	**N17**	**0.97**	**2**	**N20**	**0.971**	**2**
N98	**0.928**	**9**	**N83**	**0.959**	**7**	**N102**	**0.966**	**8**	**N53**	**0.97**	**2**	**N18**	**0.97**	**2**	**N86**	**0.973**	**2**
N89	**0.928**	**9**	**N4**	**0.961**	**8**	**N100**	**0.97**	**2**	**N59**	**0.97**	**2**	**N19**	**0.97**	**2**	**N1**	**0.982**	**10**
N92	**0.929**	**9**	**N79**	**0.962**	**8**	**N42**	**0.97**	**2**	**N40**	**0.97**	**2**	**N35**	**0.97**	**2**			
N95	**0.931**	**1**	**N70**	**0.962**	**8**	**N61**	**0.97**	**2**	**N43**	**0.97**	**2**	**N38**	**0.97**	**2**			
N93	**0.931**	**1**	**N69**	**0.962**	**8**	**N63**	**0.97**	**2**	**N48**	**0.97**	**2**	**N54**	**0.97**	**2**			

positive autocorrelation and are consequently persistent/trending. One of the essential amino acid A_5(H) is not present in the protein sequences N3, N80, N97, N98 and N99 of the SARS-COV-2. The spatial organization of amino acid H is random (neither trending nor negatively autocorrelated) in the protein sequences N5, N15, N88, N89, N90, N91, N92, N93, N94, and N95, which belong to cluster 2 as shown in Table 6 (Appendix A). Cluster 2 contains ten sequences (N68, N88, N89, N90, N91, N92, N93, N94, N95, and N99) with no HE (*), which indicates that the corresponding binary sequences $B_{2_{68}}$, $B_{2_{88}}$, $B_{2_{89}}$, $B_{2_{90}}$, $B_{2_{91}}$, $B_{2_{92}}$, $B_{2_{93}}$, $B_{2_{94}}$ and $B_{2_{95}}$ are completely free from amino acid A_2(C). Protein sequences N68 and N81 lack amino acid A_4(G) (conditionally essential), as can be seen in Table 5 (Appendix A), while N99 is the only sequence that does not have essential amino acid A_6(I). The spatial distribution of amino acid A_6(I) over the protein sequence N102 is truly random since the HE is 0.509, whereas the other 104 sequences are trending with HEs greater than 0.5. The spatial arrangements of amino acid A_7(L) over these proteins are neither random nor trending as the HE is greater than 0.5 but less than 0.6.

The HE of the binary representation of the amino acids forming eight clusters ranges from 0.483 to 0.724 with a standard deviation between 0.04 and 0.111. The binary representation B_{12_7} of the spatial organization of nonessential amino acid A_{12}(S) over the protein sequence N7 is negatively autocorrelated, whereas the other 104 binary representations corresponding to the protein sequences are positively trending (HE $>$ 0.5). The largest cluster 2, contains 62 sequences for amino acid A_{12}, cluster 1 has 48 sequences for amino acid A_{18}, cluster 3 contains 58 protein sequences for amino acid A_{19}, and cluster 1 consists of 70 protein sequences and sequences N98 and N102 for amino acid A_{20}, which are

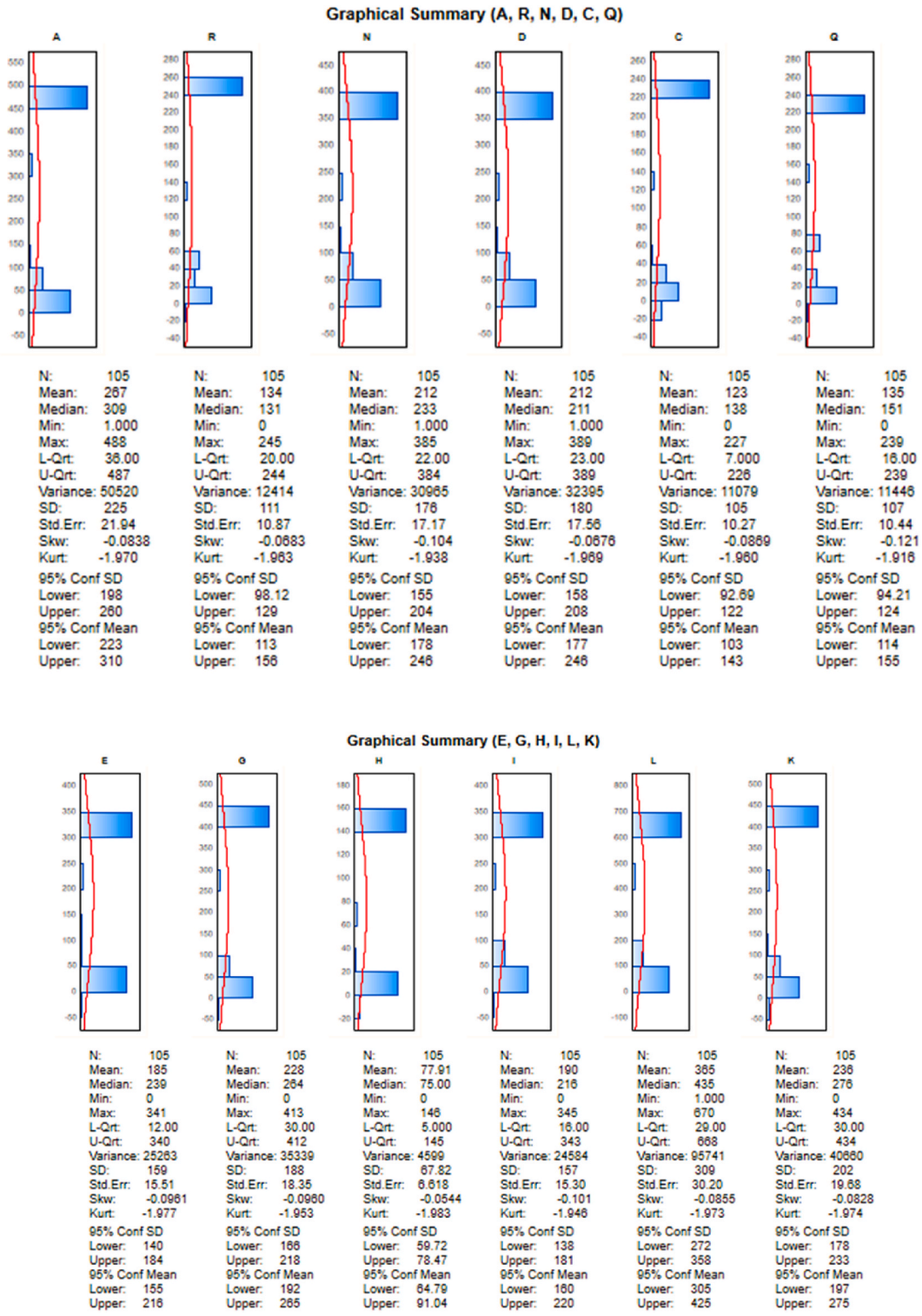

Fig. 10. Comparative statistical details frequencies of the amino acids A, R, N, D, C, Q, E, G, H, I, L, and K over proteins.

positively trending, spatially. It is noteworthy that the spatial representations of amino acid S over the protein sequences N56, N13, N44, and N67 (belonging to cluster 2) all have an HE equal to 0.6, implying positive autocorrelation, while non-essential amino acid A_{18}(E) does not appear in the protein sequences N80 and N99. The protein sequences N80, N81 and N99 are free from amino acid A_{19}(K). The spatial organization of amino acid K over the protein sequence N103 is negatively trending due to an HE of 0.483, which is less than 0.5. The conditionally essential amino acid A_{20}(R) is not at all present in protein sequences N81 and N99, and consequently, the HE is not enumerable.

The HE of the binary representation of the amino acids forming six clusters ranges from 0.479 to 0.692 with a standard deviation between 0.0434 and 0.884. The largest cluster, 1, contains 68 and 60 protein sequences for amino acids A_{16}(Y) and A_{17}(D), respectively, and is spatially spread with a positive trend. The conditional amino acid Y is absent from protein sequences N99 and N103. The spatial distribution of amino acid Y over the only protein N80 belonging to cluster 6 is not trending as its HE is 0.479, which is less than 0.5. The spatial

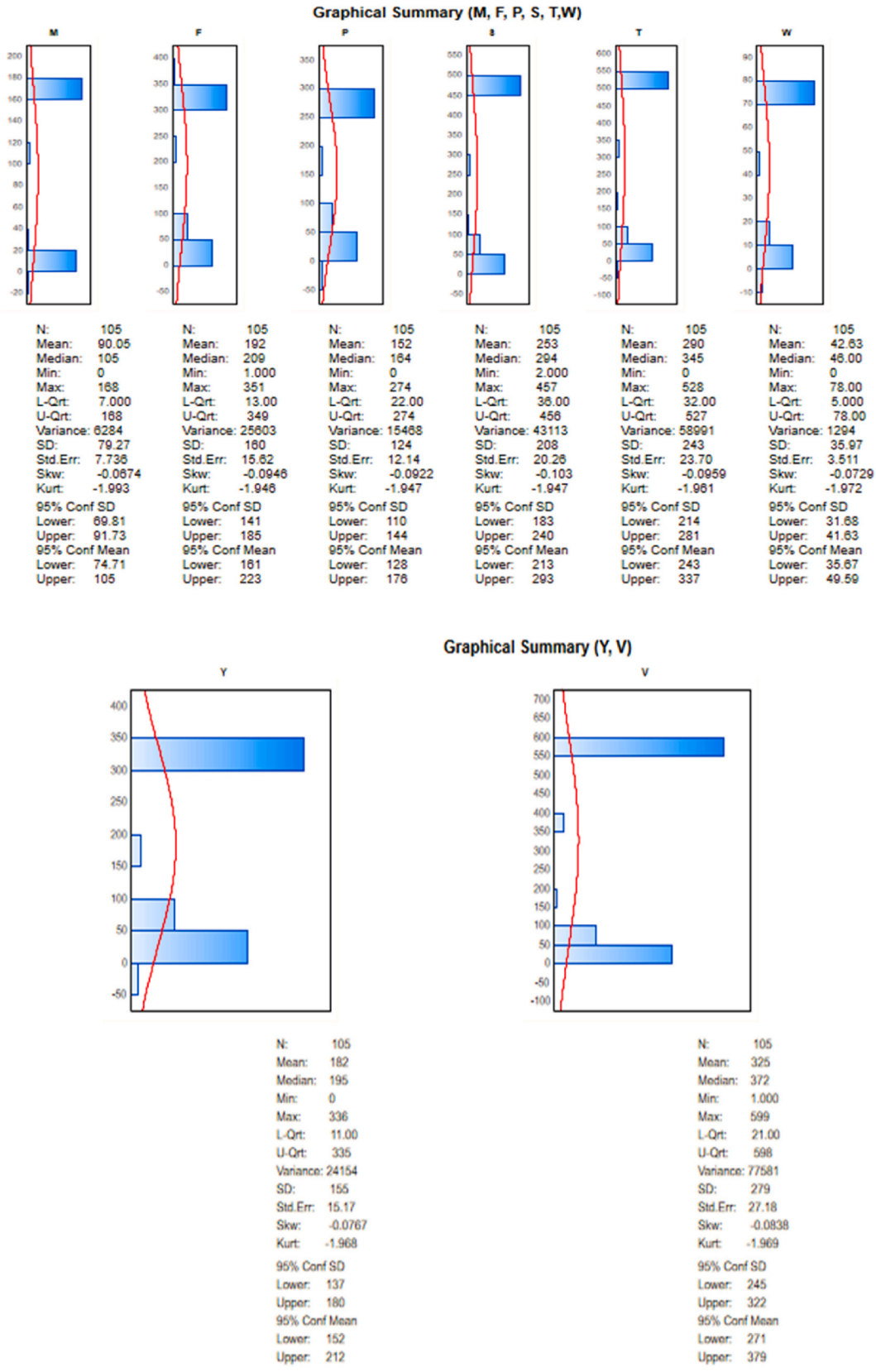

Fig. 11. Statistical comparison between the frequencies of amino acids of M, P, S, T, W, Y and V over the protein sequences.

Table 8
Correlation matrix of the frequencies of amino acids.

	L	K	M	F	P	S	T	W	Y	V
A	0.999	1.000	0.996	0.997	0.998	0.998	0.999	0.997	0.998	0.998
R	0.995	0.997	0.993	0.994	0.997	0.996	0.996	0.995	0.995	0.993
N	0.996	0.996	0.990	0.999	0.998	0.999	0.998	0.993	0.997	0.996
D	0.997	0.998	0.996	0.997	0.998	0.997	0.998	0.996	0.999	0.998
C	0.998	0.996	0.994	0.999	0.995	0.996	0.998	0.993	0.999	0.999
Q	0.989	0.992	0.982	0.993	0.998	0.997	0.994	0.987	0.989	0.988
E	0.999	0.999	0.997	0.995	0.994	0.996	0.998	0.994	0.998	0.998
G	0.997	0.998	0.992	0.997	0.999	0.999	0.999	0.995	0.996	0.995
H	0.996	0.996	0.997	0.994	0.992	0.992	0.995	0.996	0.998	0.997
I	0.998	0.996	0.991	0.999	0.997	0.998	0.998	0.996	0.998	0.998

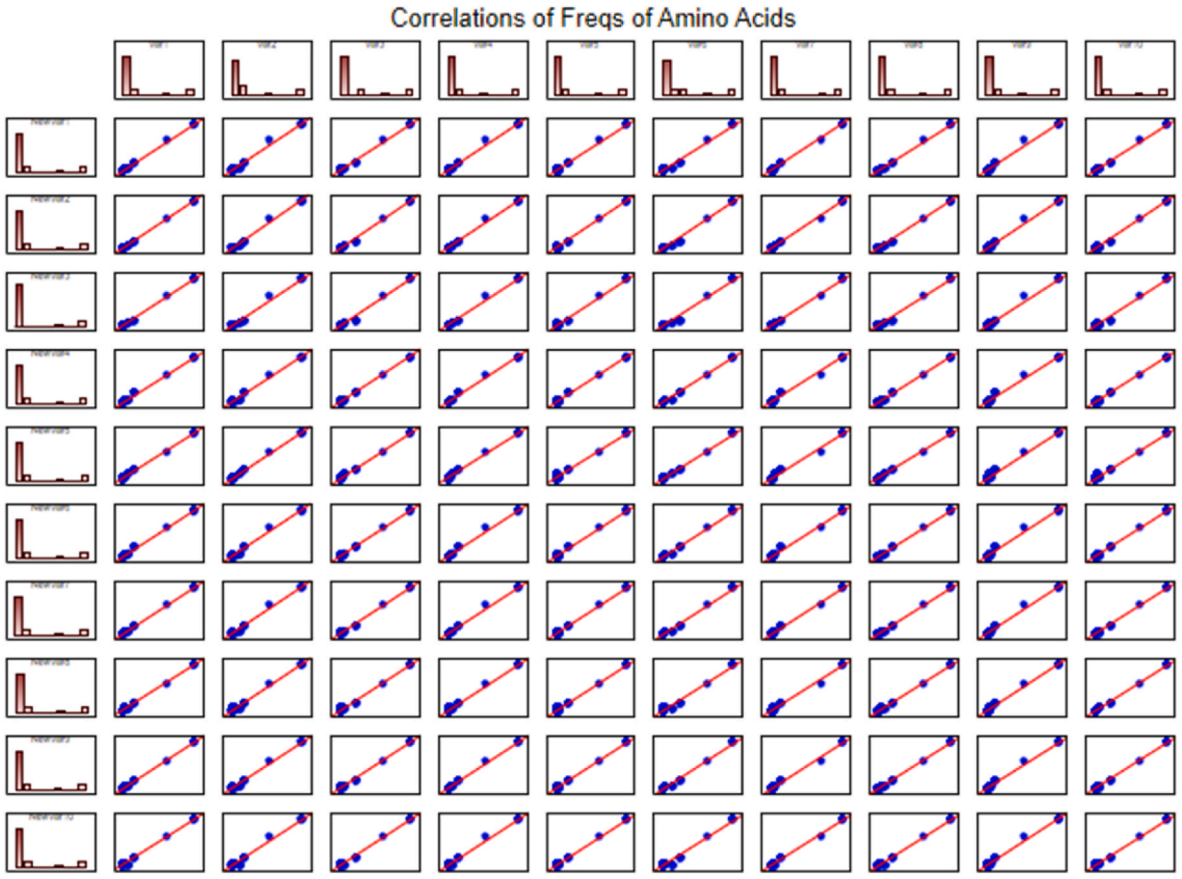

Fig. 12. Correlation graphs for the amino acid frequencies.

distribution B_{17_2} of amino acid D over the protein, sequence N2 is random since its HE is 0.501.

The HE of the binary representation of the amino acids forming five clusters ranges from 0.495 to 0.703 with a standard deviation between 0.0450 and 0.0903. Cluster 3 contains 80 sequences for amino acid A_8(M) over the protein sequences, which has an HE of 0.61 (approx) indicating the trending behavior. The spatial distribution of the amino acid A_9(N) (a non-essential amino acid) over the protein sequence N2 is reverse trending (negatively autocorrelated, HE = 0.488) as observed. In cluster 1 there are 54 sequences having a slow positive trend (HE = 0.55), whereas clusters 3, 4, and 5 contain positively trending spatial representations of amino acid A_9(N) over the protein sequences. Cluster 1 contains 84 B_{10_j} for 74 different protein sequences, where amino acid A_{10}(P) is distributed spatially in a positively trending manner since the HE is approximately 0.56. There is only one binary representation $B_{11_{100}}$ of amino acid A_{11}(Q) over protein sequence N100 that is negatively trending. In cluster 1, protein sequences N96 and N97 are absolutely free from amino acid Q. The spatial distributions of amino acid T over the 76 protein sequences (belonging to cluster 1) are positively trending. The largest cluster 2 contains 61 binary representations B_{14_j} of the spatial distribution of the amino acid A_{14}(V) over the corresponding protein sequences, which are random as the HE turned out to be 0.51(approx). The binary representation B_{14_8} is random as the HE is 0.5 which depicts positive trending behaviour of the binary representation B_{14_8} of the amino acid V over the protein sequence N8. The essential amino acid A_{15}(W) is absent from protein sequences N80, N87, N96 and N99 and consequently, the binary representations $B_{15_{80}}$, $B_{15_{87}}$, $B_{15_{96}}$ and $B_{15_{99}}$ contain only zeros, and HE is in-computable as depicted in table 16 (Appendix A).

3.2. Collective view of HEs

The protein sequences of different lengths, ranging from 13 to 419, are provided below. Table 4 lists the amino acid(s) that are not present

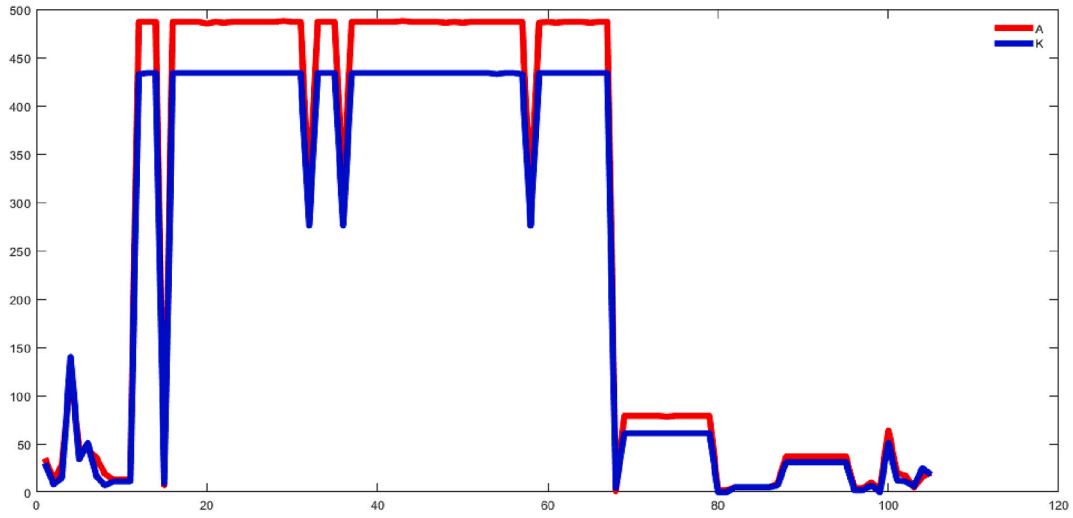

Fig. 13. Frequency plots of amino acids A and K over 105 proteins.

Table 9
List of SARS-CoV proteins with their Accession and length.

Accession ID	Seq	Length
ACU31036	S1	221
ACU31045	S2	63
ACU31034	S3	274
ACU31035	S4	76
ACU31038	S5	44
ACU31041	S6	70
ACU31042	S7	4189
ACU31039	S8	422
ACU31037	S9	122
ACU31033	S10	114
ACU31040	S11	98
ACU31043	S12	121
ACU31044	S13	6880
ACU31032	S14	1241

in the sequences.

The protein sequence N99 of length 13 does not contain some essential, conditionally essential, and non-essential amino acids, including C, H, M, P, T, W, Y, E, K and R. The largest sequences N88, N89, N90, N91, N92, N93, N94, N95 of length 419 do not contain amino acid C. It is noted that amino acid M is present over all the protein sequences, except N99, which has the smallest length of 13. Also, it is has been observed that the essential amino acids L, M, F and V as well as non-essential amino acids A, D, N and S are present in all the protein sequences of SARS-CoV-2. In addition, the six conditionally essential amino acids were not found to be essential for all the proteins of SARS-CoV-2. Proteins that have a length greater than 419 contain all 20 amino acids. It is reported that the presence of amino acid I, G and V is of primordial importance, in this study it has also been found that N99 does not contain I and amino acid G is not present in N68, N81 sequences.

It is also noted that amino acid H is randomly spatially distributed over protein sequences N5, N15, N88, N89, N90, N91, N92, N93, N94 and N95, as observed in the previous subsections. The essential hydroxyl amino acid M is randomly arranged over proteins N80 and N102. Also, amino acid L is distributed over the protein sequence N102 randomly, while only amino acid K is randomly spread over N104. In sequences N98 and N102, amino acid R is distributed with a negative trend ($HE <$ 0.5). Also, the amino acids K, Y, S, Q, N, and F are negatively trending over the protein sequences N103, N80, N7, N100, N2, and N5, respectively. Therefore, amino acids C, G, P, T, W, and E are distributed over all 105 proteins with positive autocorrelation (positively trending).

Here, we explore the correlation (of trending behaviors) of the amino acid distribution over 105 proteins of SARS-CoV-2. The correlation matrix of ten amino acids, A, C, F, G, H, I, L, M, N and P, versus another ten amino acids Q, S, T, V, W, Y, D, E, K and R, is presented below.

The spatial distribution of amino acid A with the same distribution of amino acids Q, T, V, W, and Y is positively correlated based on the HEs shown in Table 5. Likewise, the HE of the spatial distribution of amino acid C is positively correlated with S, T, Y, D, E, K and R. Similarly, the positive correlations of the spatial distributions of amino acids F, G, H, I, L, M, N and P with the spatial distribution of other amino acids are established in the correlation matrix in Table 5. The correlation-based on HEs of the spatial distribution is also demonstrated in the graphs in Fig. 4. It is worth mentioning that the correlation matrix (presented in Table 5) also displays the negative correlations of the spatial distribution of the proteins.

An example of the correlation (correlation coefficient r: 0.443) between the spatial distribution (autocorrelation) of amino acid M and the spatial distribution of amino acid L is given below in Fig. 5.

The following subsection discuss the amount of uncertainty/certainty of the presence of amino acids over the protein sequences.

3.3. Shannon entropy results

For amino acids $A_n (n = 1\ to\ 20)$, the Shannon entropy (SE) was determined for the 105 binary sequences B_{ij} for i = 1 to 20 and $j = 1, 2, \ldots 105$. Results reveal that five clusters (C) formed for amino acids A_1, A_{12}, A_{13}, A_{14}, A_{15}, A_{16}, A_{17}, A_{18}, A_{19}, and A_{20}; six clusters for A_4, A_7,A_8, A_9, A_{10}, and A_{11}; seven clusters for A_2 and A_3; and eight clusters for A_5 and A_6, as presented in **Appendix B**. The SE plot for the binary sequences and the corresponding histogram for amino acid A_1 is given in Figs. 6 and 7(a) and (b) and for the rest of the amino acids it is shown in **Appendix B**. It was anticipated that the SE of the binary representations of the ordering of the amino acids A_n over all the primary protein sequences would reveal the amount of uncertainty of the amino acids.

The SE of the binary representation of the amino acids forming five clusters ranges from 0 to 0.779 with a standard deviation between 0.0448 and 0.0919. The SE of the spatial distribution of amino acid A_1 in protein sequence N68 was determined to be 0.121, which is the lowest amount of uncertainly compared to the SE of other amino acids. In clusters 4 and 1, almost all the protein sequences had an SE less than 0.5, indicating the definite presence and absence of a particular amino acid over the protein sequences. The amount of uncertainly is high for protein sequences N3 and N99 with lengths of 198 and 13, respectively. Amino acids $A_{12}(S)$ and $A_{13}(T)$ are absent from protein sequence N99, with an SE less than 0.5, as shown in Tables 35 and 36, respectively. The

Table 10
HEs and SEs of 14 proteins of the SARS-CoV.

Hurst Exponent (HEs)																			
Seq	A	C	F	G	H	I	L	M	N	P	Q	S	T	V	W	Y	D	E	K	R
S1	0.585	0.571	0.693	0.594	0.621	0.522	0.647	0.593	0.650	0.626	0.638	0.614	0.578	0.599	0.671	0.634	0.685	0.621	0.621	0.619
S2	0.633		0.557		0.598	0.805	0.520	0.620	0.598	0.649	0.500	0.676	0.552	0.596	0.598	0.633	0.662	0.724	0.777	0.663
S3	0.712	0.705	0.540	0.627	0.567	0.506	0.735	0.648	0.602	0.690	0.550	0.588	0.689	0.531	0.595	0.687	0.698	0.627	0.566	0.606
S4	0.709	0.733	0.694	0.625		0.589	0.700	0.593	0.641	0.615		0.647	0.603	0.574		0.610	0.593	0.687	0.651	0.590
S5	0.608	0.586	0.701			0.659	0.676	0.508	0.693	0.608	0.608	0.608	0.608	0.508	0.608	0.608	0.574	0.717	0.608	
S6	0.690	0.728	0.595	0.549	0.646	0.700	0.666	0.595	0.595	0.584	0.655	0.646	0.595	0.683	0.595	0.660		0.601	0.555	0.634
S7	0.605	0.610	0.663	0.623	0.573	0.581	0.589	0.615	0.558	0.590	0.599	0.618	0.576	0.515	0.555	0.635	0.578	0.727	0.631	0.588
S8	0.554		0.604	0.648	0.573	0.600	0.609	0.604	0.614	0.596	0.641	0.695	0.516	0.536	0.549	0.644	0.689	0.548	0.700	0.623
S9	0.622	0.585	0.583	0.645	0.566	0.736	0.631	0.583	0.650	0.660	0.627	0.566	0.622	0.607		0.569	0.629	0.624	0.610	0.649
S10	0.540	0.585	0.521	0.549	0.549	0.680	0.673	0.604	0.585	0.531	0.655	0.654	0.581	0.666		0.511		0.585	0.664	0.527
S11	0.514		0.612	0.632	0.622	0.637	0.644	0.566	0.506	0.589	0.558	0.665	0.627	0.641		0.588	0.553	0.644	0.612	0.665
S12	0.654	0.616	0.511	0.612	0.530	0.475	0.682	0.594	0.643	0.658	0.625	0.488	0.531	0.691	0.583	0.555	0.660	0.583	0.621	0.602
S13	0.601	0.620	0.622	0.589	0.608	0.610	0.614	0.608	0.586	0.582	0.562	0.611	0.584	0.506	0.554	0.615	0.609	0.711	0.607	0.585
S14	0.688	0.619	0.610	0.579	0.635	0.555	0.627	0.615	0.592	0.551	0.649	0.585	0.576	0.535	0.564	0.627	0.598	0.558	0.577	0.584
Shannon Entropy (SEs)																				
Seq	**A**	**C**	**F**	**G**	**H**	**I**	**L**	**M**	**N**	**P**	**Q**	**S**	**T**	**V**	**W**	**Y**	**D**	**E**	**K**	**R**
S1	0.423	0.104	0.285	0.358	0.104	0.407	0.585	0.203	0.323	0.156	0.131	0.323	0.304	0.375	0.203	0.246	0.156	0.225	0.180	0.375
S2	0.203	0.000	0.341	0.000	0.118	0.631	0.503	0.276	0.118	0.276	0.203	0.276	0.276	0.341	0.118	0.203	0.400	0.400	0.341	0.276
S3	0.350	0.172	0.275	0.291	0.208	0.390	0.498	0.152	0.226	0.275	0.243	0.350	0.390	0.428	0.152	0.321	0.275	0.190	0.259	0.110
S4	0.297	0.240	0.297	0.176	0.000	0.240	0.689	0.101	0.350	0.176	0.000	0.443	0.350	0.689	0.000	0.297	0.101	0.240	0.176	0.176
S5	0.156	0.267	0.575	0.000	0.000	0.511	0.811	0.267	0.267	0.156	0.156	0.156	0.156	0.267	0.156	0.156	0.267	0.439	0.156	0.000
S6	0.554	0.316	0.108	0.187	0.255	0.255	0.661	0.108	0.108	0.255	0.371	0.255	0.108	0.469	0.108	0.187	0.000	0.422	0.255	0.255
S7	0.385	0.208	0.260	0.338	0.139	0.276	0.479	0.173	0.276	0.226	0.209	0.364	0.372	0.407	0.081	0.259	0.282	0.305	0.322	0.215
S8	0.404	0.000	0.198	0.490	0.093	0.186	0.334	0.122	0.305	0.379	0.412	0.412	0.387	0.174	0.093	0.174	0.305	0.198	0.370	0.379
S9	0.409	0.283	0.380	0.208	0.247	0.349	0.561	0.069	0.121	0.283	0.208	0.317	0.437	0.283	0.000	0.247	0.121	0.349	0.283	0.283
S10	0.219	0.073	0.176	0.127	0.297	0.367	0.670	0.333	0.073	0.127	0.398	0.485	0.608	0.333	0.000	0.176	0.000	0.073	0.398	0.127
S11	0.408	0.000	0.144	0.144	0.144	0.291	0.507	0.197	0.197	0.408	0.332	0.371	0.443	0.507	0.000	0.082	0.332	0.291	0.246	0.291
S12	0.121	0.382	0.285	0.285	0.248	0.382	0.439	0.210	0.210	0.351	0.248	0.319	0.121	0.411	0.069	0.351	0.285	0.382	0.210	0.248
S13	0.377	0.209	0.271	0.328	0.155	0.275	0.457	0.169	0.291	0.233	0.208	0.349	0.362	0.412	0.086	0.273	0.307	0.281	0.321	0.229
S14	0.360	0.197	0.316	0.320	0.084	0.336	0.399	0.124	0.336	0.255	0.290	0.404	0.396	0.387	0.068	0.262	0.306	0.229	0.283	0.213

Table 11
Correlation matrix of the HEs (Pairwise).

r	Q	S	T	V	W	Y	D	E	K	R
A	−0.141	−0.385	0.514	0.004	−0.244	0.283	0.260	−0.592	−0.845	−0.092
C	−0.706	−0.101	0.814	−0.288	−0.316	0.535	0.307	−0.046	−0.752	−0.077
F	0.263	0.807	−0.159	−0.431	0.305	0.253	−0.346	0.437	0.417	0.018
G	−0.503	−0.159	0.409	0.083	−0.052	0.257	0.285	0.313	0.091	0.264
H	0.298	0.680	0.037	−0.525	0.181	0.335	−0.261	−0.058	−0.239	−0.171
I	−0.256	0.723	−0.039	−0.806	−0.497	0.190	−0.758	0.696	0.120	−0.694
L	−0.302	−0.457	0.575	0.371	0.342	0.243	0.865	−0.497	−0.558	0.581
M	−0.654	0.264	0.908	−0.583	−0.286	0.796	0.138	0.096	−0.758	−0.144
N	0.408	−0.513	−0.229	0.824	0.774	−0.367	0.761	−0.614	0.118	0.798
P	−0.392	−0.418	0.456	0.457	0.412	0.153	0.854	−0.164	−0.143	0.712

Table 12
Correlation matrix of the SEs of the spatial distributions of amino acids.

r	Q	S	T	V	W	Y	D	E	K	R
A	0.245	0.109	0.119	0.123	0.032	−0.190	−0.273	−0.094	0.108	0.500
C	−0.311	−0.355	−0.553	0.237	−0.009	0.572	−0.318	0.464	−0.492	−0.350
F	−0.589	−0.554	−0.270	−0.287	0.297	0.164	0.281	0.399	−0.428	−0.490
G	0.203	0.425	0.152	−0.150	0.140	0.379	0.100	−0.426	0.198	0.526
H	0.566	0.151	0.173	−0.128	−0.247	0.108	−0.391	−0.124	0.430	0.117
I	−0.253	−0.536	−0.233	−0.262	0.407	−0.029	0.298	0.351	−0.133	−0.294
L	−0.363	−0.363	−0.190	0.229	0.030	−0.245	−0.594	0.214	−0.474	−0.591
M	0.123	−0.101	0.079	−0.237	0.162	−0.308	0.112	−0.089	0.168	−0.345
N	−0.468	0.145	−0.080	0.188	0.268	0.309	0.342	−0.176	−0.391	0.060
P	0.438	0.025	−0.079	−0.103	−0.210	−0.134	0.518	0.199	0.162	0.500

amino acid A_{14}(V) is present over all 105 proteins, and hence, none of the binary representations has SE = 0. For the amino acid V, the SE of N74 and N77 is 0.391, which implies the presence of this amino acid over the proteins has good certainty, and N96 and N97 have the maximum uncertainty of SE = 0.665. Cluster 1 contains five protein sequences, in which amino acid A_{15} is absent, and hence, SE = 0. Also, SE = 0 for the binary spatial representations of N99 and N103 for amino acid A_{16}, N80 and N99 (belonging to cluster 2) for amino acid A_{18}, N80, N81 and N99 for amino acid A_{19}, and N81 and N99 amino acid A_{20} due to the absence of these amino acids. It is pertinent to note that amino acids $A_{17}$and A_{18} are present over all 105 proteins with certainty ($HEs < 0.5$). Most of the proteins in the largest cluster 2 including other clusters contain amino acid A_{15} that is spatially distributed with certainty.

The SE of the binary representation of the amino acids forming six clusters ranges from 0 to 0.644 with a standard deviation between 0.0749 and 0.852. Amino acid $A_4(G)$ is absent from the primary protein sequences N68 and N81, and consequently, SE = 0 implies no uncertainty. Similarly, SE = 0 for the binary spatial representations of protein sequence N99 for amino acid $A_8(M)$, sequences N81, N99 and N103 for amino acid A_{10}(P), and sequences N96 and N97 for amino acid A_{11}(Q). Amino acid $A_7(L)$ is spread spatially with certainty over the proteins N2 (length of 138) and N89, N90, N91, N92, N93, N94 and N95 (lengths of 419) in cluster 3. Clusters 1 and 5 for amino acid A_7 and cluster 1 for amino acids A_8 and A_{10} contain the majority of the protein sequences, where the presence of these amino acids is spread over the proteins with almost certainty. Comparatively, clusters 2 and 6 contain five protein sequences, where the absence of the amino acid A_7 is spread with almost certainty. Cluster 3 contains one protein sequence N80 where the spatial distribution $B_{9_{80}}$ has SE = 0.562, which indicates that the absence of amino acid A_9 over the protein is without uncertainty.

The SE of the binary representation of the amino acids forming seven clusters each ranges from 0 to 0.619 with a standard deviation between 0.0667 and 0.0765. It was found that SE = 0 for the spatial distribution of amino acid A_2 in the protein sequences N68, N88, N89, N90, N91, N92, N93, N94, N95 and N99, which indicates the amount of uncertainty is zero. In other words, the absolute absence of amino acid $A_2(C)$ over these proteins and the spatial presence of amino acid C over the protein sequences of other clusters have low uncertainty (high certainty). The SE is greater than 0.5 for the binary representations of amino acid A_3 over the proteins N81 and N99, and consequently, the amount of uncertainty is lowering. In other clusters containing the other protein sequences, the spatial presence of amino acid A_3 over the protein sequences has low uncertainty (high certainty).

The SE of the binary representation of the amino acids forming eight clusters ranges from 0 to 0.644 with a standard deviation between 0.0459 and 0.0749. Because amino acid $A_5(H)$ is absent from proteins N3, N80, N97, N98 N99 and amino acid $A_6(I)$ is absent from N99 (smallest length of 13), SE = 0 for the amino acids, implying there is no uncertainty. In addition, SE = 0.078 for the spatial representation of the presence and absence of amino acid A_5 over the proteins N88, N89, N90, N91, N92, N94 and N95 (lengths of 419) belonging to cluster 4); hence, the spatial distribution is more certain/orderly. All the clusters except cluster 6 contain only protein sequences over which amino acid $A_6(I)$ is spatially distributed with certainty, whereas cluster 6 contains two sequences N81 (length of 43) and N68 (length of 61), where the absence of the amino acid dominates the presence with certainty.

3.4. Collective view of SE

It is pertinent to mention that SE = 0 for the binary representations B_{i_j} of amino acid A_i that is absent from protein sequence *Nj*, which has been demonstrated in this study. It was also observed that maximum SE was obtained for the spatial distribution of amino acids over lengthy sequences, such as N99, N80, etc. Interestingly, for some given amino acid A_i, the same SE was obtained for some spatial distributions B_{i_j} of some protein sequences *Nj*, irrespective of their lengths, for many values of *j*. This essentially suggest that the probability of the presence of amino acid A_i over these protein sequences is the same.

Further, we explored the correlation in the amount of uncertainty between the spatial distributions of the 20 amino acids over the proteins of SARS-CoV-2. Table 6 presents the correlation matrix of ten amino acids (A, C, F, G, H, I, L, M, N and P) versus another ten amino acids (Q, S, T, V, W, Y, D, E, K and R).

Based on the SEs, the spatial distribution of amino acid A was found to be positively correlated with the distributions of amino acids Q, S, D,

K and R, as shown in Table 6. Likewise, the spatial distribution of amino acid C is positively correlated with amino acids T, V, Y and E. Similarly, the positive correlations between the spatial distributions of amino acids F, G, H, I, L, M, N and P and the other amino acids are established in the correlation matrix in Table 6, which also shows negative correlations.

The correlation-based on SEs of the spatial distribution is also demonstrated in the graphs in Fig. 9. An example of the correlation-based on SEs (the correlation coefficient r: 0.646) of the spatial distribution (autocorrelation) of amino acid R with the spatial distribution of amino acid P is given in Fig. 8.

3.5. Amino acid conservation shannon entropy

For each of the 105 protein sequences, the amino acid conservation information was determined through HE measurement, as described earlier. Based on the Shannon entropy (*SE_T*2) for each sequence, the clusters (C) were formed, and the respective SE plots and histograms for the 105 protein sequences are provided in Table 7.

It can be observed that the Shannon entropy of amino acid conservation along the protein sequences of SARS-CoV-2 ranges from 0.7 to 0.982. Since the SE is close to 1, meaning uncertainty is at a maximum, all amino acids must be uniformly distributed over the protein sequences. More than 50% of the proteins sequences (54) belonging to cluster 2 of SARS-CoV-2 have SE = 0.970, which further implies that the amino acids are almost uniformly spread over the sequences. Subsequently, the frequency analysis of the amino acids over the proteins is given in the following subsection.

3.6. Frequency distribution of amino acids over the SARS-CoV-2 proteins

In this section, the frequencies of the amino acids in the 105 SARS-CoV-2 protein sequences are statistically compared, as shown in Figs. 10 and 11.

A correlation matrix between the frequency distribution of amino acids over the 105 SARS-CoV-2 protein sequences is provided in Table 8, and the respective correlation graphs are illustrated in Fig. 12.

It can be observed that the correlation coefficient is very close to 1, which indicates significant correlations between the frequencies of each amino acid over the proteins. For instance, the correlation coefficient between the frequency distributions of amino acids A (Aliphatic) and K (Basic) is 1, as illustrated in Fig. 13, means strong correlation.

Overall, it is observed that protein sequences of the same length have very similar frequency distributions of the twenty amino acids.

4. Spatial organization of proteins of SARS-COV

In 2003, the SARS coronavirus (SARS-CoV) had caused an epidemic in China including the other 22 countries [56,57]. There are 14 protein sequences available in the NCBI database (taxid: 722424). The list of proteins (S1, S2, … S11) with their accessions are given here in Table 9.

It is noted that the protein with the accession ACU31032 (S14) is a spike protein of length 1241 as mentioned in the NCBI database. The spike protein (S-protein) is a large type I transmembrane protein of length not exceeding 1400 amino acids. The spike protein has an important function in the case of SARS-CoV [58,59]. Among all other proteins of SARS-CoV, spike protein is the main antigenic component that is responsible for inducing host immune responses, neutralizing antibodies, and/or protective immunity against virus infection [60]. We, therefore illuminate here the spatial representations of the amino acids over the spike protein including the other 13 proteins as mentioned in Table 10. The HE, SE, and frequency distributions are given in the following and compared with the SARS-CoV2 proteins.

It is observed that the spatial representations of the presence of all the amino acids over the spike protein S14 follow the positive autocorrelation (positively trending) as well as with the least amount of uncertainty of presence of the amino acids. It seems that the presence of all the amino acids is necessary to make a spike protein. It is worth mentioning that yet there are no identified spike proteins in the domain of 105 distinct proteins of SARS-CoV2. The amino acids A, F, I, L, M, N, P, S, T, V, Y, E, and K are all present over all these 14 proteins unlike in the case of SARS-CoV2 proteins as mentioned in subsection 3.21. It is worth mentioning that all the spatial distributions corresponding to different amino acids over the 14 proteins are positively autocorrelated with $HE \geq 0.5$, except for the spatial distribution of the amino acid I and S over the protein S12 which is a hypothetical protein. It is noted that the HE is kept blank for the cases where the spatial distribution of an amino acid is completely a sequence of zeros i,e. absence of the amino acid over the protein. Below in Table 11, we derive the correlation coefficients of the HEs of the spatial representations of the amino acids over the 14 SARS-CoV proteins.

It is observed from Table 11 that the correlation coefficient (r) is 0.908 for the HEs of spatial representations of the amino acid M and T over all the 14 SARS-CoV proteins. Noted that overall the proteins, the presence of amino acid M and T are ensured. There is also another positive correlation that exists as can be seen in Table 11. It is noted that the SE is turned out to be zero for the cases where the spatial distribution corresponding to an amino acid that is absent over a protein. The spatial distribution of amino acids over the proteins of SARS-CoV is all without much uncertainty except for three cases where the SEs are greater than 0.5 where the absence of amino acids dominates in terms of certainty. The correlation coefficients of the SEs of the spatial distributions of the amino acids over the 14 SARS-CoV proteins are given in Table 12. It is observed that the correlations among the SEs of the spatial distributions of the amino acids over the proteins are not significantly up as tabulated in Table 12. The highest positive correlation based on SEs of the spatial distributions of the amino acid C with that of Y is turned up as 0.572.

5. Discussion

Previous reports state that the genomes of SARS-CoV and SARS-CoV-2 exhibit similar protein sequences. However, we found that the spatial arrangement of amino acids over the studied protein sequences is certainly different, contributing to differences between proteins. This study reveals the hidden spatial arrangement of the amino acids of SARS-CoV-2 and SARS-CoV1. Specifically, the spatial arrangements of amino acids over the primary protein sequences of SARS-CoV-2 were examined according to the autocorrelation via Hurst exponent measurements and the presence/absence of the amino acids via Shannon entropy. Also, the frequency distribution of amino acids was analyzed to categorize the protein sequences. Based on a comparative analysis, the spatial distribution of 14 protein sequences of SARS-CoV demonstrated a significant difference from those of SARS-CoV-2. Conclusions are based on the calculated HE and SE, which provide information about the spatial arrangement of the amino acids over the primary protein sequences of SARS-CoV-2 as well as SARS-CoV. The obtained results, present in section 4, reveal the differences between the proteins of the two types of CoV. We firmly believe that our findings on the spatial distribution of the present/absent amino acids over the proteins enable a better understanding of the PPIs of SARS-CoV-2. For instance, the spatial arrangements reveal the similarities and dissimilarities among the important structural proteins E, M, N and S, which further helps to establish a more complete evolutionary tree among the other CoV strains. Despite our promising results, the present study is limited, as it did not consider the three-dimensional spatial structure of associate proteins, such as RdRp, E, M, N and S.

Authors' contribution

SH had initiated the problem for the study, and RKR and SH executed the results from the data. SH, RKR, SS, SU, KSS, and AHG analyzed and interpreted the results. SH was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.105024.

References

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, B. Cao, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (2020) 497–506, https://doi.org/10.1016/S0140-6736(20)30183-5.

[2] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G.F. Gao, W. Tan, A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. 382 (2020) 727–733, https://doi.org/10.1056/nejmoa2001017.

[3] W. Hua, L. Xiaofeng, B. Zhenqiang, R. Jun, W. Ban, L. Liming, Consideration on the strategies during epidemic stage changing from emergency response to continuous prevention and control, Chin. J. Endemiol. 41 (2020) 297–300, https://doi.org/10.3760/cma.j.issn.0254-6450.2020.02.003.

[4] S.S. Hassan, A. Moitra, R.K. Rout, P.P. Choudhury, P. Pramanik, S.S. Jana, On spatial molecular arrangements of SARS-CoV2 genomes of Indian patients, BioRxiv (2020), https://doi.org/10.1101/2020.05.01.071985.

[5] R.K. Rout, S.S. Hassan, Spatial Distribution of Amino Acids of the SARS-CoV2 Proteins, 2020, https://doi.org/10.20944/PREPRINTS202004.0034.V2.

[6] S. Perlman, Another decade, another coronavirus, N. Engl. J. Med. 382 (2020) 760–762, https://doi.org/10.1056/nejme2001126.

[7] C. Wang, P.W. Horby, F.G. Hayden, G.F. Gao, A novel coronavirus outbreak of global health concern, Lancet 395 (2020) 470–473, https://doi.org/10.1016/S0140-6736(20)30185-9.

[8] C. Ceraolo, F.M. Giorgi, Genomic variance of the 2019-nCoV coronavirus, J. Med. Virol. 92 (2020) 522–528, https://doi.org/10.1002/jmv.25700.

[9] Z.W. Ye, S. Yuan, K.S. Yuen, S.Y. Fung, C.P. Chan, D.Y. Jin, Zoonotic origins of human coronaviruses, Int. J. Biol. Sci. 16 (2020) 1686–1697, https://doi.org/10.7150/ijbs.45472.

[10] A.E. Gorbalenya, S.C. Baker, R.S. Baric, R.J. de Groot, C. Drosten, A.A. Gulyaeva, B. L. Haagmans, C. Lauber, A.M. Leontovich, B.W. Neuman, D. Penzar, S. Perlman, L. L.M. Poon, D.V. Samborskiy, I.A. Sidorov, I. Sola, J. Ziebuhr, The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, Nat. Microbiol. 5 (2020) 536–544, https://doi.org/10.1038/s41564-020-0695-z.

[11] Y.Z. Zhang, E.C. Holmes, A genomic perspective on the origin and emergence of SARS-CoV-2, Cell 181 (2020) 223–227, https://doi.org/10.1016/j.cell.2020.03.035.

[12] K.G. Andersen, A. Rambaut, W.I. Lipkin, E.C. Holmes, R.F. Garry, The proximal origin of SARS-CoV-2, Nat. Med. 26 (2020) 450–452, https://doi.org/10.1038/s41591-020-0820-9.

[13] X. Tang, C. Wu, X. Li, Y. Song, X. Yao, X. Wu, Y. Duan, H. Zhang, Y. Wang, Z. Qian, J. Cui, J. Lu, On the origin and continuing evolution of SARS-CoV-2, Natl. Sci. Rev. 7 (2020) 1012–1023, https://doi.org/10.1093/nsr/nwaa036.

[14] E.W. Sayers, J. Beck, J.R. Brister, E.E. Bolton, K. Canese, D.C. Comeau, K. Funk, A. Ketter, S. Kim, A. Kimchi, P.A. Kitts, A. Kuznetsov, S. Lathrop, Z. Lu, K. McGarvey, T.L. Madden, T.D. Murphy, N. O'Leary, L. Phan, V.A. Schneider, F. Thibaud-Nissen, B.W. Trawick, K.D. Pruitt, J. Ostell, Database resources of the national center for biotechnology information, Nucleic Acids Res. 48 (2020) D9–D16, https://doi.org/10.1093/nar/gkz899.

[15] E.L. Hatcher, S.A. Zhdanov, Y. Bao, O. Blinkova, E.P. Nawrocki, Y. Ostapchuck, A. A. Schaffer, J. Rodney Brister, Virus Variation Resource-improved response to emergent viral outbreaks, Nucleic Acids Res. 45 (2017) D482–D490, https://doi.org/10.1093/nar/gkw1065.

[16] C. Liu, Q. Zhou, Y. Li, L.V. Garner, S.P. Watkins, L.J. Carter, J. Smoot, A.C. Gregg, A.D. Daniels, S. Jervey, D. Albaiu, Research and development on therapeutic agents and vaccines for COVID-19 and related human coronavirus diseases, ACS Cent. Sci. 6 (2020) 315–331, https://doi.org/10.1021/acscentsci.0c00272.

[17] K. Dhama, K. Sharun, R. Tiwari, M. Dadar, Y.S. Malik, K.P. Singh, W. Chaicumpa, COVID-19, an emerging coronavirus infection: advances and prospects in designing and developing vaccines, immunotherapeutics, and therapeutics, Hum. Vaccines Immunother. 16 (2020) 1232–1238, https://doi.org/10.1080/21645515.2020.1735227.

[18] M.A. Alves, G.Z. Castro, B.A.S. Oliveira, L.A. Ferreira, J.A. Ramírez, R. Silva, F. G. Guimarães, Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs, Comput, Biol. Med. 132 (2021), https://doi.org/10.1016/j.compbiomed.2021.104335.

[19] J. Liu, X. Zheng, Q. Tong, W. Li, B. Wang, K. Sutter, M. Trilling, M. Lu, U. Dittmer, D. Yang, Overlapping and discrete aspects of the pathology and pathogenesis of the emerging human pathogenic coronaviruses SARS-CoV, MERS-CoV, and 2019-nCoV, J. Med. Virol. 92 (2020) 491–494, https://doi.org/10.1002/jmv.25709.

[20] Y.C. Wang, C.H. Cheng, A multiple combined method for rebalancing medical data with class imbalances, Comput. Biol. Med. 134 (2021) 104527, https://doi.org/10.1016/j.compbiomed.2021.104527.

[21] N. Goodacre, P. Devkota, E. Bae, S. Wuchty, P. Uetz, Protein-protein interactions of human viruses, Semin. Cell Dev. Biol. 99 (2020) 31–39, https://doi.org/10.1016/j.semcdb.2018.07.018.

[22] X. Yang, S. Yang, Q. Li, S. Wuchty, Z. Zhang, Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method, Comput. Struct. Biotechnol. J. 18 (2020) 153–161, https://doi.org/10.1016/j.csbj.2019.12.005.

[23] S. Srinivasan, H. Cui, Z. Gao, M. Liu, S. Lu, W. Mkandawire, O. Narykov, M. Sun, D. Korkin, Structural genomics of SARS-COV-2 indicates evolutionary conserved functional regions of viral proteins, Viruses 12 (2020), https://doi.org/10.3390/v12040360.

[24] D.E. Gordon, G.M. Jang, M. Bouhaddou, J. Xu, K. Obernier, M.J. O'Meara, J. Z. Guo, D.L. Swaney, T.A. Tummino, R. Huettenhain, R.M. Kaake, A.L. Richards, B. Tutuncuoglu, H. Foussard, J. Batra, K. Haas, M. Modak, M. Kim, P. Haas, B. J. Polacco, et al., A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing, BioRxiv, 2020, https://doi.org/10.1101/2020.03.22.002386.

[25] R. Kolodny, D. Petrey, B. Honig, Protein structure comparison: implications for the nature of "fold space", and structure and function prediction, Curr. Opin. Struct. Biol. 16 (2006) 393–398, https://doi.org/10.1016/j.sbi.2006.04.007.

[26] E. Krissinel, K. Henrick, Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, Acta Crystallogr. Sect. D Biol. Crystallogr. 60 (2004) 2256–2268, https://doi.org/10.1107/S0907444904026460.

[27] R.K. Rout, S. Ghosh, P.P. Choudhury, Classification of mer proteins in a quantitative manner, Int. J. Comput. Appl. Eng. Sci. II (2014), https://doi.org/10.1371/journal.pone.0031635.

[28] X. Pennec, N. Ayache, A geometric algorithm to find small but highly similar 3D substructures in proteins, Bioinformatics 14 (1998) 516–522, https://doi.org/10.1093/bioinformatics/14.6.516.

[29] R. Kumar, H. Sarif, SindhwaniSanchit, P. Mohan, UmerSaiyed, Intelligent classification and analysis of essential genes using quantitative methods, ACM Trans. Multimed Comput. Commun. Appl 16 (2020), https://doi.org/10.1145/3343856.

[30] Y.S. Chiang, T.I. Gelfand, A.E. Kister, I.M. Gelfand, New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage, Proteins Struct. Funct. Genet. 68 (2007) 915–921, https://doi.org/10.1002/prot.21473.

[31] M. Michael Gromiha, P.K. Ponnuswamy, Hydrophobie distribution and spatial arrangement of amino acid residues in membrane proteins, Int. J. Pept. Protein Res. 48 (1996) 452–460, https://doi.org/10.1111/j.1399-3011.1996.tb00863.x.

[32] T. Kollár, I. Pálinkó, Z. Kónya, I. Kiricsi, Intercalating amino acid guests into montmorillonite host, in: J. Mol. Struct., Elsevier, 2003, pp. 335–340, https://doi.org/10.1016/S0022-2860(03)00109-1.

[33] R.K. Rout, S. Umer, S. Sheikh, S. Sindhwani, S. Pati, EightyDVec: a method for protein sequence similarity analysis using physicochemical properties of amino acids, https://doi.org/10.1080/21681163.2021.1956369.

[34] S.S. Hassan, R.K. Rout, V. Sharma, A Quantitative Genomic View of the Coronaviruses: SARS-COV2, 2020, pp. 1–33, https://doi.org/10.20944/PREPRINTS202003.0344.V1.

[35] J.R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, NCBI viral Genomes resource, Nucleic Acids Res. 43 (2015) D571–D577, https://doi.org/10.1093/nar/gku1207.

[36] V.K. Shah, P. Firmal, A. Alam, D. Ganguly, S. Chattopadhyay, Overview of immune response during SARS-CoV-2 infection: lessons from the past, Front. Immunol. 11 (2020), https://doi.org/10.3389/FIMMU.2020.01949.

[37] K.L. Schierhorn, F. Jolmes, J. Bespalowa, S. Saenger, C. Peteranderl, J. Dzieciolowski, M. Mielke, M. Budt, S. Pleschka, A. Herrmann, S. Herold, T. Wolff, Influenza A virus virulence depends on two amino acids in the N-terminal domain of its NS1 protein to facilitate inhibition of the RNA-dependent protein kinase PKR, J. Virol. 91 (2017), https://doi.org/10.1128/jvi.00198-17.

[38] U.A. Ashfaq, T. Javed, S. Rehman, Z. Nawaz, S. Riazuddin, An overview of HCV molecular biology, replication and immune responses, Virol. J. 8 (2011), https://doi.org/10.1186/1743-422X-8-161.

[39] W. Luytjes, L.S. Sturman, P.J. Bredenbee, J. Charite, B.A.M. van der Zeijst, M. C. Horzinek, W.J.M. Spaan, Primary structure of the glycoprotein E2 of coronavirus MHV-A59 and identification of the trypsin cleavage site, Virology 161 (1987) 479–487, https://doi.org/10.1016/0042-6822(87)90142-5.

[40] R.K. Rout, P.P. Choudhury, S.P. Maity, B.S.D. Sagar, S.S. Hassan, Fractal and mathematical morphology in intricate comparison between tertiary protein structures, https://doi.org/10.1080/21681163.2016.1214850.

[41] J. Vlasblom, S.J. Wodak, Markov clustering versus affinity propagation for the partitioning of protein interaction graphs, BMC Bioinf. 10 (2009) 1–14, https://doi.org/10.1186/1471-2105-10-99, 2009 101.

[42] T. Bhadra, S. Bandyopadhyay, Unsupervised feature selection using an improved version of Differential Evolution, Expert Syst. Appl. 42 (2015) 4042–4053, https://doi.org/10.1016/J.ESWA.2014.12.010.

[43] A. Likas, N. Vlassis, J. Verbeek, J.J. Verbeek, The global k-means clustering algorithm, (n.d.). ïhttps://doi.org/10.1016/S0031-3203(02)00060-2ï.

[44] G. Bouvier, N. Desdouits, M. Ferber, A. Blondel, M. Nilges, An automatic tool to analyze and cluster macromolecular conformations based on self-organizing maps, Bioinformatics 31 (2015) 1490–1492, https://doi.org/10.1093/BIOINFORMATICS/BTU849.

[45] V.C. De Souza, L. Goliatt, P.V.Z.C. Goliatt, Clustering algorithms applied on analysis of protein molecular dynamics, IEEE Lat. Am. Conf. Comput. Intell. LA-CCI 2017 - Proc. 2017-Novem (2017) 1–6, https://doi.org/10.1109/LA-CCI.2017.8285695, 2018.

[46] J.L. Phillips, M.E. Colvin, S. Newsam, Validating clustering of molecular dynamics simulations using polymer models, BMC Bioinf. 12 (2011) 1–23, https://doi.org/10.1186/1471-2105-12-445, 2011 121.

[47] J.P. Banerjee, J.K. Das, P. Pal Choudhury, S. Mukherjee, S.S. Hassan, P. Basu, The variations of human miRNAs and Ising like base pairing models, BioRxiv (2018) 319301, https://doi.org/10.1101/319301.
[48] J.K. Das, P.P. Choudhury, N. Chaturvedi, M. Tayyab, S.S. Hassan, Ranking and clustering of Drosophila olfactory receptors using mathematical morphology, Genomics 111 (2019) 549–559, https://doi.org/10.1016/j.ygeno.2018.03.010.
[49] J.K. Das, P.P. Choudhury, A. Chaudhuri, S.S. Hassan, P. Basu, Analysis of purines and pyrimidines distribution over miRNAs of human, Gorilla, chimpanzee, Mouse and Rat, Sci. Rep. 8 (2018) 1–19, https://doi.org/10.1038/s41598-018-28289-x.
[50] M. Kale, F. Butar Butar, Fractal analysis of time series and distribution properties of Hurst exponent, J. Math. Sci. Math. Educ. 5 (n.d.).
[51] J. Mielniczuk, P. Wojdyłło, Estimation of Hurst exponent revisited, Comput. Stat. Data Anal. 51 (2007) 4510–4525, https://doi.org/10.1016/j.csda.2006.07.033.
[52] M.J. Sánchez-Granero, M. Fernández-Martínez, J.E. Trinidad-Segovia, Introducing fractal dimension algorithms to calculate the Hurst exponent of financial time series, Eur. Phys. J. B. 85 (2012) 1–13, https://doi.org/10.1140/epjb/e2012-20803-2.
[53] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theor. 37 (1991) 145–151, https://doi.org/10.1109/18.61115.
[54] B.J. Strait, T.G. Dewey, The Shannon information entropy of protein sequences, Biophys. J. 71 (1996) 148–155, https://doi.org/10.1016/S0006-3495(96)79210-X.
[55] L.R. Nemzer, Shannon information entropy in the canonical genetic code, J. Theor. Biol. 415 (2017) 158–170, https://doi.org/10.1016/j.jtbi.2016.12.010.
[56] X. Xiao, S. Chakraborti, A.S. Dimitrov, K. Gramatikoff, D.S. Dimitrov, The SARS-CoV S glycoprotein: expression and functional characterization, Biochem. Biophys. Res. Commun. 312 (2003) 1159–1164, https://doi.org/10.1016/j.bbrc.2003.11.054.
[57] G. Simmons, J.D. Reeves, A.J. Rennekamp, S.M. Amberg, A.J. Piefer, P. Bates, Characterization of severe acute respiratory syndrome-associated coronavirus (SARS-CoV) spike glycoprotein-mediated viral entry, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 4240–4245, https://doi.org/10.1073/pnas.0306446101.
[58] L. Du, Y. He, Y. Zhou, S. Liu, B.J. Zheng, S. Jiang, The spike protein of SARS-CoV - a target for vaccine and therapeutic development, Nat. Rev. Microbiol. 7 (2009) 226–236, https://doi.org/10.1038/nrmicro2090.
[59] Y. He, Y. Zhou, S. Liu, Z. Kou, W. Li, M. Farzan, S. Jiang, Receptor-binding domain of SARS-CoV spike protein induces highly potent neutralizing antibodies: implication for developing subunit vaccine, Biochem. Biophys. Res. Commun. 324 (2004) 773–781, https://doi.org/10.1016/j.bbrc.2004.09.106.
[60] J. Cinatl, B. Morgenstern, G. Bauer, P. Chandra, H. Rabenau, H.W. Doerr, Treatment of SARS with human interferons, Lancet 362 (2003) 293–294, https://doi.org/10.1016/S0140-6736(03)13973-6.