Spatial Distribution of Amino Acids of the SARS-CoV2 Proteins

Ranjeet Kumar Rout[a], Sk. Sarif Hassan[b,*]

[a] *Department of Computer Science & Engineering, National Institute of Technology, Hazratbal, Srinagar, India*
[b] *Department of Mathematics, Pingla Thana Mahavidyalaya, Paschim Medinipur-721140, India*

Abstract

The world is now undergoing through a global emergency due to COVID-19 which needs immediate remedies in order to strengthen the healthcare facility to save the nations. Looking towards to the remedies, research on different aspects including the genomic and proteomic level characterizations of the SARS-CoV2 are necessarily important. In this present study, the spatial representation/composition of twenty amino acids across the primary protein sequences of SARS-CoV2 have been looked into through different parameters viz. Shannon entropy, Hurst exponent in order to fetch the autocorrelation and amount of information over the spatial representations. Also frequency distribution of each of the amino acids over the protein sequences have been chalked out.

Keywords: Shannon entropy, Hurst exponent, Amino acid, Frequency distribution, & SARS-CoV2.

---

1. Introduction

Global emergency due to the COVID-19 is making life hard throughout the globe. The largest genomes (of size approximately 30 kb) for RNA viruses so far is known as the SARS-CoV2 [1, 2, 3, 4, 5]. CoVs are classified into three

---

[*] Corresponding author

Email addresses: `ranjeetkumarrout@nitsri.net` (Ranjeet Kumar Rout), `sarimif@gmail.com` (Sk. Sarif Hassan)

different classes such as α-CoVs, β-CoVs and γ-CoVs based on the genetic and antigenic criteria[6, 7]. The SARS-COV2 is classified into the β-CoV group [8]. Quite a good number of untiring research activities across the world have been carried out [9, 10, 11]. Everyday, new genome sequences of SARS-CoV2 are being included in the databases viz. NCBI virus database [12, 13]. In the current scenario, no antiviral drugs with proven efficacy nor are there vaccines for the CoV2 prevention exist [14, 15]. Also, the researchers have little knowledge of the molecular biology of SARS-CoV2 infection [16]. In the present state of art, viral infection mechanism is not fully understood though various protein-protein interactions (PPIs) of virus and host are known [17, 18]. So identifying interactions between the SARS-CoV2 virus proteins and host proteins helps understand the mechanism of viral infection and develop treatments and vaccines [19]. Understanding these SARS-CoV2 proteins is one of the primary aims to get a clarity of the PPIs between the virus proteins and host proteins [20]. Biologists yet to understand the spatial arrangement of secondary structure elements (SSEs) [21, 22]. The geometric three dimensional structure of a protein depends on the spatial arrangement of the SSEs which has been studied in [23]. So the spatial distribution as well as presence/absence of different amino acids over a primary protein sequence of SARS-CoV2 are significantly important to reveal. It is needless to mention that the spatial arrangement uncovers the rules that govern the folding of polypeptide chains [24]. Alternation of amino acids over the primary sequence might affect the function of a protein. Also the primary sequence of a protein reveals the molecular events in evolution. The spatial arrangement of amino acids determines the conformability of proteins too [25, 26]. In this present study, spatial composition of twenty amino acids across the primary proteins of SARS-CoV2 have been looked into through parameters viz. Hurst exponent and Shannon entropy. Also frequency analysis of the amino acids over the proteins have been chalked out. It is noted that authors have done similar analysis for the 89 genomes of SARS-CoV2 [27].

1.1. Database used and Specifications

For this present study, as on date 24th March, 2020, there are 944 primary protein sequences of the SARS-CoV2 in the NCBI Virus Database ($https://www.ncbi.nlm.nih.gov/labs/virus/vssi/$), which have been considered [28]. We have named all these 944 complete SARS-CoV2 primary protein sequences as $P1, P2, P3, \ldots, P944$ corresponding to their accession IDs (Ref: Additional file-1). It is noted that out of all these 944 sequences only 105 sequences are distinct (Ref: Additional file-2) although these sequence data have been taken from wide ranges of geographic locations over the world. So we are ending up to the 105 primary protein sequences starting from $N1$, $N2$, $\ldots$, $N105$ which have been considered for the present study.

There are twenty amino acids of three different types as described 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, S.

Replication of a virus depends on the availability of amino acids [29]. Amino acids play a crucial role in virus-related infections as amino acids are necessarily required for protein synthesis [30]. The absence of essential amino acids may result in empty virus particles that are free of viral nucleic acids [31]. The conditionally essential amino acid Arginine (R) is very much needed for the replication of viruses and progression of viral infections. The basic backbone of the amino acids are carbon atom (C) attached to a carboxyl group (-COOH), an amino group, (-NH2), a Hydrogen, and another group of atoms (R). The R group gives the amino acid its unique characteristics, and allows it to react with other amino acids in unique ways. Based on the structure and general chemical characteristics of R groups of the amino acids the following well known classification is given in the following:

- **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

Hereby we order these amino acids as $A_1, A_2, A_3, \dots A_{20}$ corresponding to A, C, F, G, H, I, L, M, N, P, Q, S, T, V, W, Y, D, E, K and R respectively.

Every primary protein sequence has been decomposed into twenty different binary sequences of $0's$ and $1's$, which is formed by the following rule:

Given a primary protein sequences of SARS-CoV2, for every amino acid $A_i \in \{A, C, F, G, H, I, L, M, N, P, Q, S, T, V, W, Y, D, E, K, R\}$ *for* $i = 1$ *to* 20*, put one wherever* A_i *is present and elsewhere put zero.*

Consequently, for every given primary protein sequence Nj for all $j = 1, 2, \dots 105$ there would be twenty binary sequences B_{i_j} corresponding to twenty different amino acids A_i for all $i = 1, 2, \dots 20$.

The length of these complete 105 primary protein sequences is widely varying from 13 to 7097. There is one complete SARS-CoV protein sequence N99 of the smallest length 13 and one protein sequence N26 of the largest length 7097. There are 6, 3, 8, 10, 3, and 48 sequences of lengths 121, 275, 419, 1273, 4405 and 7096 respectively and rest sequences are of unique length ranges from the smallest to largest. So typically all these 105 sequences can be grouped into six groups excluding individual sequences of different unique length.

Table 1: Length of the 105 primary protein sequences

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

2. Methods

In characterizing the amino acids spatial distribution over the primary protein sequences of SARS-CoV2, the three parameters Hurst Exponent, Shannon Entropy, Amino Acid Density are considered. Following these methods are described briefly. Similar works based on these methods are done in [32, 33, 34].

2.1. Hurst Exponent

Fractality (an organized form of nonlinearity) is naturally characterised using fractal dimension. In one dimensional sequence the fractal dimension (D) and the Hurst Exponent (HE) are linearly related as $D + H = 2$ [35, 36]. The Hurst exponent measures the autocorrelation in the sequences [37]. The HE lies in the interval $(0, 1)$. For rough anti-correlated sequence HE is strictly less than

0.5 and for positively correlated sequences the HE ranges between 0.5 to 1. If HE=0.5, then the sequence clearly depicts its randomness with white noise.

The HE of a binary sequence s_n is defined as

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

where

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

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

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

and

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

The auto correlation of the binary representations of each of the amino acids over the SARS-CoV2 protein sequences are obtained through the Hurst exponent.

2.2. Shannon entropy

There are two kinds of Shannon entropy we wish to determine in this present study.

- **Binary Shannon Entropy**: The Shannon entropy (SE) measures information entropy of a Bernoulli process with probability p of the two outcomes (0/1). It is defined as

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

  where $p_1 = \frac{k}{2^l}$ and $p_2 = \frac{l-k}{2^l}$; here l is the length of the binary sequence and k is the number of 1's in the binary sequence of length l [38].The binary Shannon entropy is a measure of the uncertainty in a binary sequence.

Whenever the probability $p = 0$, the event is certain never to occur, and so there is no uncertainty, leading to an entropy of 0. Similarly, if the probability $p = 1$, the result is certain, so the 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) important biological mechanism for expanding the genetic code [39, 40]. To the find the conservation of amino acids in primary protein sequences, Shannon entropy is deployed. The SE is widely used for predicting PTMs. For a given protein sequence, the SE is calculated as follows:

$$SE = -\sum_{i=1}^{20} p_{A_i} log_2(p_{A_i})$$

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

2.3. Amino Acid Density

Over the primary protein sequences of SARS-CoV2 protein sequences, we wish to explore the amino acid frequency distributions and corresponding statistical descriptions [41]. The density of the amino acids over a primary protein sequence can also be found using the following formula:

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

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

3. Results

In this section first the the positive/negative trend of the spatial distribution of each of the twenty amino acids over the protein sequences of SARS-CoV2 through Hurst exponent are reported. The Hurst exponent also would imply the fractality (organized non-linearity) of the spatial representations. In addition, the amount of uncertainty of presence/absence of the amino acids over the protein sequences are determined through Shannon entropy. Also the amino acid conservation information is determined through the Shannon entropy. At last, the frequency distributions of all the amino acids over the protein sequences of SARS-CoV2 are given.

3.1. Hurst Exponent of B_{1_j} for $j = 1, 2, \dots 105$ and Classification

For the amino acid $A_1(A)$ the Hurst exponent(HE) for the 105 binary sequences B_{1_j} for $j = 1, 2, \dots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 1. The HE of the binary representations of ordering of the amino acid A_1 over all the primary protein sequences would reveal the autocorrelation of the amino acid.

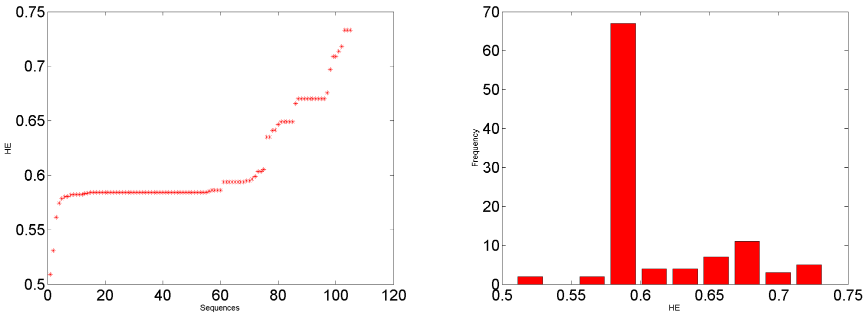


Figure 1: Plot of the HEs and corresponding histogram of all the binary sequences B_{1_j} for $j = 1, 2, \dots 105$ corresponding to the amino acid $A_1(A)$.

Table 2: HE of 105 B_{1_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_1(A)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N80	0.509	3	N31	0.584	7	N26	0.584	7	N83	0.649	5
N4	0.531	3	N33	0.584	7	N8	0.585	7	N84	0.649	5
N103	0.562	6	N34	0.584	7	N22	0.586	7	N85	0.649	5
N87	0.574	7	N35	0.584	7	N59	0.586	7	N86	0.649	5
N105	0.578	7	N37	0.584	7	N65	0.586	7	N74	0.666	1
N20	0.580	7	N38	0.584	7	N29	0.586	7	N70	0.670	1
N7	0.581	7	N39	0.584	7	N88	0.594	2	N69	0.670	1
N81	0.582	7	N40	0.584	7	N89	0.594	2	N71	0.670	1
N48	0.582	7	N41	0.584	7	N90	0.594	2	N72	0.670	1
N50	0.582	7	N42	0.584	7	N91	0.594	2	N73	0.670	1
N61	0.582	7	N45	0.584	7	N92	0.594	2	N75	0.670	1
N43	0.582	7	N46	0.584	7	N93	0.594	2	N76	0.670	1
N12	0.583	7	N47	0.584	7	N94	0.594	2	N77	0.670	1
N13	0.584	7	N49	0.584	7	N95	0.594	2	N78	0.670	1
N44	0.584	7	N51	0.584	7	N32	0.595	2	N79	0.670	1
N14	0.584	7	N52	0.584	7	N36	0.595	2	N101	0.676	1
N16	0.584	7	N53	0.584	7	N58	0.597	2	N98	0.697	8
N17	0.584	7	N54	0.584	7	N68	0.599	2	N96	0.709	10
N18	0.584	7	N55	0.584	7	N1	0.603	2	N97	0.709	10
N19	0.584	7	N56	0.584	7	N5	0.604	2	N2	0.714	9
N21	0.584	7	N57	0.584	7	N6	0.605	2	N99	0.718	9
N23	0.584	7	N60	0.584	7	N100	0.635	5	N9	0.733	4
N24	0.584	7	N62	0.584	7	N104	0.635	5	N10	0.733	4
N25	0.584	7	N63	0.584	7	N3	0.641	5	N11	0.733	4
N27	0.584	7	N64	0.584	7	N102	0.642	5			
N28	0.584	7	N66	0.584	7	N15	0.647	5			
N30	0.584	7	N67	0.584	7	N82	0.649	5			

Here the HE of the 105 binary representation of the amino acid A_1 is ranging from 0.509 to 0.7331 with standard deviation 0.04512. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, ten clustered (C) are formed as presented in the Table 2.

The cluster 3 contains two sequences having HE approximately 0.5 which depicts the random walk/Brownian motion like character of the ordering of the amino acid A_1 over the the sequences N4 and N80. It is noted that all the B_{1_j} for $j = 1, 2, \ldots 105$ except N4 and N80 are trending (persistent) sequences of ordering of the amino acid A_1 over all the 103 primary protein sequences. Out of all, there are 57 B_{1_j} for different values of j having HE at fixed at 0.58 and belonging to the cluster 7. So the amino acid A_1 is orderly placed in a persistent manner (not too persistent through) over most of the primary protein sequences of SARS-CoV2 proteins. There are three clusters 4, 9 and 10 consisting seven binary representations having HE approximately 0.7 which indicates the positive autocorrelation (more persistent). It is noted that there are other sequences having same HE belonging to different other clusters as shown in the Table 2.

3.2. Hurst Exponent of B_{2_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_2(C)$ the Hurst exponent(HE) for the 105 binary sequences B_{2_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 2. The HE of the binary representations of ordering of the amino acid A_2 over all the primary protein sequences would reveal the autocorrelation of the amino acid.

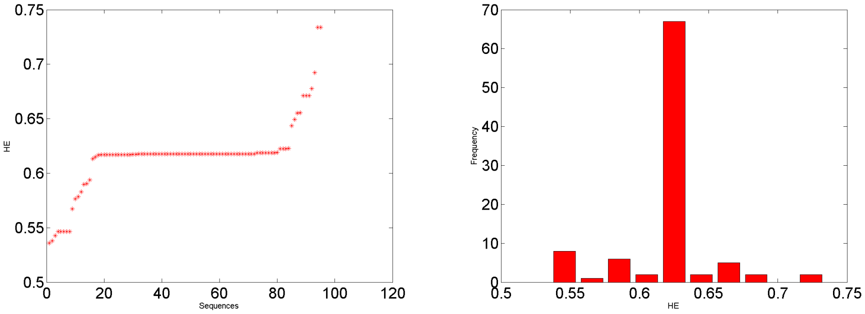


Figure 2: Plot of the HEs and corresponding histogram of all the binary sequences B_{2_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_2(A)$.

Table 3: HE of 105 B_{2_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_2(C)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N68	*	2	N73	0.617	1	N45	0.618	1	N32	0.623	1
N88	*	2	N74	0.617	1	N46	0.618	1	N36	0.623	1
N89	*	2	N75	0.617	1	N47	0.618	1	N58	0.623	1
N90	*	2	N76	0.617	1	N48	0.618	1	N102	0.623	1
N91	*	2	N77	0.617	1	N49	0.618	1	N4	0.644	8
N92	*	2	N78	0.617	1	N50	0.618	1	N2	0.649	8
N93	*	2	N79	0.617	1	N51	0.618	1	N1	0.655	8
N94	*	2	N70	0.617	1	N52	0.618	1	N6	0.655	8
N95	*	2	N13	0.617	1	N53	0.618	1	N9	0.671	5
N99	*	2	N44	0.617	1	N54	0.618	1	N10	0.671	5
N100	0.536	3	N3	0.617	1	N55	0.618	1	N11	0.671	5
N105	0.538	3	N14	0.618	1	N56	0.618	1	N5	0.678	10
N103	0.543	3	N16	0.618	1	N57	0.618	1	N101	0.692	9
N82	0.547	3	N17	0.618	1	N59	0.618	1	N96	0.734	4
N83	0.547	3	N18	0.618	1	N60	0.618	1	N97	0.734	4
N84	0.547	3	N19	0.618	1	N61	0.618	1			
N85	0.547	3	N20	0.618	1	N62	0.618	1			
N86	0.547	3	N21	0.618	1	N63	0.618	1			
N7	0.567	6	N23	0.618	1	N64	0.618	1			
N15	0.576	6	N24	0.618	1	N65	0.618	1			
N8	0.578	6	N27	0.618	1	N66	0.618	1			
N87	0.583	7	N28	0.618	1	N67	0.618	1			
N98	0.590	7	N29	0.618	1	N22	0.619	1			
N104	0.590	7	N30	0.618	1	N25	0.619	1			
N81	0.594	7	N33	0.618	1	N31	0.619	1			
N80	0.613	1	N34	0.618	1	N39	0.619	1			
N72	0.615	1	N35	0.618	1	N40	0.619	1			
N12	0.617	1	N37	0.618	1	N41	0.619	1			
N69	0.617	1	N38	0.618	1	N42	0.619	1			
N71	0.617	1	N43	0.618	1	N26	0.619	1			

Here the HE of the 105 binary representation of the amino acid A_2 is ranging from 0.536 to 0.7338 with standard deviation 0.04512. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, ten clustered (C) are formed as presented in the Table 3.

The cluster 2 contains ten sequences (N68, N88, N89, N90, N91, N92, N93, N94, N95 AND N99) having 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 $B_{2_{95}}$ respectively are completely free from the amino acid A_2(C). The cluster 3 and 6 consist of 8 and 3 sequences respectively, whose HEs are approximately 0.5 which depicts the random walk/Brownian motion like character of the ordering of the amino acid A_2 over the the corresponding protein sequences. It is

doi:10.20944/preprints202004.0034.v1

noted that almost all of these 105 SARS-CoV2 protein sequences are trending (persistent) sequences. Out of all, there are 69 B_{2_j} for different values of j having HE at fixed at 0.61 varied by 0.01 and belonging to the cluster 1. So the amino acid A_2 is orderly placed in a persistent manner over most of the primary protein sequences of SARS-CoV2 proteins. There is a cluster 4 consisting two binary representations having HE approximately 0.734 which indicates the positive autocorrelation (more persistent). It is noted that there are other sequences having same HE belonging to different other clusters as shown in the Table 3.

3.3. Hurst Exponent of B_{3_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_3(F)$ the Hurst exponent(HE) for the 105 binary sequences B_{3_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 3. The HE of the binary representations of ordering of the amino acid A_3 over all the primary protein sequences would reveal the autocorrelation of the amino acid.

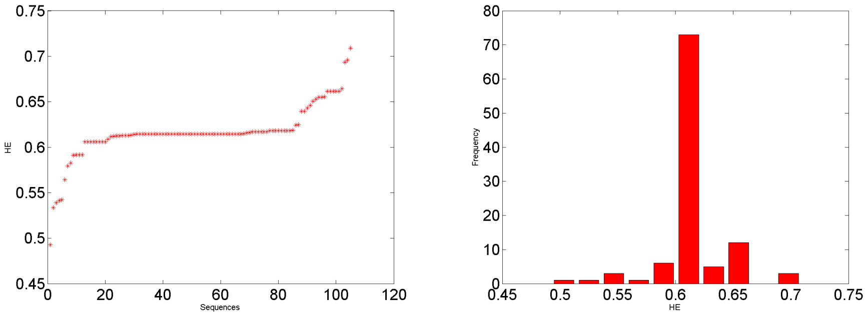


Figure 3: Plot of the HEs and corresponding histogram of all the binary sequences B_{3_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_3(F)$.

Table 4: HE of 105 B_{3_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_3(F)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N5	0.493	3	N13	0.614	8	N55	0.615	8	N96	0.640	1
N1	0.534	6	N44	0.615	8	N56	0.615	8	N97	0.640	1
N6	0.539	6	N14	0.615	8	N57	0.615	8	N7	0.643	10
N3	0.541	6	N16	0.615	8	N61	0.615	8	N68	0.646	10
N2	0.542	6	N17	0.615	8	N63	0.615	8	N101	0.650	10
N15	0.564	6	N18	0.615	8	N64	0.615	8	N36	0.653	10
N102	0.579	5	N19	0.615	8	N65	0.615	8	N32	0.655	10
N104	0.582	5	N21	0.615	8	N66	0.615	8	N58	0.655	10
N103	0.591	5	N23	0.615	8	N67	0.615	8	N99	0.656	10
N9	0.591	5	N25	0.615	8	N4	0.615	8	N82	0.662	4
N10	0.591	5	N28	0.615	8	N12	0.616	8	N83	0.662	4
N11	0.591	5	N29	0.615	8	N42	0.616	8	N84	0.662	4
N88	0.606	7	N30	0.615	8	N20	0.617	8	N85	0.662	4
N89	0.606	7	N31	0.615	8	N27	0.617	8	N86	0.662	4
N90	0.606	7	N35	0.615	8	N33	0.617	8	N105	0.664	4
N91	0.606	7	N37	0.615	8	N34	0.617	8	N8	0.694	2
N92	0.606	7	N38	0.615	8	N59	0.617	8	N81	0.696	2
N93	0.606	7	N39	0.615	8	N62	0.617	8	N80	0.709	9
N94	0.606	7	N41	0.615	8	N69	0.618	8			
N95	0.606	7	N43	0.615	8	N71	0.618	8			
N87	0.609	7	N45	0.615	8	N73	0.618	8			
N26	0.612	8	N46	0.615	8	N74	0.618	8			
N78	0.612	8	N47	0.615	8	N75	0.618	8			
N51	0.612	8	N48	0.615	8	N76	0.618	8			
N60	0.612	8	N49	0.615	8	N77	0.618	8			
N22	0.613	8	N50	0.615	8	N79	0.618	8			
N24	0.613	8	N52	0.615	8	N70	0.618	8			
N40	0.613	8	N53	0.615	8	N72	0.624	1			
N98	0.613	8	N54	0.615	8	N100	0.625	1			

Here the HE of the 105 binary representation of the amino acid A_3 is ranging from 0.493 to 0.709 with standard deviation 0.0296. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, ten clustered (C) are formed as presented in the Table 4.

The largest cluster contains 65 sequences having HE approximately 0.61 which says the binary representations B_{3_j} for $j = 1, 2, \ldots 105$ for the amino acid A_3 are positively autocorrelated/persistent. There are other sequences belonging to the the cluster 7 and 10 having same HEs 0.606 and 0.655 respectively as shown in the Table 3.

3.4. Hurst Exponent of B_{4_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_4(G)$ the Hurst exponent(HE) for the 105 binary sequences B_{4_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 4. The HE of the binary representations of ordering of the amino acid A_4 over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_4 is ranging from 0.546 to 0.664 with standard deviation 0.0876. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, ten clustered (C) are formed as presented in the Table 5.

There are two protein sequences N68 and N81 without any amino acid G (conditionally essential) as it can be seen in the Table 5. All the binary spatial distributions of the 105 proteins are having positive auto correlation and consequently persistent/trending. The most of the sequences (54 in number) are having HE 0.6 belonging to a cluster as shown in the Table 5. The cluster 8 contains N7 and N99 which has the amino acid representations B_{4_7} and $B_{4_{99}}$ respectively having the highest HE 0.664.

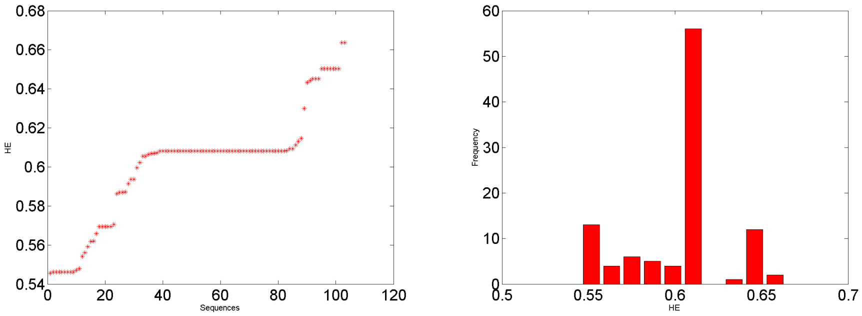


Figure 4: Plot of the HEs and corresponding histogram of all the binary sequences B_{4_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_4(G)$.

Table 5: HE of 105 B_{4_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_4(G)$.

Seq	**HE**	**C**	**Seq**	**HE**	**C**	**Seq**	**HE**	**C**	**Seq**	**HE**	**C**
N68	*	4	N96	0.594	5	N39	0.608	3	N5	0.630	9
N81	*	4	N97	0.594	5	N40	0.608	3	N95	0.643	10
N70	0.546	6	N10	0.600	7	N41	0.608	3	N87	0.644	10
N69	0.546	6	N8	0.602	7	N42	0.608	3	N32	0.645	10
N71	0.546	6	N14	0.606	7	N43	0.608	3	N36	0.645	10
N72	0.546	6	N60	0.606	7	N45	0.608	3	N58	0.645	10
N73	0.546	6	N13	0.606	3	N46	0.608	3	N88	0.650	2
N74	0.546	6	N12	0.607	3	N47	0.608	3	N89	0.650	2
N76	0.546	6	N26	0.607	3	N48	0.608	3	N90	0.650	2
N78	0.546	6	N15	0.607	3	N49	0.608	3	N91	0.650	2
N79	0.546	6	N16	0.608	3	N50	0.608	3	N92	0.650	2
N105	0.547	6	N17	0.608	3	N51	0.608	3	N93	0.650	2
N3	0.548	6	N18	0.608	3	N52	0.608	3	N94	0.650	2
N77	0.554	6	N19	0.608	3	N55	0.608	3	N7	0.664	8
N75	0.556	6	N20	0.608	3	N56	0.608	3	N99	0.664	8
N98	0.559	1	N21	0.608	3	N57	0.608	3			
N100	0.562	1	N22	0.608	3	N59	0.608	3			
N102	0.562	1	N23	0.608	3	N61	0.608	3			
N2	0.566	1	N24	0.608	3	N62	0.608	3			
N82	0.569	1	N25	0.608	3	N63	0.608	3			
N83	0.569	1	N27	0.608	3	N64	0.608	3			
N84	0.569	1	N28	0.608	3	N65	0.608	3			
N85	0.569	1	N29	0.608	3	N66	0.608	3			
N86	0.569	1	N30	0.608	3	N67	0.608	3			
N1	0.571	1	N31	0.608	3	N44	0.608	3			
N4	0.586	5	N33	0.608	3	N53	0.609	3			
N9	0.587	5	N34	0.608	3	N54	0.609	3			
N11	0.587	5	N35	0.608	3	N101	0.611	3			
N6	0.587	5	N37	0.608	3	N80	0.613	3			
N103	0.591	5	N38	0.608	3	N104	0.615	3			

3.5. Hurst Exponent of B_{5_j} *for* $j = 1, 2, \ldots 105$ *and Classification*

For the amino acid $A_5(H)$ the Hurst exponent(HE) for the 105 binary sequences B_{5_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 5. The HE of the binary representations of ordering of the amino acid A_5 over all the primary protein sequences would reveal the autocorrelation of the amino acid.

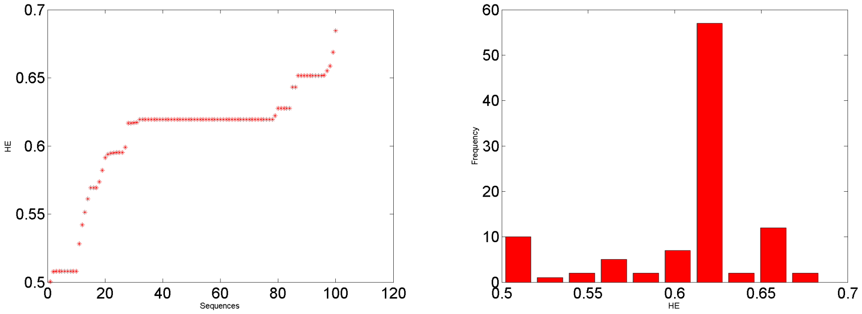


Figure 5: Plot of the HEs and corresponding histogram of all the binary sequences B_{5_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_5(G)$.

Here the HE of the 105 binary representation of the amino acid A_5 is ranging from 0.5 to 0.685 with standard deviation 0.136. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, ten clustered (C) are formed as presented in the Table 6.

One of the essential amino acids H does not present in the protein sequences N3, N80, N97, N98 and N99 of the SARS-CoV2 as found. The spatial organization of the amino acid H is absolutely random(does not either trending nor negatively autocorrelated) in the protein sequences N5, N15, N88, N89, N90, N91, N92, N93, N94 and N95 which belong to the cluster 2 as shown in the Table 6. The largest cluster 8 contains 54 protein sequences where the amino acid H is spatially distributed with positive autocorrelation (i.e. trending sequence) with HE 0.619. Such a organized spatial distributions of the essential amino acid H over most of the protein sequences belonging to the cluster 8 is certainly noteworthy.

Table 6: HE of 105 B_{5_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_5(H)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N3	*	3	N9	0.595	5	N38	0.619	8	N82	0.628	1
N80	*	3	N10	0.595	5	N39	0.619	8	N83	0.628	1
N97	*	3	N11	0.595	5	N40	0.619	8	N84	0.628	1
N98	*	3	N68	0.599	5	N41	0.619	8	N85	0.628	1
N99	*	3	N21	0.617	8	N42	0.619	8	N86	0.628	1
N5	0.500	2	N64	0.617	8	N43	0.619	8	N79	0.643	9
N15	0.508	2	N13	0.617	8	N45	0.619	8	N2	0.643	9
N88	0.508	2	N12	0.617	8	N46	0.619	8	N69	0.652	9
N89	0.508	2	N26	0.619	8	N47	0.619	8	N71	0.652	9
N90	0.508	2	N14	0.619	8	N48	0.619	8	N72	0.652	9
N91	0.508	2	N16	0.619	8	N50	0.619	8	N73	0.652	9
N92	0.508	2	N17	0.619	8	N51	0.619	8	N74	0.652	9
N93	0.508	2	N18	0.619	8	N52	0.619	8	N75	0.652	9
N94	0.508	2	N19	0.619	8	N53	0.619	8	N76	0.652	9
N95	0.508	2	N20	0.619	8	N54	0.619	8	N77	0.652	9
N7	0.528	6	N22	0.619	8	N55	0.619	8	N78	0.652	9
N100	0.542	6	N23	0.619	8	N56	0.619	8	N70	0.652	9
N1	0.551	6	N24	0.619	8	N57	0.619	8	N6	0.655	9
N4	0.561	7	N25	0.619	8	N59	0.619	8	N8	0.659	9
N32	0.569	7	N27	0.619	8	N60	0.619	8	N101	0.669	4
N36	0.569	7	N28	0.619	8	N61	0.619	8	N81	0.685	10
N58	0.569	7	N29	0.619	8	N62	0.619	8			
N87	0.574	7	N30	0.619	8	N63	0.619	8			
N105	0.582	7	N31	0.619	8	N65	0.619	8			
N103	0.591	5	N33	0.619	8	N66	0.619	8			
N96	0.594	5	N34	0.619	8	N67	0.619	8			
N104	0.595	5	N35	0.619	8	N44	0.619	8			
N102	0.595	5	N37	0.619	8	N49	0.622	8			

3.6. *Hurst Exponent of* B_{6_j} *for* $j = 1, 2, \ldots 105$ *and Classification*

For the amino acid $A_6(I)$ the Hurst exponent(HE) for the 105 binary sequences B_{6_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 6. The HE of the binary representations of ordering of the amino acid A_6 over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_6 is ranging from 0.509 to 0.736 with standard deviation 0.0661. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, ten clustered (C) are formed as presented in the Table 7.

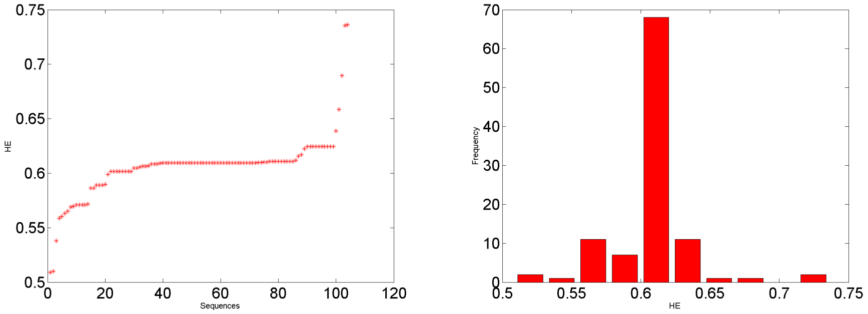


Figure 6: Plot of the HEs and corresponding histogram of all the binary sequences B_{6_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_6(I)$.

Table 7: HE of 105 B_{6_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_6(I)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N99	*	4	N94	0.602	5	N43	0.610	5	N51	0.611	5
N102	0.509	6	N95	0.602	5	N45	0.610	5	N59	0.611	5
N80	0.510	6	N32	0.605	5	N46	0.610	5	N28	0.612	5
N100	0.538	6	N36	0.605	5	N47	0.610	5	N105	0.616	5
N4	0.559	7	N63	0.606	5	N48	0.610	5	N70	0.617	3
N104	0.561	7	N98	0.606	5	N49	0.610	5	N6	0.622	3
N1	0.563	7	N58	0.607	5	N50	0.610	5	N69	0.625	3
N8	0.565	7	N35	0.607	5	N52	0.610	5	N71	0.625	3
N103	0.569	7	N55	0.608	5	N53	0.610	5	N72	0.625	3
N5	0.570	7	N61	0.608	5	N54	0.610	5	N73	0.625	3
N83	0.571	7	N65	0.608	5	N56	0.610	5	N74	0.625	3
N84	0.571	7	N2	0.609	5	N57	0.610	5	N75	0.625	3
N85	0.571	7	N26	0.610	5	N60	0.610	5	N76	0.625	3
N86	0.571	7	N14	0.610	5	N62	0.610	5	N77	0.625	3
N82	0.572	7	N16	0.610	5	N64	0.610	5	N78	0.625	3
N96	0.586	1	N21	0.610	5	N66	0.610	5	N79	0.625	3
N97	0.586	1	N22	0.610	5	N67	0.610	5	N15	0.639	10
N9	0.589	1	N23	0.610	5	N44	0.610	5	N81	0.659	9
N10	0.589	1	N24	0.610	5	N13	0.610	5	N3	0.689	8
N11	0.589	1	N30	0.610	5	N12	0.610	5	N68	0.736	2
N101	0.590	1	N31	0.610	5	N38	0.610	5	N87	0.736	2
N7	0.599	5	N33	0.610	5	N17	0.611	5			
N88	0.602	5	N34	0.610	5	N18	0.611	5			
N89	0.602	5	N37	0.610	5	N19	0.611	5			
N90	0.602	5	N39	0.610	5	N20	0.611	5			
N91	0.602	5	N40	0.610	5	N25	0.611	5			
N92	0.602	5	N41	0.610	5	N27	0.611	5			
N93	0.602	5	N42	0.610	5	N29	0.611	5			

There is one sequence N99 which does not have any the essential amino acid I. The spatial distribution of the amino acid I over the the protein sequence

N102 is truly random since the HE is turned out to be 0.509 whereas the other 104 sequences are trending by having HEs greater than 0.5. The largest cluster 5 contains 71 protein sequences whose spatial distribution of the essential amino acid I are positively autocorrelated (trending) with HE 0.61 (approx.). There are other sequences which belong to different clusters, having same HEs greater than 0.5.

3.7. Hurst Exponent of B_{7_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_7(L)$ the Hurst exponent(HE) for the 105 binary sequences B_{7_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 7. The HE of the binary representations of ordering of the amino acid A_7 over all the primary protein sequences would reveal the autocorrelation of the amino acid.

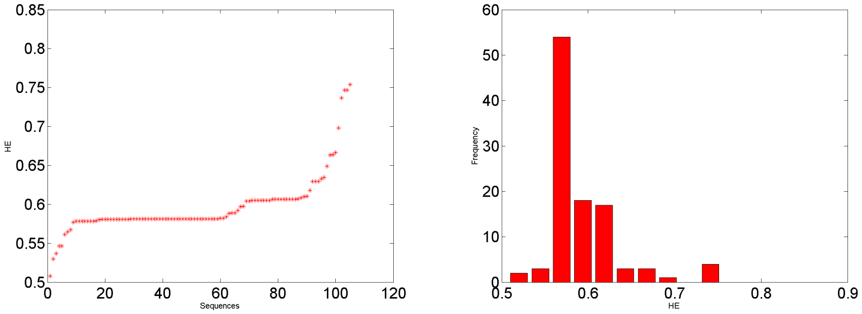


Figure 7: Plot of the HEs and corresponding histogram of all the binary sequences B_{7_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_7(L)$.

Here the HE of the 105 binary representation of the amino acid A_7 is ranging from 0.508 to 0.754 with standard deviation 0.0395. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, ten clustered (C) are formed as presented in the Table 8.

The binary representations B_{7_2}, $B_{7_{68}}$ and $B_{7_{15}}$ of the spatial arrangement of the amino acid L over the protein sequences N2, N68 and N15 are random as the HEs of these sequences is 0.5 (approx.). There are 54 sequences in the cluster 5 are having the HE 0.58. The spatial arrangements of the amino acid L over these proteins are not random but not either too trending as the HE is greater than 0.5 but less than 0.6. There are as usual other clusters having sequences with positive autocorrelation (trending) as given in the Table 8.

Table 8: HE of 105 B_{7_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_7(L)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N2	0.508	3	N16	0.581	5	N24	0.582	5	N105	0.618	7
N68	0.530	3	N17	0.581	5	N30	0.584	5	N82	0.629	8
N15	0.537	3	N18	0.581	5	N101	0.589	5	N85	0.629	8
N5	0.546	6	N19	0.581	5	N32	0.589	5	N86	0.629	8
N3	0.547	6	N21	0.581	5	N58	0.589	5	N6	0.633	8
N104	0.561	6	N22	0.581	5	N36	0.592	5	N102	0.635	8
N1	0.565	6	N28	0.581	5	N83	0.597	1	N7	0.649	8
N4	0.568	6	N29	0.581	5	N98	0.597	1	N96	0.663	4
N100	0.577	5	N31	0.581	5	N84	0.604	1	N99	0.664	4
N20	0.579	5	N35	0.581	5	N87	0.604	1	N81	0.667	4
N27	0.579	5	N38	0.581	5	N89	0.605	1	N97	0.698	10
N33	0.579	5	N39	0.581	5	N90	0.605	1	N80	0.737	9
N34	0.579	5	N43	0.581	5	N91	0.605	1	N10	0.747	2
N47	0.579	5	N45	0.581	5	N92	0.605	1	N11	0.747	2
N59	0.579	5	N46	0.581	5	N93	0.605	1	N9	0.754	2
N62	0.579	5	N48	0.581	5	N94	0.605	1			
N103	0.579	5	N49	0.581	5	N95	0.605	1			
N42	0.580	5	N50	0.581	5	N69	0.607	1			
N25	0.581	5	N51	0.581	5	N71	0.607	1			
N37	0.581	5	N52	0.581	5	N72	0.607	1			
N40	0.581	5	N53	0.581	5	N73	0.607	1			
N41	0.581	5	N54	0.581	5	N74	0.607	1			
N60	0.581	5	N55	0.581	5	N75	0.607	1			
N63	0.581	5	N56	0.581	5	N76	0.607	1			
N23	0.581	5	N57	0.581	5	N77	0.607	1			
N66	0.581	5	N61	0.581	5	N79	0.607	1			
N26	0.581	5	N64	0.581	5	N70	0.607	1			
N12	0.581	5	N65	0.581	5	N8	0.609	7			
N13	0.581	5	N67	0.581	5	N78	0.610	7			
N14	0.581	5	N44	0.582	5	N88	0.610	7			

3.8. Hurst Exponent of B_{8_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_8(M)$ the Hurst exponent(HE) for the 105 binary sequences B_{8_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 8. The HE of the binary representations of ordering of the amino acid A_8 over all the primary protein sequences would reveal the autocorrelation of the amino acid.

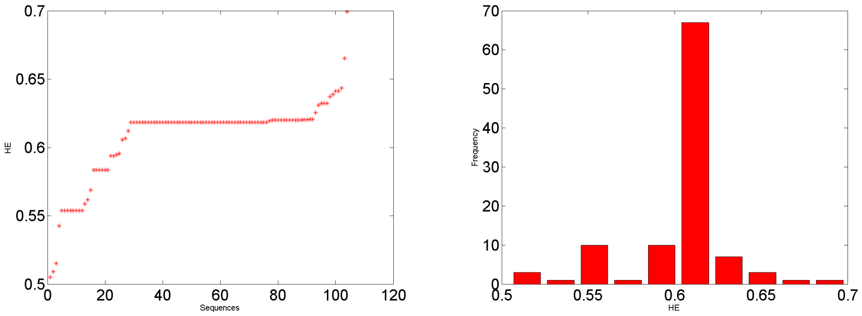


Figure 8: Plot of the HEs and corresponding histogram of all the binary sequences B_{8_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_8(M)$.

Here the HE of the 105 binary representation of the amino acid A_8 is ranging from 0.505 to 0.699 with standard deviation 0.0665. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clustered (C) are formed as presented in the Table 9.

Here the cluster 3 contains most of the sequences (80 in number) for which the spatial distributions of the amino acid M over the protein sequences are having the HE 0.61 (approx) which indicates the trending behaviour. Clearly, the spatial organizations of the amino acid M over the protein sequences N102, N80 and N81 are random. Rest all as usual having the trending trend as seen before.

Table 9: HE of 105 B_{8_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_8(M)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N99	*	4	N18	0.618	3	N49	0.618	3	N13	0.620	3
N102	0.505	5	N14	0.618	3	N50	0.618	3	N52	0.620	3
N80	0.509	5	N16	0.618	3	N51	0.618	3	N59	0.620	3
N81	0.515	5	N17	0.618	3	N53	0.618	3	N8	0.626	3
N2	0.543	5	N19	0.618	3	N54	0.618	3	N68	0.631	3
N88	0.554	5	N20	0.618	3	N55	0.618	3	N9	0.632	3
N89	0.554	5	N21	0.618	3	N56	0.618	3	N10	0.632	3
N90	0.554	5	N22	0.618	3	N57	0.618	3	N11	0.632	3
N91	0.554	5	N23	0.618	3	N60	0.618	3	N100	0.637	3
N92	0.554	5	N24	0.618	3	N61	0.618	3	N103	0.639	3
N93	0.554	5	N25	0.618	3	N62	0.618	3	N32	0.641	3
N94	0.554	5	N27	0.618	3	N63	0.618	3	N36	0.641	3
N95	0.554	5	N28	0.618	3	N64	0.618	3	N58	0.643	3
N5	0.559	5	N29	0.618	3	N65	0.618	3	N101	0.665	2
N4	0.562	5	N30	0.618	3	N66	0.618	3	N3	0.699	2
N104	0.569	1	N31	0.618	3	N67	0.618	3			
N82	0.583	1	N33	0.618	3	N26	0.619	3			
N83	0.583	1	N34	0.618	3	N105	0.619	3			
N84	0.583	1	N35	0.618	3	N69	0.620	3			
N85	0.583	1	N37	0.618	3	N71	0.620	3			
N86	0.583	1	N38	0.618	3	N72	0.620	3			
N87	0.583	1	N39	0.618	3	N73	0.620	3			
N96	0.594	1	N40	0.618	3	N74	0.620	3			
N97	0.594	1	N41	0.618	3	N75	0.620	3			
N7	0.595	1	N42	0.618	3	N76	0.620	3			
N15	0.596	1	N43	0.618	3	N77	0.620	3			
N1	0.606	3	N45	0.618	3	N78	0.620	3			
N6	0.607	3	N46	0.618	3	N79	0.620	3			
N98	0.612	3	N47	0.618	3	N70	0.620	3			
N44	0.618	3	N48	0.618	3	N12	0.620	3			

3.9. Hurst Exponent of B_{9_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_9(N)$ the Hurst exponent(HE) for the 105 binary sequences B_{9_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 9. The HE of the binary representations of ordering of the amino acid A_9 over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_9 is ranging from 0.505 to 0.699 with standard deviation 0.0665. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clustered (C) are formed as presented in the Table 10.

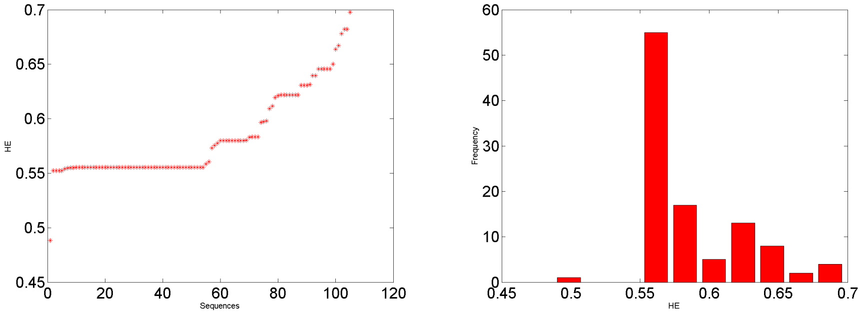


Figure 9: Plot of the HEs and corresponding histogram of all the binary sequences B_{9_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_9(N)$.

Table 10: HE of 105 B_{9_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_9(N)$.

Seq	**HE**	**C**	**Seq**	**HE**	**C**	**Seq**	**HE**	**C**	**Seq**	**HE**	**C**
N2	0.488	1	N40	0.555	1	N72	0.580	4	N89	0.631	5
N51	0.552	1	N41	0.555	1	N73	0.580	4	N96	0.640	3
N16	0.552	1	N42	0.555	1	N74	0.580	4	N97	0.640	3
N46	0.552	1	N43	0.555	1	N75	0.580	4	N82	0.646	3
N57	0.552	1	N45	0.555	1	N76	0.580	4	N83	0.646	3
N12	0.554	1	N47	0.555	1	N77	0.580	4	N84	0.646	3
N13	0.555	1	N48	0.555	1	N78	0.580	4	N85	0.646	3
N5	0.555	1	N49	0.555	1	N79	0.580	4	N86	0.646	3
N44	0.555	1	N50	0.555	1	N70	0.580	4	N87	0.650	3
N14	0.555	1	N52	0.555	1	N103	0.583	4	N99	0.664	3
N17	0.555	1	N53	0.555	1	N32	0.583	4	N3	0.667	2
N18	0.555	1	N54	0.555	1	N36	0.583	4	N104	0.678	2
N19	0.555	1	N55	0.555	1	N58	0.583	4	N98	0.682	2
N20	0.555	1	N56	0.555	1	N101	0.597	4	N15	0.682	2
N21	0.555	1	N59	0.555	1	N102	0.597	4	N68	0.698	2
N22	0.555	1	N60	0.555	1	N4	0.598	4			
N23	0.555	1	N61	0.555	1	N81	0.609	5			
N24	0.555	1	N62	0.555	1	N6	0.612	5			
N25	0.555	1	N63	0.555	1	N8	0.619	5			
N27	0.555	1	N64	0.555	1	N80	0.621	5			
N28	0.555	1	N65	0.555	1	N88	0.622	5			
N29	0.555	1	N66	0.555	1	N90	0.622	5			
N30	0.555	1	N67	0.555	1	N91	0.622	5			
N31	0.555	1	N26	0.555	1	N92	0.622	5			
N33	0.555	1	N7	0.558	1	N93	0.622	5			
N34	0.555	1	N105	0.561	1	N94	0.622	5			
N35	0.555	1	N69	0.573	4	N95	0.622	5			
N37	0.555	1	N1	0.576	4	N9	0.631	5			
N38	0.555	1	N100	0.577	4	N10	0.631	5			
N39	0.555	1	N71	0.580	4	N11	0.631	5			

Here the spatial distribution of the amino acid N (a non-essential amino acid) over the protein sequence N2 is reverse trending (negatively autocorrelated, HE: 0.488) as observed. In the cluster 1 there are 54 sequences having a slow positive trend (HE: 0.55) whereas the clusters 3,4 and 5 contain all the positive trendy spatial representations of the amino acid N over the protein sequences.

3.10. Hurst Exponent of B_{10_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{10}(P)$ the Hurst exponent(HE) for the 105 binary sequences B_{10_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 10. The HE of the binary representations of ordering of the amino acid A_{10} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_{10} is ranging from 0.505 to 0.699 with standard deviation 0.0665. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clustered (C) are formed as presented in the Table 11.

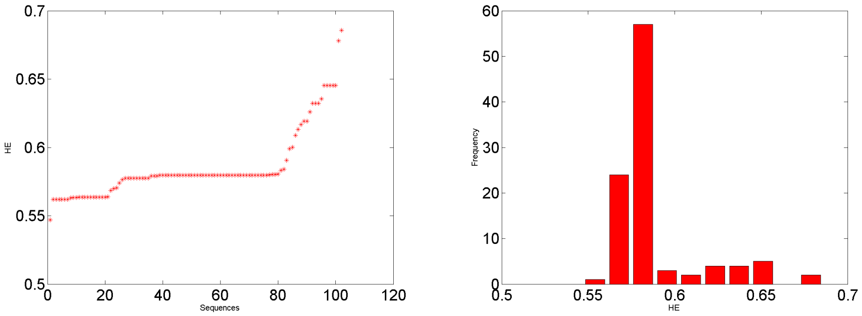


Figure 10: Plot of the HEs and corresponding histogram of all the binary sequences B_{10_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{10}(P)$.

One of the conditionally essential amino acids P does not arise in the protein sequences N81, N99 and N103 of SARS-CoV2. The cluster 1 contains 84 B_{10_j} for 74 different protein sequences where the amino acid P is distributed spatially in a positive trendy manner since the HE is approximately 0.56. As usual there are other clusters containing positive trendy sequence of spatial distribution of the amino acid P.

Table 11: HE of 105 B_{10_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{10}(P)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N81	*	2	N25	0.578	1	N43	0.580	1	N87	0.617	4
N99	*	2	N40	0.578	1	N45	0.580	1	N96	0.619	4
N103	*	2	N41	0.578	1	N46	0.580	1	N97	0.619	4
N94	0.547	1	N42	0.578	1	N47	0.580	1	N8	0.626	4
N88	0.562	1	N56	0.578	1	N48	0.580	1	N9	0.632	5
N89	0.562	1	N23	0.578	1	N49	0.580	1	N10	0.632	5
N90	0.562	1	N66	0.578	1	N50	0.580	1	N11	0.632	5
N92	0.562	1	N67	0.578	1	N51	0.580	1	N3	0.636	5
N93	0.562	1	N32	0.579	1	N52	0.580	1	N82	0.645	5
N95	0.562	1	N36	0.579	1	N53	0.580	1	N83	0.645	5
N102	0.563	1	N58	0.579	1	N54	0.580	1	N84	0.645	5
N5	0.563	1	N14	0.580	1	N55	0.580	1	N85	0.645	5
N15	0.563	1	N16	0.580	1	N57	0.580	1	N86	0.645	5
N69	0.564	1	N17	0.580	1	N59	0.580	1	N6	0.678	3
N71	0.564	1	N18	0.580	1	N60	0.580	1	N98	0.686	3
N72	0.564	1	N19	0.580	1	N62	0.580	1			
N73	0.564	1	N20	0.580	1	N64	0.580	1			
N74	0.564	1	N21	0.580	1	N65	0.580	1			
N75	0.564	1	N22	0.580	1	N63	0.580	1			
N76	0.564	1	N24	0.580	1	N44	0.580	1			
N77	0.564	1	N27	0.580	1	N61	0.580	1			
N78	0.564	1	N28	0.580	1	N13	0.580	1			
N79	0.564	1	N29	0.580	1	N12	0.581	1			
N70	0.564	1	N30	0.580	1	N37	0.583	1			
N104	0.568	1	N31	0.580	1	N91	0.584	1			
N7	0.570	1	N33	0.580	1	N4	0.591	1			
N2	0.571	1	N34	0.580	1	N68	0.599	4			
N105	0.574	1	N35	0.580	1	N1	0.600	4			
N100	0.577	1	N38	0.580	1	N101	0.609	4			
N26	0.578	1	N39	0.580	1	N80	0.613	4			

3.11. Hurst Exponent of B_{11_j} *for* $j = 1, 2, \ldots 105$ *and Classification*

For the amino acid $A_{11}(Q)$ the Hurst exponent(HE) for the 105 binary sequences B_{11_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 11. The HE of the binary representations of ordering of the amino acid A_{11} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_{11} is ranging from 0.495 to 0.671 with standard deviation 0.0903. Based on the HEs of the

335 binary sequences all these 105 primary protein sequences of SARS-CoV2, five clustered (C) are formed as presented in the Table 12.

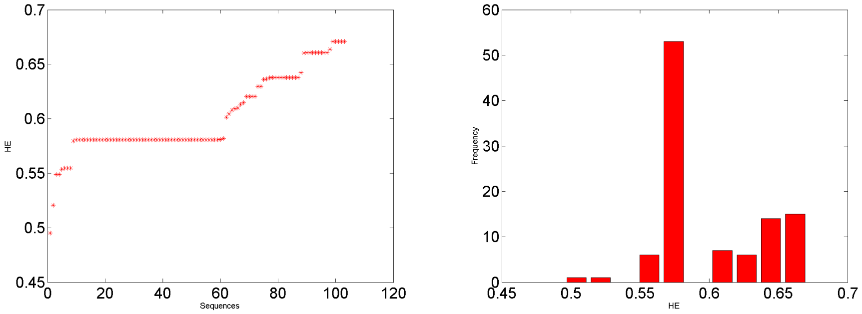


Figure 11: Plot of the HEs and corresponding histogram of all the binary sequences B_{11_j} for $j = 1, 2, \dots 105$ corresponding to the amino acid $A_{11}(Q)$.

Table 12: HE of 105 B_{11_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{11}(Q)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N96	*	3	N37	0.581	2	N13	0.581	2	N68	0.660	4
N97	*	3	N38	0.581	2	N12	0.581	2	N88	0.661	4
N100	0.495	5	N39	0.581	2	N5	0.582	2	N89	0.661	4
N2	0.521	5	N40	0.581	2	N105	0.601	2	N90	0.661	4
N102	0.549	5	N41	0.581	2	N7	0.604	2	N91	0.661	4
N1	0.549	5	N42	0.581	2	N15	0.608	1	N92	0.661	4
N6	0.554	5	N43	0.581	2	N81	0.609	1	N93	0.661	4
N9	0.555	5	N45	0.581	2	N3	0.610	1	N94	0.661	4
N10	0.555	5	N46	0.581	2	N80	0.613	1	N95	0.661	4
N11	0.555	5	N47	0.581	2	N101	0.615	1	N99	0.664	4
N21	0.580	2	N48	0.581	2	N98	0.620	1	N82	0.671	4
N26	0.581	2	N49	0.581	2	N32	0.621	1	N83	0.671	4
N14	0.581	2	N50	0.581	2	N36	0.621	1	N84	0.671	4
N16	0.581	2	N51	0.581	2	N58	0.621	1	N85	0.671	4
N17	0.581	2	N52	0.581	2	N87	0.630	1	N86	0.671	4
N18	0.581	2	N53	0.581	2	N4	0.630	1			
N19	0.581	2	N54	0.581	2	N103	0.636	1			
N20	0.581	2	N55	0.581	2	N8	0.637	1			
N22	0.581	2	N56	0.581	2	N70	0.637	1			
N23	0.581	2	N57	0.581	2	N69	0.638	1			
N24	0.581	2	N59	0.581	2	N71	0.638	1			
N25	0.581	2	N60	0.581	2	N72	0.638	1			
N27	0.581	2	N61	0.581	2	N73	0.638	1			
N28	0.581	2	N62	0.581	2	N74	0.638	1			
N29	0.581	2	N63	0.581	2	N75	0.638	1			
N30	0.581	2	N64	0.581	2	N76	0.638	1			
N31	0.581	2	N65	0.581	2	N77	0.638	1			
N33	0.581	2	N66	0.581	2	N78	0.638	1			
N34	0.581	2	N67	0.581	2	N79	0.638	1			
N35	0.581	2	N44	0.581	2	N104	0.642	1			

In the cluster 1, there are two protein sequences N96 and N97 which are absolutely free from the amino acid Q. The cluster 2 contains 45 sequences having HE 0.58 and so the spatial organization of the amino acid Q is positively trending. As usual there are other three clusters 1, 4 and 5 which contain positive autocorrelated sequences of the spatial distribution of the amino acid Q over the protein sequences. There is only one binary representation $B_{11_{100}}$ of the amino acid Q over the protein sequence N100 having negatively trending.

3.12. Hurst Exponent of B_{12_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{12}(S)$ the Hurst exponent(HE) for the 105 binary sequences B_{12_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given

in the Fig 12. The HE of the binary representations of ordering of the amino acid A_{12} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_{12} is ranging from 0.486 to 0.694 with standard deviation 0.04. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, eight clustered (C) are formed as presented in the Table 13.

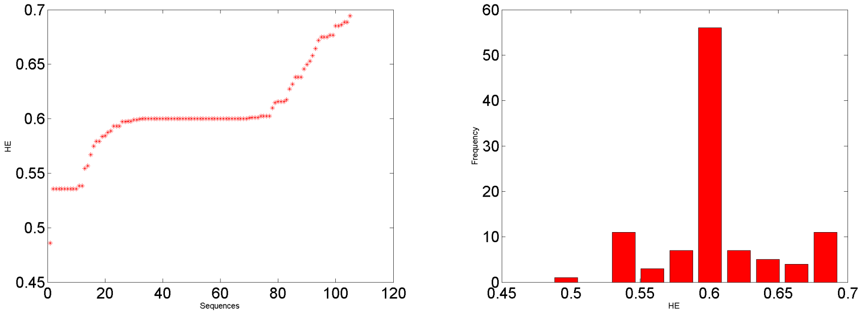


Figure 12: Plot of the HEs and corresponding histogram of all the binary sequences B_{12_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{12}(S)$.

Table 13: HE of 105 B_{12_j} for $j = 1, 2, \dots 105$ corresponding to the amino acid $A_{12}(S)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N7	0.486	6	N56	0.599	2	N53	0.600	2	N80	0.653	5
N70	0.536	1	N13	0.600	2	N55	0.600	2	N68	0.658	5
N69	0.536	1	N44	0.600	2	N57	0.600	2	N103	0.664	5
N72	0.536	1	N17	0.600	2	N59	0.600	2	N3	0.672	3
N73	0.536	1	N18	0.600	2	N63	0.600	2	N88	0.675	3
N74	0.536	1	N19	0.600	2	N64	0.600	2	N89	0.675	3
N75	0.536	1	N21	0.600	2	N65	0.600	2	N93	0.675	3
N77	0.536	1	N22	0.600	2	N66	0.600	2	N96	0.677	3
N78	0.536	1	N23	0.600	2	N67	0.600	2	N97	0.677	3
N79	0.536	1	N24	0.600	2	N12	0.601	2	N91	0.685	4
N71	0.538	1	N25	0.600	2	N40	0.601	2	N94	0.685	4
N76	0.538	1	N27	0.600	2	N16	0.601	2	N8	0.686	4
N5	0.554	1	N29	0.600	2	N62	0.601	2	N90	0.689	4
N81	0.557	1	N31	0.600	2	N26	0.602	2	N92	0.689	4
N98	0.567	1	N33	0.600	2	N51	0.603	2	N95	0.694	4
N101	0.575	2	N34	0.600	2	N54	0.603	2			
N84	0.579	2	N35	0.600	2	N61	0.603	2			
N86	0.579	2	N37	0.600	2	N100	0.610	8			
N2	0.584	2	N38	0.600	2	N6	0.615	8			
N99	0.584	2	N39	0.600	2	N32	0.616	8			
N15	0.588	2	N41	0.600	2	N36	0.616	8			
N1	0.589	2	N42	0.600	2	N58	0.616	8			
N82	0.593	2	N43	0.600	2	N102	0.617	8			
N83	0.593	2	N45	0.600	2	N104	0.627	8			
N85	0.593	2	N46	0.600	2	N105	0.632	7			
N20	0.597	2	N47	0.600	2	N9	0.638	7			
N30	0.597	2	N48	0.600	2	N10	0.638	7			
N14	0.598	2	N49	0.600	2	N11	0.638	7			
N28	0.598	2	N50	0.600	2	N4	0.646	7			
N60	0.599	2	N52	0.600	2	N87	0.650	5			

The binary representation B_{12_7} of spatial organization of the non essential amino acid 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 which are positively trending. It noteworthy that the spatial representations of the amino acid S over the protein sequences N56, N13, N44, . . . , N67 (belonging to the cluster 2) are having HEs 0.6 exactly, implying the positive autocorrelation.

3.13. Hurst Exponent of B_{13_j} for $j = 1, 2, \dots 105$ and Classification

For the amino acid $A_{13}(T)$ the Hurst exponent(HE) for the 105 binary sequences B_{13_j} for $j = 1, 2, \dots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given

in the Fig 13. The HE of the binary representations of ordering of the amino acid A_{13} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_{13} is ranging from 0.523 to 0.685 with standard deviation 0.0665. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clustered (C) are formed as presented in the Table 14.

The spatial representation $B_{13_{99}}$ of the essential amino acid T is a null sequences having only zeros which imply the absence of the amino acid over the protein sequence N99. The spatial distributions of the amino acid T over the 76 protein sequences (belong to cluster 1) are positively trending.

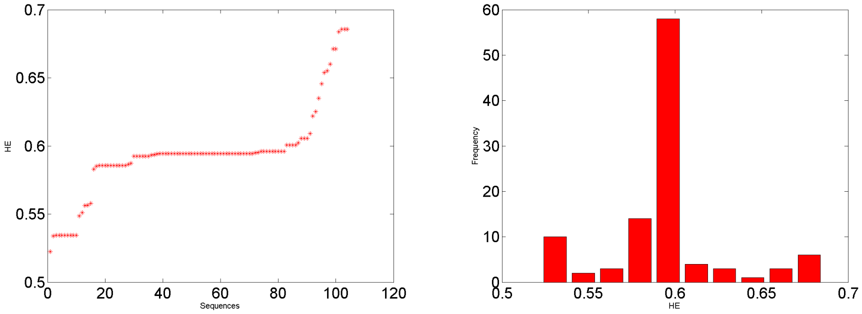


Figure 13: Plot of the HEs and corresponding histogram of all the binary sequences B_{13_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{13}(T)$.

Table 14: HE of 105 B_{13_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{13}(T)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N99	*	3	N25	0.592	1	N52	0.594	1	N58	0.606	1
N7	0.523	2	N27	0.592	1	N53	0.594	1	N81	0.609	1
N8	0.534	2	N29	0.592	1	N54	0.594	1	N102	0.622	5
N88	0.534	2	N34	0.592	1	N56	0.594	1	N2	0.625	5
N89	0.534	2	N35	0.592	1	N57	0.594	1	N4	0.635	5
N90	0.534	2	N65	0.593	1	N59	0.594	1	N1	0.646	5
N91	0.534	2	N12	0.593	1	N60	0.594	1	N105	0.654	5
N92	0.534	2	N13	0.594	1	N62	0.594	1	N87	0.655	5
N93	0.534	2	N44	0.594	1	N64	0.594	1	N98	0.660	5
N94	0.534	2	N16	0.594	1	N66	0.594	1	N96	0.671	4
N95	0.534	2	N21	0.594	1	N67	0.594	1	N97	0.671	4
N68	0.549	2	N22	0.594	1	N26	0.594	1	N3	0.684	4
N101	0.551	2	N23	0.594	1	N6	0.595	1	N9	0.686	4
N82	0.556	2	N24	0.594	1	N61	0.595	1	N10	0.686	4
N15	0.557	2	N30	0.594	1	N17	0.596	1	N11	0.686	4
N100	0.558	2	N31	0.594	1	N18	0.596	1			
N5	0.583	1	N33	0.594	1	N19	0.596	1			
N70	0.585	1	N37	0.594	1	N20	0.596	1			
N69	0.586	1	N39	0.594	1	N28	0.596	1			
N71	0.586	1	N40	0.594	1	N38	0.596	1			
N72	0.586	1	N41	0.594	1	N63	0.596	1			
N73	0.586	1	N42	0.594	1	N14	0.596	1			
N74	0.586	1	N43	0.594	1	N55	0.596	1			
N75	0.586	1	N45	0.594	1	N83	0.601	1			
N76	0.586	1	N46	0.594	1	N84	0.601	1			
N77	0.586	1	N47	0.594	1	N85	0.601	1			
N78	0.586	1	N48	0.594	1	N86	0.601	1			
N79	0.586	1	N49	0.594	1	N80	0.602	1			
N103	0.586	1	N50	0.594	1	N32	0.606	1			
N104	0.587	1	N51	0.594	1	N36	0.606	1			

3.14. *Hurst Exponent of* B_{14_j} *for* $j = 1, 2, \ldots 105$ *and Classification*

For the amino acid $A_{14}(V)$ the Hurst exponent(HE) for the 105 binary sequences B_{14_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 14. The HE of the binary representations of ordering of the amino acid A_{14} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_{14} is ranging from 0.507 to 0.723 with standard deviation 0.0450. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clustered (C) are formed as presented in the Table 15.

doi:10.20944/preprints202004.0034.v1

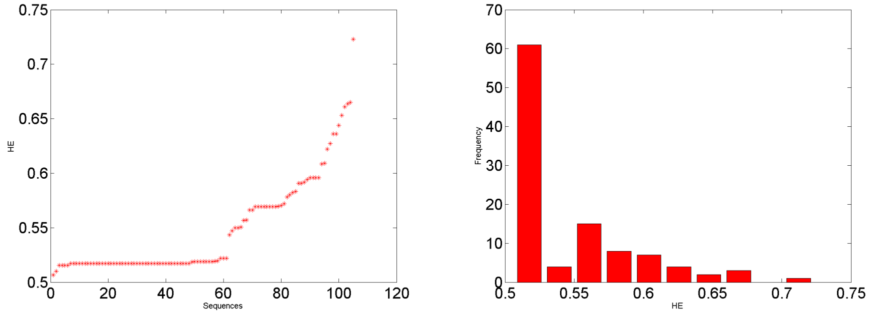


Figure 14: Plot of the HEs and corresponding histogram of all the binary sequences B_{14_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{14}(V)$.

Table 15: HE of 105 B_{14_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{14}(V)$.

Seq	**HE**	**C**	**Seq**	**HE**	**C**	**Seq**	**HE**	**C**	**Seq**	**HE**	**C**
N8	0.507	2	N42	0.517	2	N53	0.522	2	N84	0.596	3
N5	0.510	2	N43	0.517	2	N100	0.544	5	N85	0.596	3
N59	0.516	2	N45	0.517	2	N58	0.547	5	N6	0.596	3
N61	0.516	2	N46	0.517	2	N32	0.550	5	N98	0.608	8
N63	0.516	2	N47	0.517	2	N36	0.550	5	N81	0.609	8
N65	0.516	2	N49	0.517	2	N1	0.551	5	N103	0.622	8
N44	0.517	2	N51	0.517	2	N92	0.557	5	N101	0.627	1
N14	0.517	2	N52	0.517	2	N74	0.557	5	N83	0.636	1
N16	0.517	2	N54	0.517	2	N96	0.566	7	N86	0.636	1
N17	0.517	2	N55	0.517	2	N97	0.566	7	N87	0.644	1
N18	0.517	2	N56	0.517	2	N69	0.569	7	N68	0.653	6
N19	0.517	2	N57	0.517	2	N71	0.569	7	N104	0.661	6
N20	0.517	2	N60	0.517	2	N72	0.569	7	N99	0.664	6
N21	0.517	2	N62	0.517	2	N73	0.569	7	N15	0.665	6
N23	0.517	2	N64	0.517	2	N75	0.569	7	N80	0.723	4
N24	0.517	2	N66	0.517	2	N76	0.569	7			
N25	0.517	2	N67	0.517	2	N78	0.569	7			
N27	0.517	2	N26	0.517	2	N79	0.569	7			
N28	0.517	2	N22	0.519	2	N70	0.570	7			
N29	0.517	2	N88	0.519	2	N102	0.570	7			
N30	0.517	2	N89	0.519	2	N77	0.572	7			
N31	0.517	2	N90	0.519	2	N2	0.578	7			
N33	0.517	2	N91	0.519	2	N10	0.580	7			
N34	0.517	2	N93	0.519	2	N3	0.582	3			
N35	0.517	2	N94	0.519	2	N7	0.583	3			
N37	0.517	2	N95	0.519	2	N9	0.591	3			
N38	0.517	2	N12	0.519	2	N11	0.591	3			
N39	0.517	2	N13	0.520	2	N4	0.592	3			
N40	0.517	2	N48	0.522	2	N105	0.594	3			
N41	0.517	2	N50	0.522	2	N82	0.596	3			

The largest cluster 2 contains 61 binary representations B_{14_j} of the spatial distribution of the amino acid V over the corresponding protein sequences, which are random as the HE is 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.

3.15. Hurst Exponent of B_{15_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{15}(W)$ the Hurst exponent(HE) for the 105 binary sequences B_{15_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 15. The HE of the binary representations of ordering of the amino acid A_{15} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_{15} is ranging from 0.535 to 0.672. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clustered (C) are formed as presented in the Table 16.

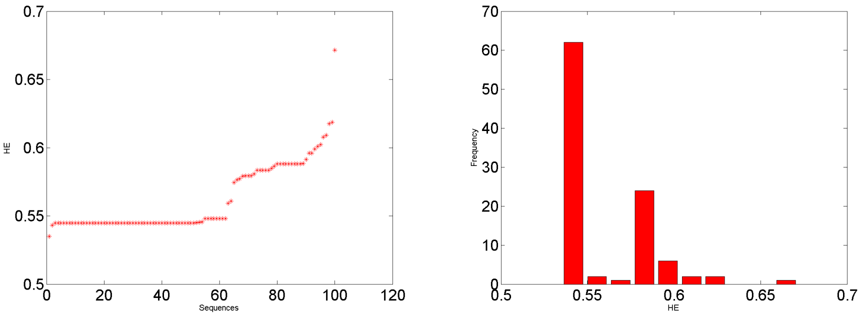


Figure 15: Plot of the HEs and corresponding histogram of all the binary sequences B_{15_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{15}(V)$.

Table 16: HE of 105 B_{15_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{15}(W)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N80	*	2	N40	0.545	3	N89	0.548	3	N77	0.588	1
N87	*	2	N41	0.545	3	N90	0.548	3	N78	0.588	1
N96	*	2	N42	0.545	3	N91	0.548	3	N79	0.588	1
N97	*	2	N43	0.545	3	N92	0.548	3	N70	0.588	1
N99	*	2	N45	0.545	3	N93	0.548	3	N103	0.591	1
N1	0.535	3	N46	0.545	3	N94	0.548	3	N10	0.596	1
N100	0.543	3	N47	0.545	3	N95	0.548	3	N11	0.596	1
N26	0.545	3	N48	0.545	3	N6	0.559	3	N68	0.599	5
N14	0.545	3	N49	0.545	3	N5	0.561	3	N101	0.601	5
N16	0.545	3	N50	0.545	3	N105	0.575	1	N76	0.602	5
N17	0.545	3	N51	0.545	3	N15	0.576	1	N104	0.608	5
N18	0.545	3	N52	0.545	3	N9	0.577	1	N81	0.609	5
N19	0.545	3	N53	0.545	3	N7	0.579	1	N3	0.617	5
N20	0.545	3	N54	0.545	3	N32	0.580	1	N4	0.619	5
N21	0.545	3	N55	0.545	3	N36	0.580	1	N8	0.672	4
N22	0.545	3	N56	0.545	3	N58	0.580	1			
N23	0.545	3	N57	0.545	3	N2	0.581	1			
N24	0.545	3	N59	0.545	3	N82	0.583	1			
N25	0.545	3	N60	0.545	3	N83	0.583	1			
N27	0.545	3	N61	0.545	3	N84	0.583	1			
N28	0.545	3	N62	0.545	3	N85	0.583	1			
N29	0.545	3	N63	0.545	3	N86	0.583	1			
N30	0.545	3	N64	0.545	3	N98	0.585	1			
N31	0.545	3	N65	0.545	3	N102	0.586	1			
N33	0.545	3	N66	0.545	3	N69	0.588	1			
N34	0.545	3	N67	0.545	3	N71	0.588	1			
N35	0.545	3	N44	0.545	3	N72	0.588	1			
N37	0.545	3	N13	0.545	3	N73	0.588	1			
N38	0.545	3	N12	0.546	3	N74	0.588	1			
N39	0.545	3	N88	0.548	3	N75	0.588	1			

The essential amino acid W is absent throughout the protein sequences of 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 the Table 16. Rest all the other binary representations of the spatial distribution of the amino acid V over the protein sequences are positively trending. There are 64 protein sequences (cluster 3) where the amino acid W is spread with positive trend as shown in the Table 16.

3.16. Hurst Exponent of B_{16_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{16}(Y)$ the Hurst exponent(HE) for the 105 binary sequences B_{16_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given

in the Fig 16. The HE of the binary representations of ordering of the amino acid A_{16} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

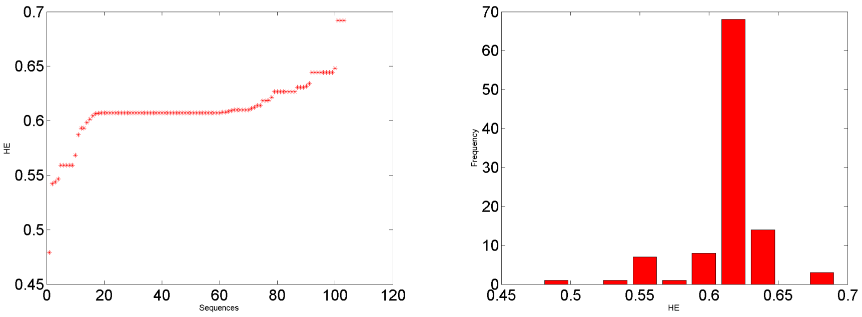


Figure 16: Plot of the HEs and corresponding histogram of all the binary sequences B_{16_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{16}(Y)$.

Here the HE of the 105 binary representation of the amino acid A_{16} is ranging from 0.479 to 0.692 with standard deviation 0.0884. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, six clustered (C) are formed as presented in the Table 17.

The conditional amino acid Y is absent in the protein sequences N99 and N103. The spatial distribution of the amino acid Y over the only protein N80 belonging to the cluster 6 is not trending as its HE $0.479 < 0.5$. The largest cluster 1 contains 68 protein sequences where the amino acid Y spatially spread with positive trend.

Table 17: HE of 105 B_{16_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{16}(Y)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N99	*	4	N27	0.607	1	N66	0.607	1	N58	0.631	5
N103	*	4	N28	0.607	1	N67	0.607	1	N69	0.632	5
N80	0.479	6	N29	0.607	1	N22	0.608	1	N8	0.634	5
N2	0.542	3	N30	0.607	1	N98	0.608	1	N88	0.644	5
N100	0.544	3	N31	0.607	1	N49	0.609	1	N89	0.644	5
N105	0.546	3	N33	0.607	1	N81	0.609	1	N90	0.644	5
N82	0.559	3	N34	0.607	1	N102	0.610	1	N91	0.644	5
N83	0.559	3	N35	0.607	1	N25	0.610	1	N92	0.644	5
N84	0.559	3	N37	0.607	1	N40	0.610	1	N93	0.644	5
N85	0.559	3	N38	0.607	1	N41	0.610	1	N94	0.644	5
N86	0.559	3	N39	0.607	1	N42	0.610	1	N95	0.644	5
N87	0.568	3	N43	0.607	1	N26	0.610	1	N101	0.648	5
N3	0.587	1	N45	0.607	1	N104	0.611	1	N9	0.692	2
N5	0.593	1	N46	0.607	1	N15	0.612	1	N10	0.692	2
N6	0.593	1	N47	0.607	1	N96	0.614	1	N11	0.692	2
N68	0.598	1	N48	0.607	1	N97	0.614	1			
N1	0.602	1	N50	0.607	1	N7	0.618	1			
N64	0.605	1	N51	0.607	1	N70	0.619	1			
N12	0.606	1	N52	0.607	1	N4	0.619	1			
N13	0.607	1	N53	0.607	1	N79	0.621	1			
N44	0.607	1	N54	0.607	1	N71	0.627	5			
N14	0.607	1	N55	0.607	1	N72	0.627	5			
N16	0.607	1	N56	0.607	1	N73	0.627	5			
N17	0.607	1	N57	0.607	1	N74	0.627	5			
N18	0.607	1	N59	0.607	1	N75	0.627	5			
N19	0.607	1	N60	0.607	1	N76	0.627	5			
N20	0.607	1	N61	0.607	1	N77	0.627	5			
N21	0.607	1	N62	0.607	1	N78	0.627	5			
N23	0.607	1	N63	0.607	1	N32	0.631	5			
N24	0.607	1	N65	0.607	1	N36	0.631	5			

3.17. Hurst Exponent of B_{17_j} *for* $j = 1, 2, \ldots 105$ *and Classification*

For the amino acid $A_{17}(D)$ the Hurst exponent(HE) for the 105 binary sequences B_{17_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 17. The HE of the binary representations of ordering of the amino acid A_{17} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_{17} is ranging from 0.501 to 0.6918 with standard deviation 0.0434. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, six clustered (C) are formed as presented in the Table 18.

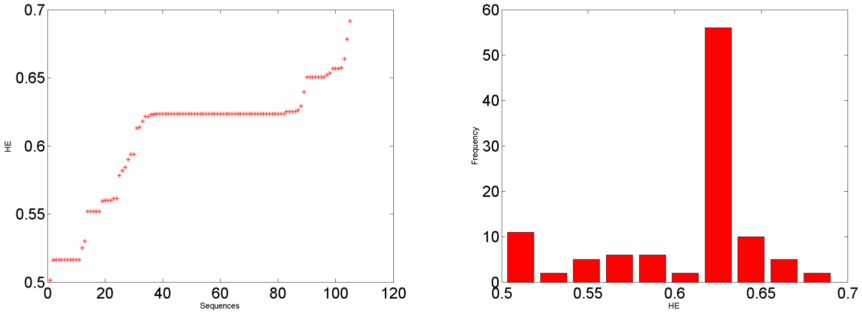


Figure 17: Plot of the HEs and corresponding histogram of all the binary sequences B_{17_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{17}(D)$.

Table 18: HE of 105 B_{17_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{17}(D)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N2	0.501	5	N80	0.613	1	N42	0.624	1	N89	0.650	2
N70	0.516	5	N68	0.614	1	N45	0.624	1	N90	0.650	2
N69	0.516	5	N4	0.618	1	N47	0.624	1	N91	0.650	2
N71	0.516	5	N7	0.622	1	N48	0.624	1	N93	0.650	2
N72	0.516	5	N43	0.622	1	N49	0.624	1	N94	0.650	2
N73	0.516	5	N12	0.623	1	N50	0.624	1	N95	0.650	2
N74	0.516	5	N13	0.623	1	N51	0.624	1	N15	0.652	2
N76	0.516	5	N14	0.624	1	N52	0.624	1	N98	0.654	2
N77	0.516	5	N16	0.624	1	N53	0.624	1	N9	0.657	2
N78	0.516	5	N17	0.624	1	N54	0.624	1	N10	0.657	2
N79	0.516	5	N18	0.624	1	N55	0.624	1	N11	0.657	2
N75	0.525	5	N19	0.624	1	N56	0.624	1	N101	0.657	2
N5	0.530	5	N20	0.624	1	N57	0.624	1	N99	0.664	2
N82	0.552	3	N21	0.624	1	N59	0.624	1	N105	0.678	4
N83	0.552	3	N22	0.624	1	N60	0.624	1	N8	0.692	4
N84	0.552	3	N23	0.624	1	N61	0.624	1			
N85	0.552	3	N24	0.624	1	N62	0.624	1			
N86	0.552	3	N25	0.624	1	N63	0.624	1			
N6	0.560	3	N27	0.624	1	N64	0.624	1			
N32	0.560	3	N28	0.624	1	N65	0.624	1			
N36	0.560	3	N30	0.624	1	N26	0.624	1			
N58	0.560	3	N31	0.624	1	N44	0.624	1			
N1	0.561	3	N33	0.624	1	N46	0.625	1			
N103	0.562	3	N34	0.624	1	N29	0.625	1			
N104	0.578	3	N35	0.624	1	N66	0.625	1			
N81	0.582	3	N37	0.624	1	N67	0.625	1			
N102	0.584	3	N38	0.624	1	N100	0.626	1			
N3	0.590	3	N39	0.624	1	N87	0.629	1			
N96	0.594	1	N40	0.624	1	N92	0.640	2			
N97	0.594	1	N41	0.624	1	N88	0.650	2			

The spatial distribution B_{17_2} of the amino acid D over the protein sequence

N2 is random since the HE of B_{17_2} is turned out to be 0.501. The largest cluster 1 contains 60 protein sequences where the amino acid D is spread with positive trend as shown in the Table 18.

3.18. Hurst Exponent of B_{18_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{18}(E)$ the Hurst exponent(HE) for the 105 binary sequences B_{18_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 18. The HE of the binary representations of ordering of the amino acid A_{18} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_{18} is ranging from 0.525 to 0.724. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, eight clustered (C) are formed as presented in the Table 19.

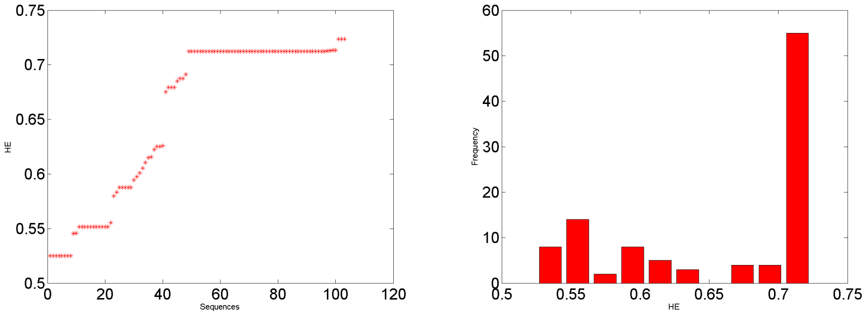


Figure 18: Plot of the HEs and corresponding histogram of all the binary sequences B_{18_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{18}(E)$.

There are 48 sequences in the cluster 1 where corresponding spatial distributions B_{18_j} of are positively trending with HE 0.712 exactly. Such a organized trend is certainly noteworthy. it is noted that, the non-essential amino acid E does not appear in the protein sequences N80 and N99.

Table 19: HE of 105 B_{18_j} for $j = 1, 2, \dots 105$ corresponding to the amino acid $A_{18}(E)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N80	*	3	N86	0.588	8	N25	0.712	1	N60	0.712	1
N99	*	3	N105	0.595	8	N27	0.712	1	N61	0.712	1
N88	0.525	4	N102	0.597	8	N28	0.712	1	N62	0.712	1
N89	0.525	4	N6	0.601	2	N29	0.712	1	N63	0.712	1
N90	0.525	4	N8	0.606	2	N30	0.712	1	N64	0.712	1
N91	0.525	4	N103	0.610	2	N31	0.712	1	N65	0.712	1
N92	0.525	4	N98	0.615	2	N33	0.712	1	N67	0.712	1
N93	0.525	4	N104	0.616	2	N34	0.712	1	N26	0.712	1
N94	0.525	4	N7	0.622	6	N35	0.712	1	N13	0.713	1
N95	0.525	4	N5	0.625	6	N37	0.712	1	N12	0.713	1
N1	0.545	7	N3	0.625	6	N38	0.712	1	N14	0.713	1
N15	0.546	7	N87	0.626	6	N39	0.712	1	N66	0.713	1
N70	0.551	7	N81	0.675	5	N40	0.712	1	N32	0.724	1
N69	0.552	7	N9	0.679	5	N41	0.712	1	N36	0.724	1
N71	0.552	7	N10	0.679	5	N42	0.712	1	N58	0.724	1
N72	0.552	7	N11	0.679	5	N43	0.712	1			
N73	0.552	7	N68	0.685	5	N45	0.712	1			
N74	0.552	7	N96	0.688	5	N46	0.712	1			
N75	0.552	7	N97	0.688	5	N47	0.712	1			
N76	0.552	7	N4	0.691	5	N48	0.712	1			
N77	0.552	7	N44	0.712	1	N49	0.712	1			
N78	0.552	7	N16	0.712	1	N50	0.712	1			
N79	0.552	7	N17	0.712	1	N51	0.712	1			
N101	0.555	7	N18	0.712	1	N52	0.712	1			
N2	0.580	8	N19	0.712	1	N53	0.712	1			
N100	0.583	8	N20	0.712	1	N54	0.712	1			
N82	0.588	8	N21	0.712	1	N55	0.712	1			
N83	0.588	8	N22	0.712	1	N56	0.712	1			
N84	0.588	8	N23	0.712	1	N57	0.712	1			
N85	0.588	8	N24	0.712	1	N59	0.712	1			

3.19. Hurst Exponent of B_{19_j} for $j = 1, 2, \dots 105$ and Classification

For the amino acid $A_{19}(K)$ the Hurst exponent(HE) for the 105 binary sequences B_{19_j} for $j = 1, 2, \dots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 19. The HE of the binary representations of ordering of the amino acid A_{19} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_{19} is ranging from 0.483 to 0.695 with standard deviation 0.111. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, eight clustered (C) are formed as presented in the Table 20.

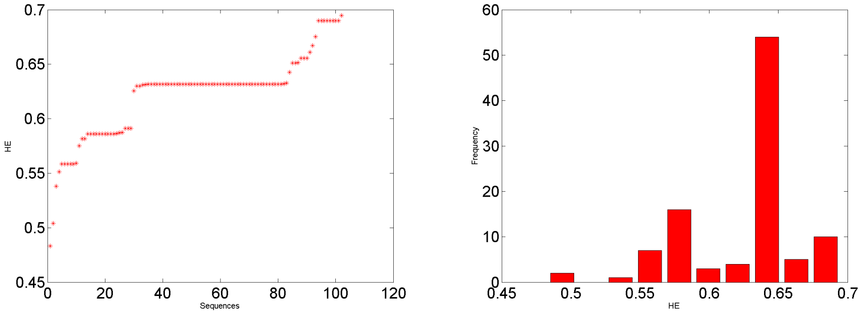


Figure 19: Plot of the HEs and corresponding histogram of all the binary sequences B_{19_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{19}(K)$.

Table 20: HE of 105 B_{19_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{19}(K)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N80	*	2	N10	0.591	8	N41	0.632	3	N32	0.655	7
N81	*	2	N11	0.591	8	N42	0.632	3	N36	0.655	7
N99	*	2	N4	0.626	3	N43	0.632	3	N58	0.655	7
N103	0.483	6	N54	0.630	3	N45	0.632	3	N98	0.661	7
N104	0.504	6	N57	0.630	3	N46	0.632	3	N101	0.667	5
N100	0.538	1	N12	0.631	3	N47	0.632	3	N7	0.675	5
N5	0.551	1	N26	0.631	3	N48	0.632	3	N88	0.690	4
N82	0.558	1	N14	0.632	3	N49	0.632	3	N89	0.690	4
N83	0.558	1	N16	0.632	3	N50	0.632	3	N90	0.690	4
N84	0.558	1	N17	0.632	3	N51	0.632	3	N91	0.690	4
N85	0.558	1	N18	0.632	3	N52	0.632	3	N92	0.690	4
N86	0.558	1	N19	0.632	3	N53	0.632	3	N93	0.690	4
N6	0.559	1	N20	0.632	3	N55	0.632	3	N94	0.690	4
N15	0.575	8	N21	0.632	3	N56	0.632	3	N95	0.690	4
N105	0.582	8	N22	0.632	3	N59	0.632	3	N3	0.695	4
N1	0.582	8	N23	0.632	3	N60	0.632	3			
N69	0.586	8	N24	0.632	3	N61	0.632	3			
N71	0.586	8	N25	0.632	3	N62	0.632	3			
N72	0.586	8	N27	0.632	3	N63	0.632	3			
N73	0.586	8	N28	0.632	3	N64	0.632	3			
N74	0.586	8	N29	0.632	3	N65	0.632	3			
N75	0.586	8	N30	0.632	3	N66	0.632	3			
N76	0.586	8	N31	0.632	3	N67	0.632	3			
N77	0.586	8	N33	0.632	3	N44	0.632	3			
N78	0.586	8	N34	0.632	3	N13	0.632	3			
N79	0.586	8	N35	0.632	3	N102	0.633	3			
N70	0.586	8	N37	0.632	3	N8	0.643	3			
N87	0.587	8	N38	0.632	3	N96	0.651	7			
N2	0.587	8	N39	0.632	3	N97	0.651	7			
N9	0.591	8	N40	0.632	3	N68	0.652	7			

The protein sequences N80, N81 and N99 are free from one of the amino acid K. The spatial organization of the amino acid K over the protein sequence N103 is negatively trending as it's HE $0.483 < 0.5$. There 58 protein sequences in the cluster 3 where the amino acid K is spatially distributed with positive trend as shown in the the Table 20. It is noted that the spatial distribution $B_{19_{104}}$ is random as the HE of N104 is turned out to be 0.5.

3.20. Hurst Exponent of B_{20_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{19}(R)$ the Hurst exponent(HE) for the 105 binary sequences B_{20_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the HE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 20. The HE of the binary representations of ordering of the amino acid A_{20} over all the primary protein sequences would reveal the autocorrelation of the amino acid.

Here the HE of the 105 binary representation of the amino acid A_{20} is ranging from 0.498 to 0.697 with standard deviation 0.0904. Based on the HEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, eight clustered (C) are formed as presented in the Table 21.

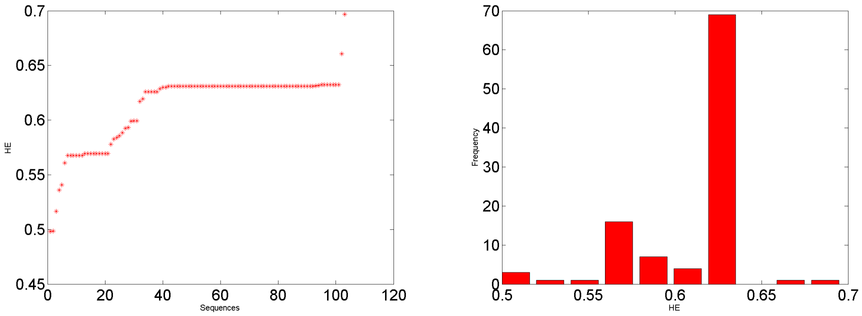


Figure 20: Plot of the HEs and corresponding histogram of all the binary sequences B_{20_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{20}(R)$.

Table 21: HE of 105 B_{20_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{20}(R)$.

Seq	HE	C	Seq	HE	C	Seq	HE	C	Seq	HE	C
N81	*	3	N68	0.599	7	N34	0.631	1	N32	0.631	1
N99	*	3	N96	0.599	7	N35	0.631	1	N36	0.631	1
N98	0.498	4	N97	0.599	7	N37	0.631	1	N58	0.631	1
N102	0.498	4	N87	0.617	1	N38	0.631	1	N13	0.631	1
N100	0.517	4	N104	0.620	1	N40	0.631	1	N12	0.631	1
N15	0.536	2	N105	0.626	1	N41	0.631	1	N54	0.632	1
N2	0.541	2	N9	0.626	1	N42	0.631	1	N88	0.632	1
N70	0.561	2	N10	0.626	1	N43	0.631	1	N89	0.632	1
N82	0.568	2	N11	0.626	1	N45	0.631	1	N90	0.632	1
N83	0.568	2	N95	0.626	1	N46	0.631	1	N91	0.632	1
N84	0.568	2	N8	0.629	1	N47	0.631	1	N92	0.632	1
N85	0.568	2	N31	0.630	1	N48	0.631	1	N93	0.632	1
N86	0.568	2	N39	0.630	1	N49	0.631	1	N94	0.632	1
N6	0.568	2	N26	0.631	1	N50	0.631	1	N3	0.661	6
N69	0.570	2	N14	0.631	1	N51	0.631	1	N80	0.697	5
N72	0.570	2	N16	0.631	1	N52	0.631	1			
N73	0.570	2	N17	0.631	1	N53	0.631	1			
N74	0.570	2	N18	0.631	1	N55	0.631	1			
N75	0.570	2	N19	0.631	1	N56	0.631	1			
N76	0.570	2	N20	0.631	1	N57	0.631	1			
N77	0.570	2	N21	0.631	1	N59	0.631	1			
N78	0.570	2	N22	0.631	1	N60	0.631	1			
N79	0.570	2	N23	0.631	1	N61	0.631	1			
N71	0.578	8	N24	0.631	1	N62	0.631	1			
N101	0.583	8	N25	0.631	1	N63	0.631	1			
N4	0.584	8	N27	0.631	1	N64	0.631	1			
N1	0.586	8	N28	0.631	1	N65	0.631	1			
N103	0.588	8	N29	0.631	1	N66	0.631	1			
N7	0.592	7	N30	0.631	1	N67	0.631	1			
N5	0.593	7	N33	0.631	1	N44	0.631	1			

The conditionally essential amino acid R is not at all present in the protein sequences N81 and N99 and consequently the HE is not enumerable. There are two sequences N98 and N102 having positive trending spatial representation of the amino acid R. The cluster 1 contains 70 protein sequences where the amino acid R is positively trended, spatially.As usual there other clusters containing positive autocorrelated spatial representations of the amino acid R.

3.21. A Collective Views of the HEs

Following we have listed the protein sequences of different lengths ranging from 13 to 419, which does not contain some amino acid(s) as listed in the following Table 22.

Table 22: Absence of amino acids on various SARS-CoV2 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

The protein sequence N99 of length 13 does not contains the amino acids C, H, M, P, T, W, Y, E, K and R which are of different types of essential, conditionally essential as well as non essential. The largest sequences N88, N89, N90, N91, N92, N93, N94, N95 of length 419 do not contain the amino acid C. It is noted that the amino acid M is present over all the proteins except N99 which of the smallest length 13. Also it is observed that the essential amino acids L, M, F and V are present in all the proteins of SARS-CoV2. The non essential amino acids A, D, N and S is also present in all the protein sequences. It is noted that all the six conditionally essential amino acids are not found essential for all the proteins of SARS-CoV2. Proteins that are of greater than 419 contains all the twenty amino acids. It is reported that the presence of the amino acid I, G including V is of primordial importance through in this study we found N99 which does not contain I, N68, N81 do not contain G.

It is found that the amino acid H is randomly spatially distributed over the proteins 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 the protein N80 and N102. Also the amino acid L is distributed over the protein N102 randomly. The only amino acid K which is randomly spread over the protein N104. There are other sequences N98 and N102 where the amino acid R is spread with negatively trend ($HE < 0.5$). Also the amino acid K, Y, S, Q, N and F are negatively trending over the protein sequences N103, N80, N7, N100, N2 and N5 respectively. Therefore the amino acid C, G, P, T, W and E is spread over all the 105 proteins with positive autocorrelation (positively trending).

Here we explore the correlation (of trending behaviours) of the distribution of the amino acids over 105 proteins of SARS-CoV2. Following is 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.

Table 23: 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

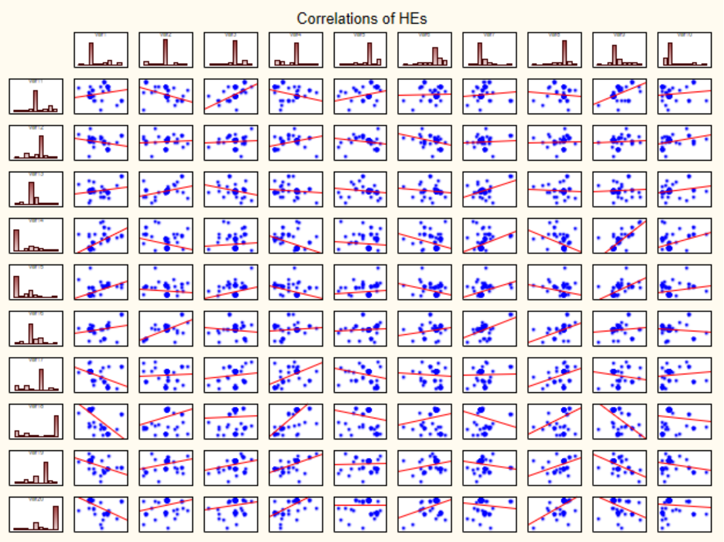


Figure 21: Plot of the correlation of the HEs of the amino acids distribution distinct pairwise.

The spatial distribution of the amino acid A with the same of the amino acids Q, T, V, W, and Y are turned out to be positively correlated based on HEs as shown in the Table 23. Likewise, the the HEs of the spatial distribution of the amino acid C is positively correlated with the S, T, Y, D, E, K and R. Similarly, the positive correlations of the spatial distributions of the amino acids F, G, H, I, L, M, N and P with the spatial distribution of the other amino acids are established in the correlation matrix, Table 23. The correlation based on HEs of the spatial distribution is also shown through graphs in the Fig 21. It is worthy mentioning that in the correlation matrix in the Table 23, the negative correlations of the spatial distribution of the proteins are also shown.

As an example of the correlation (the correlation coefficient r: 0.443) of the spatial distribution (autocorrelation) of the amino acid M with the spatial distribution of the amino acid Y is given below in the Fig. 22.

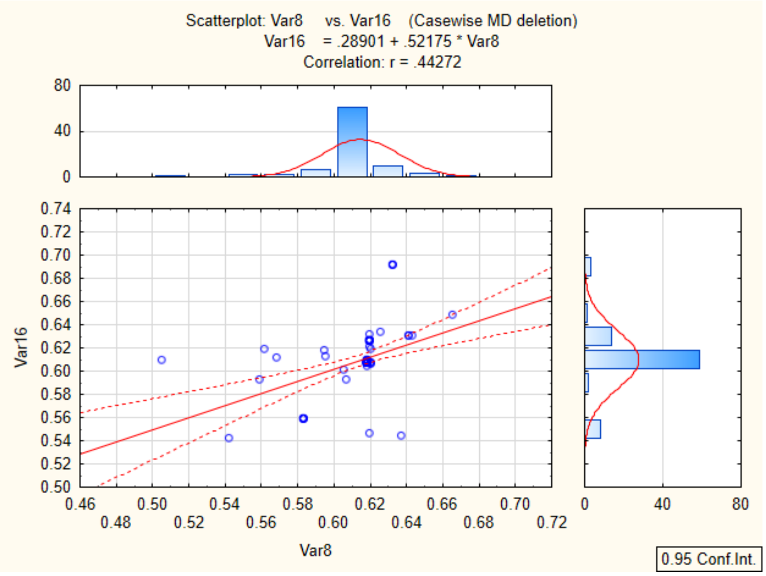


Figure 22: Plot of the correlation of the HEs of the amino acid M with that of L.

Now we move on the amount of uncertainty/certainty of presence of amino acids over the protein sequences in the following subsections.

3.22. Shannon Entropy of B_{1_j} *for* $j = 1, 2, \ldots 105$ *and Classification*

For the amino acid $A_1(A)$ the Shannon Entropy (SE) for the 105 binary sequences B_{1_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 23. The SE of the binary representations of ordering of the amino acid A_1 over all the primary protein sequences would reveal the amount of uncertainty of the amino acid.

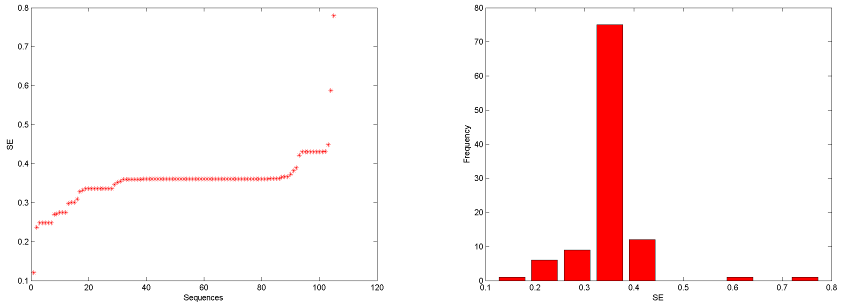


Figure 23: Plot of the SEs and corresponding histogram of all the binary sequences B_{1_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_1(A)$.

Here the SE of the 105 binary representation of the amino acid A_1 is ranging from 0.121 to 0.779 with standard deviation 0.0679. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clustered (C) are formed as presented in the Table 24.

It is found that the SE of the spatial distribution of amino acid A_1 in the protein sequence N68 is 0.121 which indicates the amount of uncertainly is lowest as compare to the others. The cluster 4 as well the the cluster 1 contains almost all the protein sequences where the SEs are less than 0.5 which says that the presence and absence over the protein sequences are mostly certain. Also amount of uncertainly is high for the proteins N3 and N99 of length 198 and 13 respectively.

Table 24: SE of 105 B_{1_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_1(A)$.

Seq	SE	C	Seq	SE	C	Seq	SE	C	Seq	SE	C
N68	0.121	3	N105	0.355	1	N42	0.361	1	N87	0.382	1
N15	0.237	4	N20	0.360	1	N45	0.361	1	N5	0.390	1
N82	0.248	4	N22	0.360	1	N46	0.361	1	N8	0.422	1
N83	0.248	4	N48	0.360	1	N47	0.361	1	N88	0.431	1
N84	0.248	4	N50	0.360	1	N49	0.361	1	N89	0.431	1
N85	0.248	4	N59	0.360	1	N51	0.361	1	N90	0.431	1
N86	0.248	4	N61	0.360	1	N52	0.361	1	N91	0.431	1
N104	0.270	4	N65	0.360	1	N53	0.361	1	N92	0.431	1
N81	0.271	4	N26	0.361	1	N54	0.361	1	N93	0.431	1
N9	0.275	4	N14	0.361	1	N55	0.361	1	N94	0.431	1
N10	0.275	4	N16	0.361	1	N56	0.361	1	N95	0.431	1
N11	0.275	4	N17	0.361	1	N57	0.361	1	N98	0.431	1
N80	0.297	4	N18	0.361	1	N60	0.361	1	N2	0.448	1
N96	0.300	4	N19	0.361	1	N62	0.361	1	N3	0.588	5
N97	0.300	4	N21	0.361	1	N63	0.361	1	N99	0.779	2
N102	0.310	4	N23	0.361	1	N64	0.361	1			
N103	0.328	1	N24	0.361	1	N66	0.361	1			
N74	0.332	1	N25	0.361	1	N67	0.361	1			
N69	0.336	1	N27	0.361	1	N44	0.361	1			
N71	0.336	1	N28	0.361	1	N100	0.361	1			
N72	0.336	1	N30	0.361	1	N13	0.361	1			
N73	0.336	1	N31	0.361	1	N12	0.361	1			
N75	0.336	1	N33	0.361	1	N29	0.361	1			
N76	0.336	1	N34	0.361	1	N43	0.361	1			
N77	0.336	1	N35	0.361	1	N6	0.362	1			
N78	0.336	1	N37	0.361	1	N101	0.362	1			
N79	0.336	1	N38	0.361	1	N58	0.366	1			
N70	0.336	1	N39	0.361	1	N32	0.366	1			
N4	0.346	1	N40	0.361	1	N36	0.366	1			
N1	0.352	1	N41	0.361	1	N7	0.373	1			

3.23. Shannon Entropy of B_{2_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_2(C)$ the Shannon Entropy (SE) for the 105 binary sequences B_{2_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 24. The SE of the binary representations of ordering of the amino acid A_2 over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

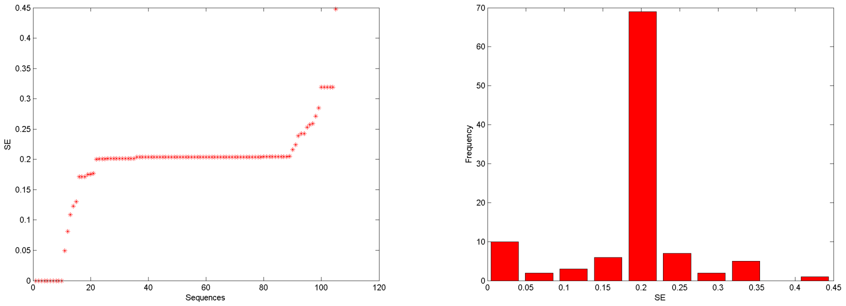


Figure 24: Plot of the SEs and corresponding histogram of all the binary sequences B_{2_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_2(C)$.

Here the SE of the 105 binary representation of the amino acid A_2 is ranging from 0 to 0.448 with standard deviation 0.0765. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, seven clusters (C) are formed as presented in the Table 25.

It is found that the SE of the spatial distribution of amino acid A_2 in the protein sequence N68, N88, N89, N90, N91, N92, N93, N94, N95 and N99 is zero which indicates the amount of uncertainly is zero i.e. the absolutely absence of the amino acid $A_2(C)$ over these proteins. The other clusters contains rest all protein sequences where the spatial presence of the amino acid over the protein sequences are having low uncertainly (high certainty) as shown in the Table 25.

Table 25: SE of 105 B_{2_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_2(C)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N68	0.000	1	N76	0.201	4	N51	0.204	4	N103	0.224	5
N88	0.000	1	N77	0.201	4	N52	0.204	4	N102	0.239	5
N89	0.000	1	N78	0.201	4	N53	0.204	4	N96	0.242	5
N90	0.000	1	N79	0.201	4	N54	0.204	4	N97	0.242	5
N91	0.000	1	N70	0.202	4	N55	0.204	4	N6	0.253	5
N92	0.000	1	N14	0.204	4	N56	0.204	4	N5	0.257	5
N93	0.000	1	N16	0.204	4	N57	0.204	4	N1	0.259	5
N94	0.000	1	N17	0.204	4	N59	0.204	4	N81	0.271	5
N95	0.000	1	N18	0.204	4	N60	0.204	4	N87	0.285	5
N99	0.000	1	N19	0.204	4	N61	0.204	4	N82	0.319	6
N15	0.050	2	N20	0.204	4	N62	0.204	4	N83	0.319	6
N3	0.081	2	N21	0.204	4	N63	0.204	4	N84	0.319	6
N104	0.109	3	N23	0.204	4	N64	0.204	4	N85	0.319	6
N105	0.123	3	N24	0.204	4	N65	0.204	4	N86	0.319	6
N8	0.130	3	N27	0.204	4	N66	0.204	4	N2	0.448	7
N9	0.171	3	N28	0.204	4	N67	0.204	4			
N10	0.171	3	N29	0.204	4	N44	0.204	4			
N11	0.171	3	N30	0.204	4	N13	0.204	4			
N4	0.175	3	N33	0.204	4	N12	0.204	4			
N80	0.176	3	N34	0.204	4	N26	0.204	4			
N98	0.177	3	N35	0.204	4	N22	0.204	4			
N100	0.200	4	N37	0.204	4	N25	0.204	4			
N32	0.201	4	N38	0.204	4	N31	0.204	4			
N36	0.201	4	N43	0.204	4	N39	0.204	4			
N58	0.201	4	N45	0.204	4	N40	0.204	4			
N69	0.201	4	N46	0.204	4	N41	0.204	4			
N71	0.201	4	N47	0.204	4	N42	0.204	4			
N73	0.201	4	N48	0.204	4	N7	0.204	4			
N74	0.201	4	N49	0.204	4	N72	0.205	4			
N75	0.201	4	N50	0.204	4	N101	0.216	5			

3.24. Shannon Entropy of B_{3_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_3(F)$ the Shannon Entropy (SE) for the 105 binary sequences B_{3_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 25. The SE of the binary representations of ordering of the amino acid A_3 over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

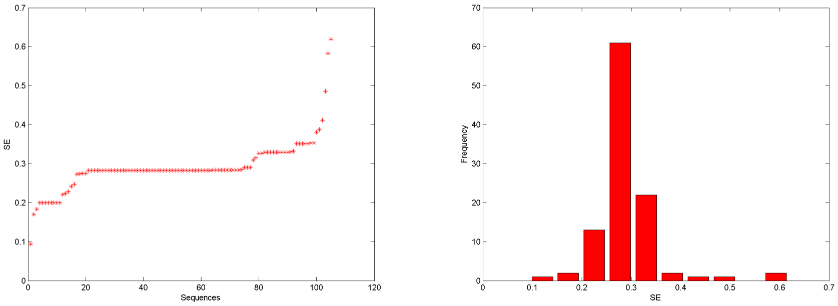


Figure 25: Plot of the SEs and corresponding histogram of all the binary sequences B_{3_j} for $j = 1, 2, \dots 105$ corresponding to the amino acid $A_3(F)$.

Table 26: SE of 105 B_{3_j} for $j = 1, 2, \dots 105$ corresponding to the amino acid $A_3(F)$.

Seq	SE	C	Seq	SE	C	Seq	SE	C	Seq	SE	C
N103	0.094	4	N21	0.283	2	N67	0.283	2	N104	0.330	7
N3	0.170	5	N23	0.283	2	N44	0.283	2	N100	0.332	7
N15	0.183	5	N25	0.283	2	N13	0.283	2	N82	0.351	7
N88	0.200	5	N28	0.283	2	N20	0.284	2	N83	0.351	7
N89	0.200	5	N29	0.283	2	N22	0.284	2	N84	0.351	7
N90	0.200	5	N30	0.283	2	N24	0.284	2	N85	0.351	7
N91	0.200	5	N31	0.283	2	N27	0.284	2	N86	0.351	7
N92	0.200	5	N35	0.283	2	N33	0.284	2	N96	0.353	7
N93	0.200	5	N37	0.283	2	N34	0.284	2	N97	0.353	7
N94	0.200	5	N38	0.283	2	N40	0.284	2	N7	0.381	1
N95	0.200	5	N39	0.283	2	N59	0.284	2	N101	0.387	1
N98	0.221	5	N41	0.283	2	N62	0.284	2	N87	0.411	1
N2	0.224	5	N43	0.283	2	N26	0.284	2	N80	0.485	6
N105	0.228	5	N45	0.283	2	N8	0.284	2	N81	0.583	8
N5	0.242	5	N46	0.283	2	N9	0.290	2	N99	0.619	3
N4	0.247	2	N47	0.283	2	N10	0.290	2			
N6	0.274	2	N48	0.283	2	N11	0.290	2			
N36	0.274	2	N49	0.283	2	N102	0.310	7			
N32	0.275	2	N50	0.283	2	N1	0.315	7			
N58	0.275	2	N52	0.283	2	N72	0.326	7			
N42	0.282	2	N53	0.283	2	N78	0.326	7			
N51	0.282	2	N54	0.283	2	N69	0.329	7			
N60	0.282	2	N55	0.283	2	N71	0.329	7			
N12	0.283	2	N56	0.283	2	N73	0.329	7			
N68	0.283	2	N57	0.283	2	N74	0.329	7			
N14	0.283	2	N61	0.283	2	N75	0.329	7			
N16	0.283	2	N63	0.283	2	N76	0.329	7			
N17	0.283	2	N64	0.283	2	N77	0.329	7			
N18	0.283	2	N65	0.283	2	N79	0.329	7			
N19	0.283	2	N66	0.283	2	N70	0.330	7			

Here the SE of the 105 binary representation of the amino acid A_3 is ranging from 0.094 to 0.619 with standard deviation 0.0667. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, seven clusters (C) are formed as presented in the Table 26.

The SE is greater than 0.5 for the binary representations of the protein A_3 over the proteins N81 and N99 and consequently the amount of uncertainty is lowering. It is noted that these two sequences are very small in length. Other clusters as usual contains rest all protein sequences where the spatial presence of the amino acid A_3 over the protein sequences are having low uncertainly (high certainty) as shown in the Table 26.

3.25. Shannon Entropy of B_{4_j} for $j = 1, 2, \dots 105$ and Classification

For the amino acid $A_4(G)$ the Shannon Entropy (SE) for the 105 binary sequences B_{4_j} for $j = 1, 2, \dots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 26. The SE of the binary representations of ordering of the amino acid A_4 over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

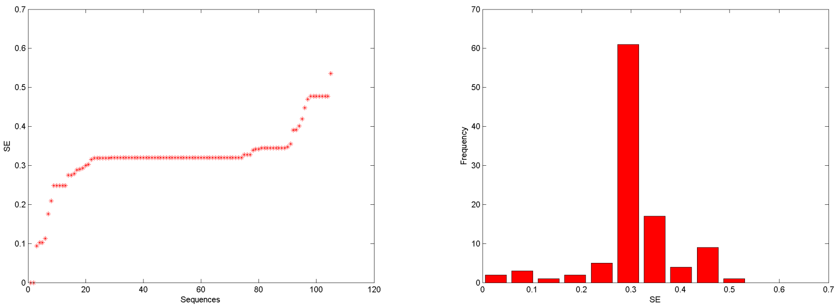


Figure 26: Plot of the SEs and corresponding histogram of all the binary sequences B_{4_j} for $j = 1, 2, \dots 105$ corresponding to the amino acid $A_4(G)$.

Here the SE of the 105 binary representation of the amino acid A_4 is ranging from 0 to 0.536 with standard deviation 0.0852. Based on the SEs of the binary

doi:10.20944/preprints202004.0034.v1

sequences all these 105 primary protein sequences of SARS-CoV2, six clusters (C) are formed as presented in the Table 27.

Table 27: SE of 105 B_{4_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_4(G)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N68	0.000	4	N18	0.320	1	N52	0.320	1	N105	0.355	6
N81	0.000	4	N19	0.320	1	N55	0.320	1	N6	0.391	5
N103	0.094	4	N20	0.320	1	N56	0.320	1	N99	0.391	5
N96	0.102	4	N21	0.320	1	N57	0.320	1	N98	0.401	5
N97	0.102	4	N22	0.320	1	N59	0.320	1	N102	0.419	5
N3	0.113	4	N23	0.320	1	N61	0.320	1	N2	0.448	3
N80	0.176	4	N24	0.320	1	N62	0.320	1	N95	0.470	3
N87	0.210	1	N25	0.320	1	N63	0.320	1	N88	0.477	3
N82	0.248	1	N27	0.320	1	N64	0.320	1	N89	0.477	3
N83	0.248	1	N28	0.320	1	N65	0.320	1	N90	0.477	3
N84	0.248	1	N29	0.320	1	N66	0.320	1	N91	0.477	3
N85	0.248	1	N30	0.320	1	N67	0.320	1	N92	0.477	3
N86	0.248	1	N31	0.320	1	N44	0.320	1	N93	0.477	3
N9	0.275	1	N33	0.320	1	N26	0.320	1	N94	0.477	3
N11	0.275	1	N34	0.320	1	N32	0.327	6	N15	0.536	2
N100	0.279	1	N35	0.320	1	N36	0.327	6			
N4	0.288	1	N37	0.320	1	N58	0.327	6			
N10	0.290	1	N38	0.320	1	N8	0.339	6			
N101	0.294	1	N39	0.320	1	N104	0.342	6			
N5	0.300	1	N40	0.320	1	N77	0.342	6			
N7	0.303	1	N41	0.320	1	N69	0.345	6			
N1	0.315	1	N42	0.320	1	N71	0.345	6			
N14	0.319	1	N43	0.320	1	N72	0.345	6			
N53	0.319	1	N45	0.320	1	N73	0.345	6			
N54	0.319	1	N46	0.320	1	N74	0.345	6			
N60	0.319	1	N47	0.320	1	N76	0.345	6			
N13	0.319	1	N48	0.320	1	N78	0.345	6			
N12	0.319	1	N49	0.320	1	N79	0.345	6			
N16	0.320	1	N50	0.320	1	N70	0.345	6			
N17	0.320	1	N51	0.320	1	N75	0.348	6			

It is noted that the amino acid $A_4(G)$ does not present in the primary protein sequences of N68 and N81 and consequently the SE is turned up as zero saying that there is no uncertainty. In the cluster 4 there are other sequences N103, N96, N97, N3 and N80 having 0.0839 as the centre cluster 4 of the SEs and so the amount of uncertainty of presence of the amino acid $A_4(G)$ is low. All these sequences are smaller than 198 in length. The largest cluster 1 which contains most of the protein sequences where the the presence of the amino acid over the protein sequences is almost certain with HE at centre of the cluster 0.3088. The other clusters also contains similar proteins where the amino acid A_3 is present with almost certainty.

3.26. Shannon Entropy of B_{5_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_5(H)$ the Shannon Entropy (SE) for the 105 binary sequences B_{5_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 27. The SE of the binary representations of ordering of the amino acid A_5 over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

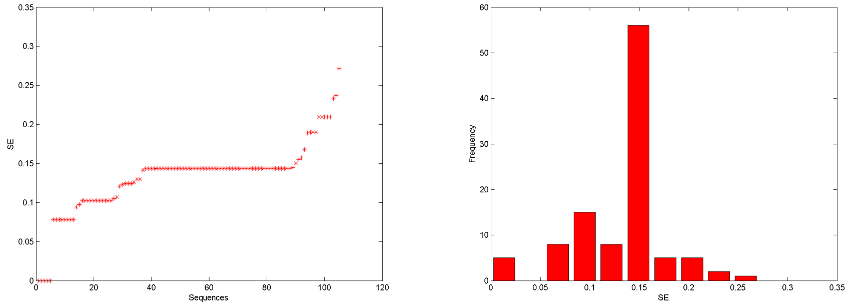


Figure 27: Plot of the SEs and corresponding histogram of all the binary sequences B_{5_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_5(H)$.

Here the SE of the 105 binary representation of the amino acid A_5 is ranging from 0 to 0.271 with standard deviation 0.0459. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, eight clusters (C) are formed as presented in the Table 28.

The amino acid $A_5(H)$ does not present in the amino acid sequence of the proteins N3, N80, N97, N98 and N99 and so the SE is turned up as zero implying there is no uncertainly at all, naturally. The SE of the spatial representation of the presence and absence of the amino acid A_5 over the proteins N88, N89, N90, N91, N92, N94 and N95 (belong to cluster 4) of length 419 is 0.078 and hence the spatial distriburion is more certainly/orderly. The largest cluster 1 contains protein sequence of various lengths including highest length where the presence of the amino acid is clearly having least of uncertainty (SE: 0.144). It is noted that the protein N81 of length 43 having highest SE 0.271 having having certainty of appearance of the amino acid A_5.

Table 28: SE of 105 B_{5_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_5(H)$.

Seq	SE	C	Seq	SE	C	Seq	SE	C	Seq	SE	C
N3	0.000	3	N32	0.124	8	N37	0.144	1	N8	0.155	6
N80	0.000	3	N36	0.124	8	N38	0.144	1	N102	0.157	6
N97	0.000	3	N58	0.124	8	N39	0.144	1	N87	0.168	6
N98	0.000	3	N104	0.126	1	N40	0.144	1	N100	0.189	2
N99	0.000	3	N4	0.130	1	N41	0.144	1	N9	0.190	2
N88	0.078	4	N7	0.130	1	N42	0.144	1	N10	0.190	2
N89	0.078	4	N5	0.141	1	N43	0.144	1	N11	0.190	2
N90	0.078	4	N49	0.143	1	N45	0.144	1	N82	0.210	2
N91	0.078	4	N64	0.143	1	N46	0.144	1	N83	0.210	2
N92	0.078	4	N13	0.143	1	N47	0.144	1	N84	0.210	2
N93	0.078	4	N12	0.143	1	N48	0.144	1	N85	0.210	2
N94	0.078	4	N26	0.144	1	N50	0.144	1	N86	0.210	2
N95	0.078	4	N14	0.144	1	N51	0.144	1	N1	0.233	7
N103	0.094	8	N16	0.144	1	N52	0.144	1	N15	0.237	7
N79	0.097	8	N17	0.144	1	N53	0.144	1	N81	0.271	5
N96	0.102	8	N18	0.144	1	N54	0.144	1			
N69	0.102	8	N19	0.144	1	N55	0.144	1			
N71	0.102	8	N20	0.144	1	N56	0.144	1			
N72	0.102	8	N22	0.144	1	N57	0.144	1			
N73	0.102	8	N23	0.144	1	N59	0.144	1			
N74	0.102	8	N24	0.144	1	N60	0.144	1			
N75	0.102	8	N25	0.144	1	N61	0.144	1			
N76	0.102	8	N27	0.144	1	N62	0.144	1			
N77	0.102	8	N28	0.144	1	N63	0.144	1			
N78	0.102	8	N29	0.144	1	N65	0.144	1			
N70	0.102	8	N30	0.144	1	N66	0.144	1			
N101	0.105	8	N31	0.144	1	N67	0.144	1			
N6	0.107	8	N33	0.144	1	N44	0.144	1			
N68	0.121	8	N34	0.144	1	N21	0.145	1			
N105	0.123	8	N35	0.144	1	N2	0.150	1			

3.27. Shannon Entropy of B_{6_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_6(I)$ the Shannon Entropy (SE) for the 105 binary sequences B_{6_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 28. The SE of the binary representations of ordering of the amino acid A_6 over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

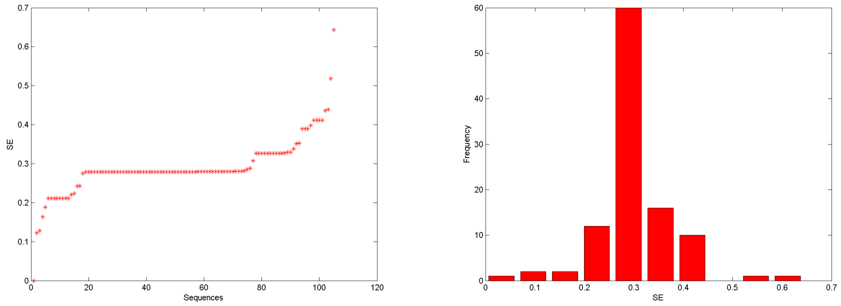


Figure 28: Plot of the SEs and corresponding histogram of all the binary sequences B_{6_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_6(I)$.

Here the SE of the 105 binary representation of the amino acid A_6 is ranging from 0 to 0.644 with standard deviation 0.0749. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, eight clusters (C) are formed as presented in the Table 29.

The amino acid $A_6(I)$ does not present in the protein N99 (of smallest length 13) and so the SE is turned up as zero implying there is no uncertainly at all, as usual. The rest all the clusters except 6 contain only those proteins where the amino acid $A_6(I)$ is spread spatially with certainty whereas the cluster 6 contains two sequences N81 (of length 43) and N68 (of length 61) where the absence of the amino acid dominate the presence with certainty.

Table 29: SE of 105 B_{6_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_6(I)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N99	0.000	1	N31	0.279	4	N18	0.280	4	N6	0.338	5
N15	0.122	1	N33	0.279	4	N19	0.280	4	N87	0.351	5
N98	0.128	1	N34	0.279	4	N20	0.280	4	N104	0.353	5
N103	0.164	7	N37	0.279	4	N25	0.280	4	N9	0.389	3
N2	0.188	7	N39	0.279	4	N27	0.280	4	N10	0.389	3
N88	0.211	7	N40	0.279	4	N29	0.280	4	N11	0.389	3
N89	0.211	7	N41	0.279	4	N35	0.280	4	N80	0.398	3
N90	0.211	7	N42	0.279	4	N38	0.280	4	N83	0.411	6
N91	0.211	7	N43	0.279	4	N51	0.280	4	N84	0.411	6
N92	0.211	7	N45	0.279	4	N59	0.280	4	N85	0.411	6
N93	0.211	7	N46	0.279	4	N28	0.281	4	N86	0.411	6
N94	0.211	7	N47	0.279	4	N32	0.281	4	N8	0.437	6
N95	0.211	7	N48	0.279	4	N36	0.281	4	N82	0.439	6
N100	0.221	8	N49	0.279	4	N58	0.282	4	N81	0.519	2
N102	0.223	8	N50	0.279	4	N4	0.286	4	N68	0.644	2
N96	0.242	8	N52	0.279	4	N3	0.289	4			
N97	0.242	8	N53	0.279	4	N101	0.308	5			
N1	0.275	4	N54	0.279	4	N69	0.326	5			
N55	0.279	4	N56	0.279	4	N71	0.326	5			
N61	0.279	4	N57	0.279	4	N72	0.326	5			
N65	0.279	4	N60	0.279	4	N73	0.326	5			
N5	0.279	4	N62	0.279	4	N74	0.326	5			
N26	0.279	4	N63	0.279	4	N75	0.326	5			
N14	0.279	4	N64	0.279	4	N76	0.326	5			
N16	0.279	4	N66	0.279	4	N77	0.326	5			
N21	0.279	4	N67	0.279	4	N78	0.326	5			
N22	0.279	4	N44	0.279	4	N79	0.326	5			
N23	0.279	4	N13	0.279	4	N7	0.327	5			
N24	0.279	4	N12	0.280	4	N105	0.329	5			
N30	0.279	4	N17	0.280	4	N70	0.330	5			

3.28. Shannon Entropy of B_{7_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_7(L)$ the Shannon Entropy (SE) for the 105 binary sequences B_{7_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 29. The SE of the binary representations of ordering of the amino acid A_7 over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

Here the SE of the 105 binary representation of the amino acid A_7 is ranging from 0 to 0.644 with standard deviation 0.0749. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, six clusters (C) are formed as presented in the Table 30.

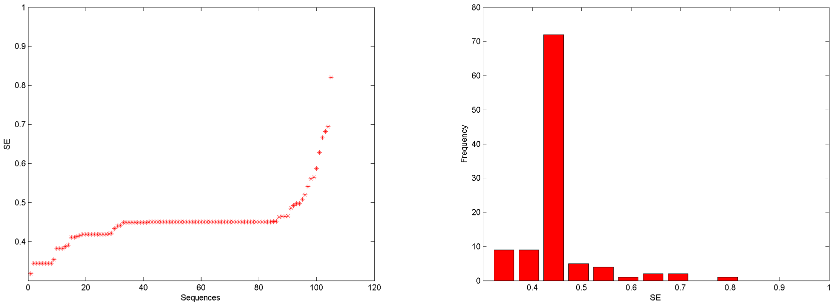


Figure 29: Plot of the SEs and corresponding histogram of all the binary sequences B_{7_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_7(L)$.

Table 30: SE of 105 B_{7_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_7(L)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N2	0.318	3	N3	0.439	1	N50	0.450	1	N80	0.485	1
N89	0.345	3	N4	0.442	1	N51	0.450	1	N105	0.492	1
N90	0.345	3	N20	0.450	1	N52	0.450	1	N10	0.497	1
N91	0.345	3	N24	0.450	1	N53	0.450	1	N11	0.497	1
N92	0.345	3	N27	0.450	1	N54	0.450	1	N9	0.508	4
N93	0.345	3	N33	0.450	1	N55	0.450	1	N15	0.520	4
N94	0.345	3	N34	0.450	1	N56	0.450	1	N87	0.541	4
N95	0.345	3	N47	0.450	1	N57	0.450	1	N68	0.561	4
N88	0.354	3	N59	0.450	1	N61	0.450	1	N98	0.565	4
N82	0.382	5	N62	0.450	1	N64	0.450	1	N101	0.588	4
N85	0.382	5	N44	0.450	1	N65	0.450	1	N8	0.629	6
N86	0.382	5	N6	0.450	1	N67	0.450	1	N96	0.665	6
N1	0.388	5	N14	0.450	1	N13	0.450	1	N103	0.682	6
N99	0.391	5	N16	0.450	1	N12	0.450	1	N97	0.694	6
N83	0.411	5	N17	0.450	1	N26	0.451	1	N81	0.820	2
N84	0.411	5	N18	0.450	1	N23	0.451	1			
N5	0.413	5	N19	0.450	1	N25	0.451	1			
N104	0.415	5	N21	0.450	1	N30	0.451	1			
N69	0.419	5	N22	0.450	1	N37	0.451	1			
N71	0.419	5	N28	0.450	1	N40	0.451	1			
N72	0.419	5	N29	0.450	1	N41	0.451	1			
N73	0.419	5	N31	0.450	1	N60	0.451	1			
N74	0.419	5	N35	0.450	1	N63	0.451	1			
N75	0.419	5	N38	0.450	1	N66	0.451	1			
N76	0.419	5	N39	0.450	1	N42	0.451	1			
N77	0.419	5	N43	0.450	1	N102	0.452	1			
N79	0.419	5	N45	0.450	1	N7	0.463	1			
N70	0.419	5	N46	0.450	1	N32	0.464	1			
N78	0.422	5	N48	0.450	1	N58	0.464	1			
N100	0.433	1	N49	0.450	1	N36	0.465	1			

The proteins N2 (of length 138) and N89, N90, N91, N92, N93, N94 and N95 of length 419 in the cluster 3, where the amino acid $A_7(L)$ is spread spatially with certainty. The clusters 1 and 5 contain majority of the protein sequences where the presence of the amino acid is spread over the proteins with almost certainty whereas the clusters 2 and 6 contain five protein sequences where the absence of the amino acid A_7 is spread with almost certainty.

3.29. Shannon Entropy of B_{8_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_8(M)$ the Shannon Entropy (SE) for the 105 binary sequences B_{8_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 30. The SE of the binary representations of ordering of the amino acid A_8 over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

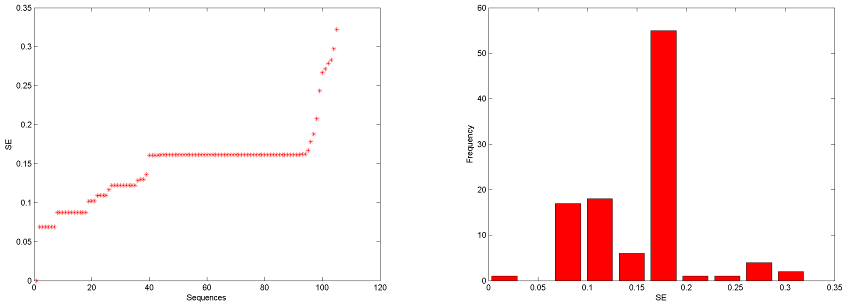


Figure 30: Plot of the SEs and corresponding histogram of all the binary sequences B_{8_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_8(M)$.

Here the SE of the 105 binary representation of the amino acid A_8 is ranging from 0 to 0.322 with standard deviation 0.0749. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, six clusters (C) are formed as presented in the Table 31.

The amino acid $A_8(M)$ does not present in the sequence N99 which is of smallest length and so the amount of uncertaintly is zero as found in the Table 31. The cluster 1 including others contains most of the proteins of SARS-CoV2 where the amino acid is present all over the proteins of various lengths with almost certainty which is validated by its SE which is 0.162.

Table 31: SE of 105 B_{8_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_8(M)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N99	0.000	3	N91	0.123	4	N31	0.162	1	N67	0.162	1
N82	0.069	6	N92	0.123	4	N33	0.162	1	N44	0.162	1
N83	0.069	6	N93	0.123	4	N34	0.162	1	N32	0.162	1
N84	0.069	6	N94	0.123	4	N35	0.162	1	N36	0.162	1
N85	0.069	6	N95	0.123	4	N37	0.162	1	N1	0.167	1
N86	0.069	6	N98	0.128	4	N38	0.162	1	N100	0.178	1
N87	0.069	6	N7	0.130	4	N39	0.162	1	N2	0.188	1
N69	0.087	5	N8	0.130	4	N40	0.162	1	N102	0.208	1
N71	0.087	5	N4	0.136	4	N41	0.162	1	N105	0.244	2
N72	0.087	5	N52	0.161	1	N42	0.162	1	N3	0.267	2
N73	0.087	5	N59	0.161	1	N43	0.162	1	N81	0.271	2
N74	0.087	5	N13	0.161	1	N45	0.162	1	N103	0.279	2
N75	0.087	5	N12	0.161	1	N46	0.162	1	N68	0.283	2
N76	0.087	5	N58	0.161	1	N47	0.162	1	N80	0.297	2
N77	0.087	5	N26	0.162	1	N48	0.162	1	N101	0.322	2
N78	0.087	5	N14	0.162	1	N49	0.162	1			
N79	0.087	5	N16	0.162	1	N50	0.162	1			
N70	0.087	5	N17	0.162	1	N51	0.162	1			
N5	0.102	5	N18	0.162	1	N53	0.162	1			
N96	0.102	5	N19	0.162	1	N54	0.162	1			
N97	0.102	5	N20	0.162	1	N55	0.162	1			
N104	0.109	4	N21	0.162	1	N56	0.162	1			
N9	0.110	4	N22	0.162	1	N57	0.162	1			
N10	0.110	4	N23	0.162	1	N60	0.162	1			
N11	0.110	4	N24	0.162	1	N61	0.162	1			
N6	0.116	4	N25	0.162	1	N62	0.162	1			
N15	0.122	4	N27	0.162	1	N63	0.162	1			
N88	0.123	4	N28	0.162	1	N64	0.162	1			
N89	0.123	4	N29	0.162	1	N65	0.162	1			
N90	0.123	4	N30	0.162	1	N66	0.162	1			

3.30. Shannon Entropy of B_{9_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_9(N)$ the Shannon Entropy (SE) for the 105 binary sequences B_{9_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 31. The SE of the binary representations of ordering of the amino acid A_9 over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

Here the SE of the 105 binary representation of the amino acid A_9 is ranging from 0.121 to 0.562. Based on the SEs of the binary sequences all these

105 primary protein sequences of SARS-CoV2, six clusters (C) are formed as presented in the Table 32.

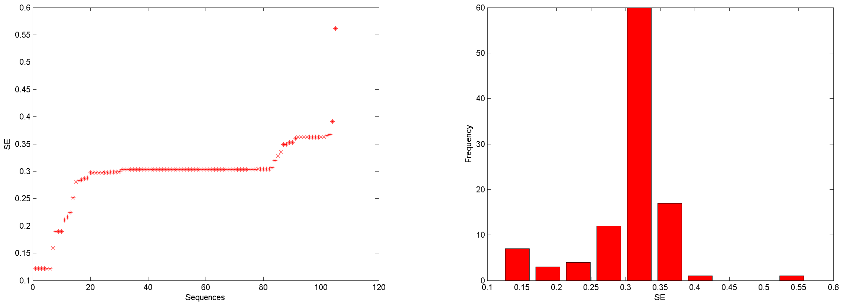


Figure 31: Plot of the SEs and corresponding histogram of all the binary sequences B_{9_j} for $j = 1, 2, \dots 105$ corresponding to the amino acid $A_9(M)$.

The cluster 3 contains one protein N80 where the spatial distribution $B_{9_{80}}$ has the SE 0.562 which says the absence of the the amino acid A_9 over the protein is without uncertainty. It is noted that total number of amino acid A_9 placed over the 38 length protein N80 is 5. The other five clusters contains rest 104 proteins where the amino acid A_9 is spread with certainty as the HE is less than 0.5.

Table 32: SE of 105 B_{9_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_9(N)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N82	0.121	5	N51	0.303	1	N49	0.304	1	N102	0.361	6
N83	0.121	5	N26	0.304	1	N50	0.304	1	N71	0.363	6
N84	0.121	5	N14	0.304	1	N52	0.304	1	N72	0.363	6
N85	0.121	5	N17	0.304	1	N53	0.304	1	N73	0.363	6
N86	0.121	5	N18	0.304	1	N54	0.304	1	N74	0.363	6
N87	0.121	5	N19	0.304	1	N55	0.304	1	N75	0.363	6
N81	0.159	2	N20	0.304	1	N56	0.304	1	N76	0.363	6
N9	0.190	2	N21	0.304	1	N59	0.304	1	N77	0.363	6
N10	0.190	2	N22	0.304	1	N60	0.304	1	N78	0.363	6
N11	0.190	2	N23	0.304	1	N61	0.304	1	N79	0.363	6
N15	0.211	2	N24	0.304	1	N62	0.304	1	N70	0.363	6
N101	0.216	2	N25	0.304	1	N63	0.304	1	N69	0.366	6
N103	0.224	2	N27	0.304	1	N64	0.304	1	N105	0.368	6
N7	0.251	1	N28	0.304	1	N65	0.304	1	N99	0.391	4
N6	0.280	1	N29	0.304	1	N66	0.304	1	N80	0.562	3
N104	0.283	1	N30	0.304	1	N67	0.304	1			
N8	0.284	1	N31	0.304	1	N44	0.304	1			
N5	0.286	1	N33	0.304	1	N13	0.304	1			
N2	0.288	1	N34	0.304	1	N12	0.304	1			
N88	0.297	1	N35	0.304	1	N16	0.304	1			
N90	0.297	1	N37	0.304	1	N46	0.304	1			
N91	0.297	1	N38	0.304	1	N57	0.304	1			
N92	0.297	1	N39	0.304	1	N89	0.307	1			
N93	0.297	1	N40	0.304	1	N4	0.320	1			
N94	0.297	1	N41	0.304	1	N100	0.328	1			
N95	0.297	1	N42	0.304	1	N98	0.335	6			
N32	0.299	1	N43	0.304	1	N68	0.349	6			
N36	0.299	1	N45	0.304	1	N3	0.350	6			
N58	0.299	1	N47	0.304	1	N96	0.353	6			
N1	0.300	1	N48	0.304	1	N97	0.353	6			

3.31. Shannon Entropy of B_{10_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{10}(P)$ the Shannon Entropy (SE) for the 105 binary sequences B_{10_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 32. The SE of the binary representations of ordering of the amino acid A_{10} over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

Here the SE of the 105 binary representation of the amino acid A_{10} is ranging from 0 to 0.644 with standard deviation 0.0749. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, six clusters (C) are formed as presented in the Table 33.

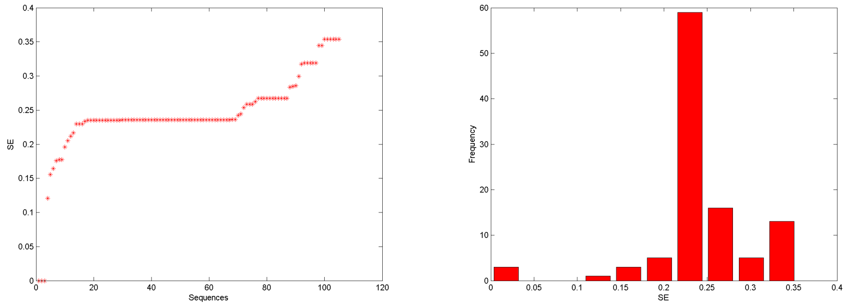


Figure 32: Plot of the SEs and corresponding histogram of all the binary sequences B_{10_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{10}(P)$.

Table 33: SE of 105 B_{10_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{10}(P)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N81	0.000	4	N16	0.236	2	N57	0.236	2	N98	0.299	6
N99	0.000	4	N17	0.236	2	N59	0.236	2	N2	0.318	6
N103	0.000	4	N18	0.236	2	N60	0.236	2	N82	0.319	6
N68	0.121	3	N19	0.236	2	N62	0.236	2	N83	0.319	6
N8	0.155	3	N20	0.236	2	N64	0.236	2	N84	0.319	6
N101	0.164	3	N21	0.236	2	N65	0.236	2	N85	0.319	6
N80	0.176	3	N22	0.236	2	N44	0.236	2	N86	0.319	6
N96	0.177	3	N24	0.236	2	N13	0.236	2	N91	0.345	1
N97	0.177	3	N27	0.236	2	N12	0.236	2	N94	0.345	1
N3	0.196	3	N28	0.236	2	N7	0.242	2	N88	0.354	1
N100	0.205	2	N29	0.236	2	N104	0.244	2	N89	0.354	1
N105	0.212	2	N30	0.236	2	N102	0.254	5	N90	0.354	1
N6	0.216	2	N31	0.236	2	N9	0.259	5	N92	0.354	1
N32	0.229	2	N33	0.236	2	N10	0.259	5	N93	0.354	1
N36	0.229	2	N34	0.236	2	N11	0.259	5	N95	0.354	1
N58	0.229	2	N35	0.236	2	N15	0.262	5			
N4	0.233	2	N38	0.236	2	N69	0.267	5			
N26	0.235	2	N39	0.236	2	N71	0.267	5			
N23	0.235	2	N43	0.236	2	N72	0.267	5			
N25	0.235	2	N45	0.236	2	N73	0.267	5			
N37	0.235	2	N46	0.236	2	N74	0.267	5			
N40	0.235	2	N47	0.236	2	N75	0.267	5			
N41	0.235	2	N48	0.236	2	N76	0.267	5			
N42	0.235	2	N49	0.236	2	N77	0.267	5			
N56	0.235	2	N50	0.236	2	N78	0.267	5			
N61	0.235	2	N51	0.236	2	N79	0.267	5			
N63	0.235	2	N52	0.236	2	N70	0.267	5			
N66	0.235	2	N53	0.236	2	N1	0.284	5			
N67	0.235	2	N54	0.236	2	N87	0.285	5			
N14	0.236	2	N55	0.236	2	N5	0.286	5			

The cluster 3 contains the proteins N81, N99 and N103 where the HE is turned out to be zero for the binary representations B_{10_j} for $j = 81, 99$ and 103 of the amino acid A_{10}. These proteins naturally absolutely free from the amino acid A_{10}. The cluster 1 contains all the protein sequences 419 where the amino acid A_{10} is present over the proteins with almost certainty.

3.32. Shannon Entropy of B_{11_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{11}(Q)$ the Shannon Entropy (SE) for the 105 binary sequences B_{11_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 33. The SE of the binary representations of ordering of the amino acid A_{11} over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

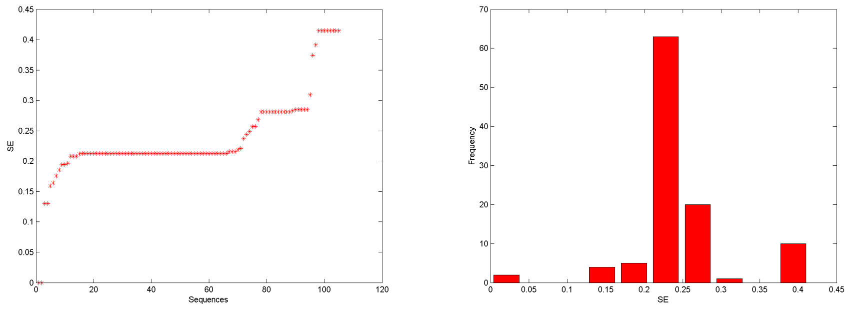


Figure 33: Plot of the SEs and corresponding histogram of all the binary sequences B_{11_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{11}(Q)$.

Here the SE of the 105 binary representation of the amino acid A_{11} is ranging from 0 to 0.415 with standard deviation 0.0749. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, six clusters (C) are formed as presented in the Table 34.

The cluster 4 contains the proteins N96 and N97 of length where the HE is turned out to be zero for the binary representations B_{10_j} for $j = 96$ and 97 of the amino acid A_{11}. It is noted that these two proteins naturally absolutely free from the amino acid A_{11}. All the rest clusters contain all the protein sequences where the amino acid A_{10} is present over the proteins with almost certainty.

Table 34: SE of 105 B_{11_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{11}(Q)$.

Seq	SE	C	Seq	SE	C	Seq	SE	C	Seq	SE	C
N96	0.000	4	N31	0.213	3	N65	0.213	3	N83	0.285	1
N97	0.000	4	N33	0.213	3	N66	0.213	3	N84	0.285	1
N7	0.130	3	N34	0.213	3	N67	0.213	3	N85	0.285	1
N8	0.130	3	N35	0.213	3	N44	0.213	3	N86	0.285	1
N81	0.159	3	N37	0.213	3	N13	0.213	3	N3	0.310	1
N101	0.164	3	N38	0.213	3	N12	0.213	3	N103	0.374	6
N80	0.176	3	N39	0.213	3	N32	0.215	3	N99	0.391	6
N6	0.185	3	N40	0.213	3	N36	0.215	3	N88	0.415	2
N5	0.194	3	N41	0.213	3	N58	0.215	3	N89	0.415	2
N100	0.195	3	N42	0.213	3	N4	0.219	3	N90	0.415	2
N1	0.196	3	N43	0.213	3	N98	0.221	3	N91	0.415	2
N9	0.208	3	N45	0.213	3	N15	0.237	5	N92	0.415	2
N10	0.208	3	N46	0.213	3	N105	0.244	5	N93	0.415	2
N11	0.208	3	N47	0.213	3	N87	0.248	5	N94	0.415	2
N21	0.212	3	N48	0.213	3	N2	0.257	5	N95	0.415	2
N26	0.213	3	N49	0.213	3	N104	0.257	5			
N14	0.213	3	N50	0.213	3	N102	0.268	1			
N16	0.213	3	N51	0.213	3	N69	0.281	1			
N17	0.213	3	N52	0.213	3	N71	0.281	1			
N18	0.213	3	N53	0.213	3	N72	0.281	1			
N19	0.213	3	N54	0.213	3	N73	0.281	1			
N20	0.213	3	N55	0.213	3	N74	0.281	1			
N22	0.213	3	N56	0.213	3	N75	0.281	1			
N23	0.213	3	N57	0.213	3	N76	0.281	1			
N24	0.213	3	N59	0.213	3	N77	0.281	1			
N25	0.213	3	N60	0.213	3	N78	0.281	1			
N27	0.213	3	N61	0.213	3	N79	0.281	1			
N28	0.213	3	N62	0.213	3	N70	0.281	1			
N29	0.213	3	N63	0.213	3	N68	0.283	1			
N30	0.213	3	N64	0.213	3	N82	0.285	1			

3.33. Shannon Entropy of B_{12_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{12}(S)$ the Shannon Entropy (SE) for the 105 binary sequences B_{12_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 34. The SE of the binary representations of ordering of the amino acid A_{12} over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

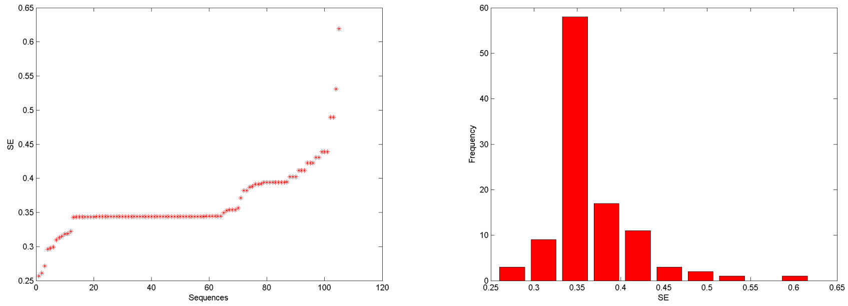


Figure 34: Plot of the SEs and corresponding histogram of all the binary sequences B_{12_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{12}(S)$.

Table 35: SE of 105 B_{12_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{12}(S)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N2	0.257	3	N29	0.344	5	N54	0.345	5	N82	0.411	1
N98	0.261	3	N31	0.344	5	N56	0.345	5	N83	0.411	1
N81	0.271	3	N33	0.344	5	N60	0.345	5	N85	0.411	1
N102	0.296	3	N34	0.344	5	N61	0.345	5	N88	0.423	1
N80	0.297	3	N35	0.344	5	N68	0.349	5	N89	0.423	1
N1	0.300	3	N37	0.344	5	N5	0.353	5	N93	0.423	1
N15	0.310	3	N38	0.344	5	N32	0.354	5	N90	0.431	1
N6	0.313	3	N39	0.344	5	N36	0.354	5	N92	0.431	1
N100	0.315	3	N41	0.344	5	N58	0.354	5	N91	0.439	1
N104	0.319	5	N42	0.344	5	N8	0.357	5	N94	0.439	1
N87	0.319	5	N43	0.344	5	N4	0.371	5	N95	0.439	1
N101	0.322	5	N45	0.344	5	N84	0.382	1	N96	0.490	4
N26	0.343	5	N46	0.344	5	N86	0.382	1	N97	0.490	4
N14	0.344	5	N47	0.344	5	N3	0.387	1	N103	0.531	4
N16	0.344	5	N48	0.344	5	N7	0.388	1	N99	0.619	2
N20	0.344	5	N49	0.344	5	N71	0.391	1			
N28	0.344	5	N50	0.344	5	N76	0.391	1			
N30	0.344	5	N52	0.344	5	N105	0.392	1			
N40	0.344	5	N53	0.344	5	N69	0.394	1			
N62	0.344	5	N55	0.344	5	N72	0.394	1			
N12	0.344	5	N57	0.344	5	N73	0.394	1			
N17	0.344	5	N59	0.344	5	N74	0.394	1			
N18	0.344	5	N63	0.344	5	N75	0.394	1			
N19	0.344	5	N64	0.344	5	N77	0.394	1			
N21	0.344	5	N65	0.344	5	N78	0.394	1			
N22	0.344	5	N66	0.344	5	N79	0.394	1			
N23	0.344	5	N67	0.344	5	N70	0.394	1			
N24	0.344	5	N44	0.344	5	N9	0.402	1			
N25	0.344	5	N13	0.344	5	N10	0.402	1			
N27	0.344	5	N51	0.345	5	N11	0.402	1			

Here the SE of the 105 binary representation of the amino acid A_{12} is ranging

from 0.257 to 0.619 with standard deviation 0.0503. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clusters (C) are formed as presented in the Table 35.

The amino acid A_{10} is present over the all the proteins except N99 with almost certainty since the SE of the spatial distributions is turned out to be less than 0.5, as shown in the Table 35. The SE of the smallest lengthy protein N99 is greater than 0.5 which imply the absence of the amino acid is spread over the protein with certainty.

3.34. Shannon Entropy of B_{13_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{13}(T)$ the Shannon Entropy (SE) for the 105 binary sequences B_{13_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 35. The SE of the binary representations of ordering of the amino acid A_{13} over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

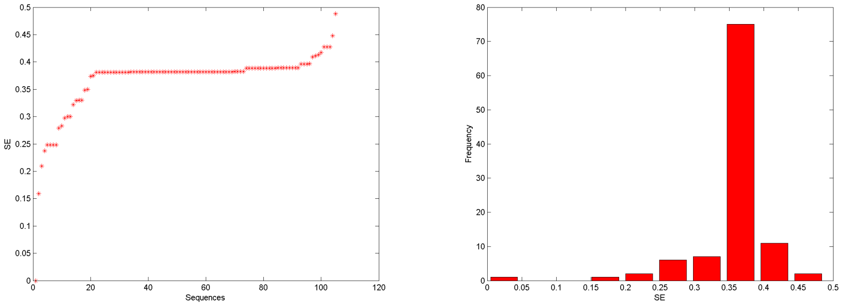


Figure 35: Plot of the SEs and corresponding histogram of all the binary sequences B_{13_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{13}(T)$.

Here the SE of the 105 binary representation of the amino acid A_{13} is ranging from 0 to 0.488 with standard deviation 0.0601. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clusters (C) are formed as presented in the Table 36.

Table 36: SE of 105 B_{13_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{13}(T)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N99	0.000	3	N35	0.381	2	N59	0.382	2	N94	0.389	2
N81	0.159	3	N38	0.381	2	N60	0.382	2	N95	0.389	2
N82	0.210	1	N63	0.381	2	N62	0.382	2	N32	0.396	2
N15	0.237	1	N26	0.382	2	N64	0.382	2	N36	0.396	2
N83	0.248	1	N16	0.382	2	N66	0.382	2	N58	0.396	2
N84	0.248	1	N21	0.382	2	N67	0.382	2	N102	0.397	2
N85	0.248	1	N22	0.382	2	N44	0.382	2	N7	0.409	5
N86	0.248	1	N23	0.382	2	N13	0.382	2	N87	0.411	5
N103	0.279	1	N24	0.382	2	N12	0.382	2	N5	0.413	5
N68	0.283	1	N30	0.382	2	N14	0.382	2	N4	0.417	5
N80	0.297	1	N31	0.382	2	N55	0.382	2	N9	0.427	5
N96	0.300	1	N33	0.382	2	N61	0.382	2	N10	0.427	5
N97	0.300	1	N37	0.382	2	N65	0.382	2	N11	0.427	5
N8	0.322	1	N39	0.382	2	N69	0.389	2	N2	0.448	5
N105	0.329	1	N40	0.382	2	N71	0.389	2	N98	0.488	4
N104	0.330	1	N41	0.382	2	N72	0.389	2			
N1	0.330	1	N42	0.382	2	N73	0.389	2			
N100	0.349	2	N43	0.382	2	N74	0.389	2			
N3	0.350	2	N45	0.382	2	N75	0.389	2			
N6	0.374	2	N46	0.382	2	N76	0.389	2			
N101	0.375	2	N47	0.382	2	N77	0.389	2			
N17	0.381	2	N48	0.382	2	N78	0.389	2			
N18	0.381	2	N49	0.382	2	N79	0.389	2			
N19	0.381	2	N50	0.382	2	N70	0.389	2			
N20	0.381	2	N51	0.382	2	N88	0.389	2			
N25	0.381	2	N52	0.382	2	N89	0.389	2			
N27	0.381	2	N53	0.382	2	N90	0.389	2			
N28	0.381	2	N54	0.382	2	N91	0.389	2			
N29	0.381	2	N56	0.382	2	N92	0.389	2			
N34	0.381	2	N57	0.382	2	N93	0.389	2			

The amino acid $A_{13}(T)$ is absent in the protein sequence of N99 and consequently the binary representation $B_{13_{99}}$ of presence and absence of the amino acid is absolutely a sequence with zeros without any uncertainty (SE=0) as shown in the Table 36. The rest proteins belonging to other clusters have the presence of the amino acid $A_{13}(T)$ with least amount of uncertainty as depicted in the Table 36.

3.35. Shannon Entropy of B_{14_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{14}(V)$ the Shannon Entropy (SE) for the 105 binary sequences B_{14_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 36. The SE of the binary representations of ordering of the amino acid A_{14} over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

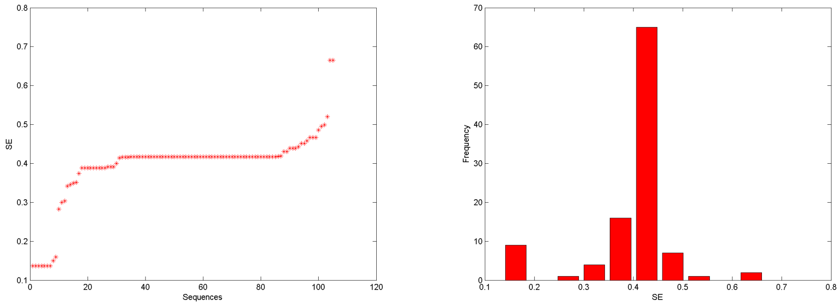


Figure 36: Plot of the SEs and corresponding histogram of all the binary sequences B_{14_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{14}(V)$.

Here the SE of the 105 binary representation of the amino acid A_{14} is ranging from 0.136 to 0.665 with standard deviation 0.0919. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clusters (C) are formed as presented in the Table 37.

The essential amino acid V is present all over the 105 porteins and hence none of the binary representations has the zero SE. The SE of the $B_{14_{99}}$ smallest lengthy protein including other proteins such as N74 and N77 of length 1273 is exactly 0.391 which imply the presence of the amino acid $A_1$4 over the proteins is with good certainty. The proteins N96 and N97 having length 75 have the the maximum about of uncertainty (SE:0.665) as compared to other proteins. The rest proteins belonging to other clusters have the presence of the amino acid A_{14} with certainty.

Table 37: SE of 105 B_{14_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{14}(V)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N88	0.136	4	N1	0.414	5	N42	0.417	5	N83	0.439	5
N89	0.136	4	N13	0.416	5	N43	0.417	5	N86	0.439	5
N90	0.136	4	N12	0.416	5	N45	0.417	5	N7	0.443	5
N91	0.136	4	N22	0.417	5	N46	0.417	5	N9	0.451	3
N93	0.136	4	N32	0.417	5	N47	0.417	5	N11	0.451	3
N94	0.136	4	N36	0.417	5	N49	0.417	5	N101	0.458	3
N95	0.136	4	N26	0.417	5	N51	0.417	5	N82	0.466	3
N92	0.150	4	N14	0.417	5	N52	0.417	5	N84	0.466	3
N81	0.159	4	N16	0.417	5	N54	0.417	5	N85	0.466	3
N68	0.283	1	N17	0.417	5	N55	0.417	5	N80	0.485	3
N98	0.299	1	N18	0.417	5	N56	0.417	5	N103	0.495	3
N8	0.303	1	N19	0.417	5	N57	0.417	5	N104	0.499	3
N105	0.342	1	N20	0.417	5	N60	0.417	5	N15	0.520	3
N2	0.346	1	N21	0.417	5	N62	0.417	5	N96	0.665	2
N3	0.350	1	N23	0.417	5	N64	0.417	5	N97	0.665	2
N87	0.351	1	N24	0.417	5	N66	0.417	5			
N6	0.374	1	N25	0.417	5	N67	0.417	5			
N69	0.389	1	N27	0.417	5	N44	0.417	5			
N71	0.389	1	N28	0.417	5	N48	0.418	5			
N72	0.389	1	N29	0.417	5	N50	0.418	5			
N73	0.389	1	N30	0.417	5	N53	0.418	5			
N75	0.389	1	N31	0.417	5	N59	0.418	5			
N76	0.389	1	N33	0.417	5	N61	0.418	5			
N78	0.389	1	N34	0.417	5	N63	0.418	5			
N79	0.389	1	N35	0.417	5	N65	0.418	5			
N70	0.389	1	N37	0.417	5	N58	0.418	5			
N99	0.391	1	N38	0.417	5	N4	0.419	5			
N74	0.391	1	N39	0.417	5	N5	0.430	5			
N77	0.391	1	N40	0.417	5	N102	0.431	5			
N100	0.400	5	N41	0.417	5	N10	0.439	5			

3.36. Shannon Entropy of B_{15_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{15}(W)$ the Shannon Entropy (SE) for the 105 binary sequences B_{15_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 37. The SE of the binary representations of ordering of the amino acid A_{15} over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

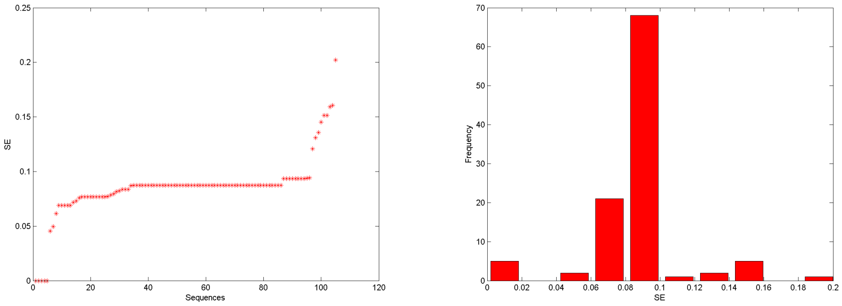


Figure 37: Plot of the SEs and corresponding histogram of all the binary sequences B_{15_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{15}(W)$.

Table 38: SE of 105 B_{15_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{15}(W)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N80	0.000	1	N32	0.084	2	N43	0.087	2	N92	0.093	2
N87	0.000	1	N36	0.084	2	N45	0.087	2	N93	0.093	2
N96	0.000	1	N58	0.084	2	N46	0.087	2	N94	0.093	2
N97	0.000	1	N6	0.087	2	N47	0.087	2	N95	0.093	2
N99	0.000	1	N26	0.087	2	N48	0.087	2	N7	0.094	2
N5	0.045	2	N14	0.087	2	N49	0.087	2	N103	0.094	2
N15	0.050	2	N16	0.087	2	N50	0.087	2	N68	0.121	5
N2	0.062	2	N17	0.087	2	N51	0.087	2	N9	0.131	5
N82	0.069	2	N18	0.087	2	N52	0.087	2	N1	0.136	5
N83	0.069	2	N19	0.087	2	N53	0.087	2	N101	0.145	4
N84	0.069	2	N20	0.087	2	N54	0.087	2	N10	0.152	4
N85	0.069	2	N21	0.087	2	N55	0.087	2	N11	0.152	4
N86	0.069	2	N22	0.087	2	N56	0.087	2	N81	0.159	4
N104	0.072	2	N23	0.087	2	N57	0.087	2	N105	0.161	4
N98	0.073	2	N24	0.087	2	N59	0.087	2	N8	0.202	3
N4	0.076	2	N25	0.087	2	N60	0.087	2			
N69	0.077	2	N27	0.087	2	N61	0.087	2			
N71	0.077	2	N28	0.087	2	N62	0.087	2			
N72	0.077	2	N29	0.087	2	N63	0.087	2			
N73	0.077	2	N30	0.087	2	N64	0.087	2			
N74	0.077	2	N31	0.087	2	N65	0.087	2			
N75	0.077	2	N33	0.087	2	N66	0.087	2			
N77	0.077	2	N34	0.087	2	N67	0.087	2			
N78	0.077	2	N35	0.087	2	N44	0.087	2			
N79	0.077	2	N37	0.087	2	N13	0.087	2			
N70	0.077	2	N38	0.087	2	N12	0.087	2			
N100	0.079	2	N39	0.087	2	N88	0.093	2			
N102	0.079	2	N40	0.087	2	N89	0.093	2			
N3	0.081	2	N41	0.087	2	N90	0.093	2			
N76	0.082	2	N42	0.087	2	N91	0.093	2			

Here the SE of the 105 binary representation of the amino acid A_{15} is ranging

doi:10.20944/preprints202004.0034.v1

from 0 to 0.202 with standard deviation 0.0919. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clusters (C) are formed as presented in the Table 38.

The cluster 1 contains five protein sequences where the amino acid A_{15} is absent and hence the SE is turned up as zero. Most of the proteins in the largest cluster 2 including other clusters contain the amino acid A_{15} spatially placed over with almost certainty.

3.37. *Shannon Entropy of* B_{16_j} *for* $j = 1, 2, \dots 105$ *and Classification*

For the amino acid $A_{16}(Y)$ the Shannon Entropy (SE) for the 105 binary sequences B_{16_j} for $j = 1, 2, \dots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 38. The SE of the binary representations of ordering of the amino acid A_{16} over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

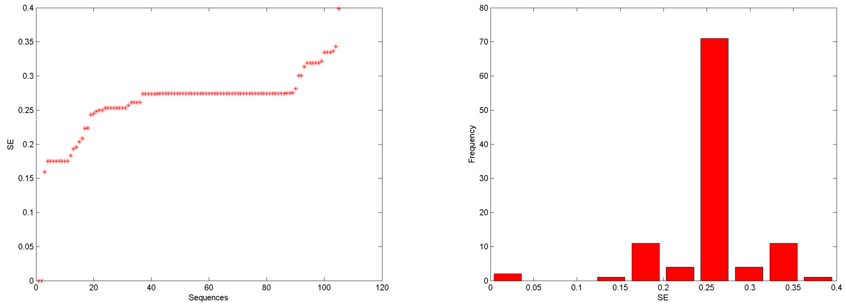


Figure 38: Plot of the SEs and corresponding histogram of all the binary sequences B_{16_j} for $j = 1, 2, \dots 105$ corresponding to the amino acid $A_{16}(Y)$.

Here the SE of the 105 binary representation of the amino acid A_{16} is ranging from 0 to 0.398 with standard deviation 0.055. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clusters (C) are formed as presented in the Table 39.

Table 39: SE of 105 B_{16_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{16}(Y)$.

Seq	SE	C	Seq	SE	C	Seq	SE	C	Seq	SE	C
N99	0.000	2	N78	0.253	4	N38	0.274	1	N96	0.300	1
N103	0.000	2	N79	0.257	4	N39	0.274	1	N97	0.300	1
N81	0.159	5	N98	0.261	4	N43	0.274	1	N5	0.314	3
N88	0.175	5	N32	0.262	4	N45	0.274	1	N82	0.319	3
N89	0.175	5	N36	0.262	4	N46	0.274	1	N83	0.319	3
N90	0.175	5	N58	0.262	4	N47	0.274	1	N84	0.319	3
N91	0.175	5	N26	0.274	1	N48	0.274	1	N85	0.319	3
N92	0.175	5	N22	0.274	1	N50	0.274	1	N86	0.319	3
N93	0.175	5	N25	0.274	1	N51	0.274	1	N101	0.322	3
N94	0.175	5	N40	0.274	1	N52	0.274	1	N9	0.335	3
N95	0.175	5	N41	0.274	1	N53	0.274	1	N10	0.335	3
N15	0.183	5	N42	0.274	1	N54	0.274	1	N11	0.335	3
N6	0.193	5	N14	0.274	1	N55	0.274	1	N100	0.336	3
N3	0.196	5	N16	0.274	1	N56	0.274	1	N7	0.343	3
N104	0.203	5	N17	0.274	1	N57	0.274	1	N80	0.398	3
N68	0.208	5	N18	0.274	1	N59	0.274	1			
N102	0.223	4	N19	0.274	1	N60	0.274	1			
N2	0.224	4	N20	0.274	1	N61	0.274	1			
N105	0.244	4	N21	0.274	1	N62	0.274	1			
N8	0.245	4	N23	0.274	1	N63	0.274	1			
N87	0.248	4	N24	0.274	1	N65	0.274	1			
N69	0.250	4	N27	0.274	1	N66	0.274	1			
N70	0.250	4	N28	0.274	1	N67	0.274	1			
N71	0.253	4	N29	0.274	1	N44	0.274	1			
N72	0.253	4	N30	0.274	1	N13	0.275	1			
N73	0.253	4	N31	0.274	1	N12	0.275	1			
N74	0.253	4	N33	0.274	1	N49	0.275	1			
N75	0.253	4	N34	0.274	1	N64	0.275	1			
N76	0.253	4	N35	0.274	1	N1	0.275	1			
N77	0.253	4	N37	0.274	1	N4	0.281	1			

The SE of the binary spatial representations of the proteins N99 and N103 is zero as the the amino acid does not present at all over these two proteins. Rest all the proteins have the amino acid A_{16} presence with certainty as presented in the Table 39.

3.38. Shannon Entropy of B_{17_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{17}(D)$ the Shannon Entropy (SE) for the 105 binary sequences B_{17_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 39. The SE of the binary representations of ordering of the amino acid A_{17} over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

doi:10.20944/preprints202004.0034.v1

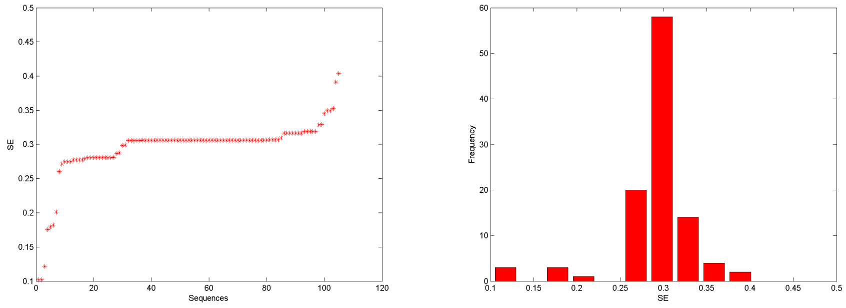


Figure 39: Plot of the SEs and corresponding histogram of all the binary sequences B_{17_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{17}(D)$.

Table 40: SE of 105 B_{17_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{17}(D)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N96	0.102	5	N98	0.299	1	N42	0.307	1	N94	0.317	1
N97	0.102	5	N29	0.306	1	N45	0.307	1	N95	0.317	1
N87	0.121	5	N43	0.306	1	N47	0.307	1	N82	0.319	1
N80	0.176	2	N46	0.306	1	N48	0.307	1	N83	0.319	1
N8	0.179	2	N66	0.306	1	N49	0.307	1	N84	0.319	1
N101	0.182	2	N67	0.306	1	N50	0.307	1	N85	0.319	1
N6	0.201	2	N26	0.306	1	N51	0.307	1	N86	0.319	1
N7	0.260	3	N14	0.307	1	N52	0.307	1	N103	0.328	1
N81	0.271	3	N16	0.307	1	N53	0.307	1	N105	0.329	1
N9	0.275	3	N17	0.307	1	N54	0.307	1	N1	0.345	4
N10	0.275	3	N18	0.307	1	N55	0.307	1	N68	0.349	4
N11	0.275	3	N19	0.307	1	N56	0.307	1	N3	0.350	4
N32	0.277	3	N20	0.307	1	N57	0.307	1	N104	0.353	4
N36	0.277	3	N21	0.307	1	N59	0.307	1	N99	0.391	4
N58	0.277	3	N22	0.307	1	N60	0.307	1	N100	0.404	4
N75	0.277	3	N23	0.307	1	N61	0.307	1			
N5	0.279	3	N24	0.307	1	N62	0.307	1			
N69	0.281	3	N25	0.307	1	N63	0.307	1			
N71	0.281	3	N27	0.307	1	N64	0.307	1			
N72	0.281	3	N28	0.307	1	N65	0.307	1			
N73	0.281	3	N30	0.307	1	N44	0.307	1			
N74	0.281	3	N31	0.307	1	N13	0.307	1			
N76	0.281	3	N33	0.307	1	N12	0.307	1			
N77	0.281	3	N34	0.307	1	N92	0.307	1			
N78	0.281	3	N35	0.307	1	N102	0.310	1			
N79	0.281	3	N37	0.307	1	N88	0.317	1			
N70	0.281	3	N38	0.307	1	N89	0.317	1			
N15	0.286	3	N39	0.307	1	N90	0.317	1			
N2	0.288	3	N40	0.307	1	N91	0.317	1			
N4	0.299	1	N41	0.307	1	N93	0.317	1			

Here the SE of the 105 binary representation of the amino acid A_{17} is ranging

from 0.102 to 0.404 with standard deviation 0.0448. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clusters (C) are formed as presented in the Table 40.

825 There are five clusters containing different proteins according to the closeness of the SE obtained for their spatial representations of the amino acids over the proteins. It is noted that the amino acid A_{17} is present over all the 105 proteins with almost certainty ($HEs < 0.5$).

3.39. Shannon Entropy of B_{18_j} for $j = 1, 2, \ldots 105$ and Classification

830 For the amino acid $A_{18}(E)$ the Shannon Entropy (SE) for the 105 binary sequences B_{18_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 40. The SE of the binary representations of ordering of the amino acid A_{18} over all the primary protein sequences would reveal the amount 835 of uncertainty of the presence or absence of the amino acid.

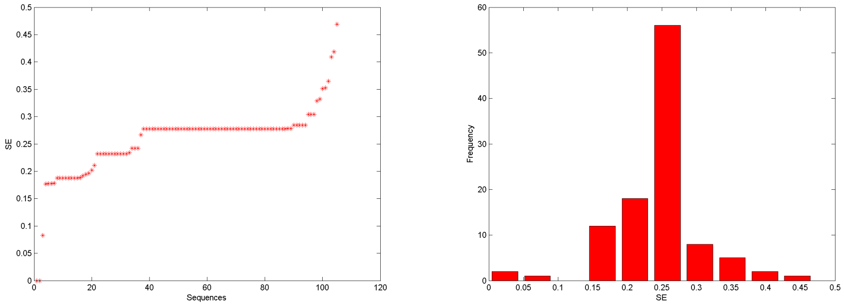


Figure 40: Plot of the SEs and corresponding histogram of all the binary sequences B_{18_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{18}(D)$.

Here the SE of the 105 binary representation of the amino acid A_{18} is ranging from 0 to 0.404 with standard deviation 0.0638. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clusters (C) are formed as presented in the Table 41.

Table 41: SE of 105 B_{18_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{18}(E)$.

Seq	SE	C	Seq	SE	C	Seq	SE	C	Seq	SE	C
N80	0.000	2	N79	0.232	5	N41	0.277	1	N83	0.285	1
N99	0.000	2	N70	0.232	5	N42	0.277	1	N84	0.285	1
N101	0.083	4	N5	0.234	5	N43	0.277	1	N85	0.285	1
N98	0.177	3	N9	0.242	5	N45	0.277	1	N86	0.285	1
N96	0.177	3	N10	0.242	5	N46	0.277	1	N32	0.304	1
N97	0.177	3	N11	0.242	5	N47	0.277	1	N36	0.304	1
N105	0.178	3	N3	0.267	1	N48	0.277	1	N58	0.304	1
N88	0.188	3	N26	0.277	1	N49	0.277	1	N103	0.328	1
N89	0.188	3	N16	0.277	1	N50	0.277	1	N4	0.332	1
N90	0.188	3	N17	0.277	1	N51	0.277	1	N87	0.351	1
N91	0.188	3	N18	0.277	1	N52	0.277	1	N104	0.353	1
N92	0.188	3	N19	0.277	1	N53	0.277	1	N81	0.365	1
N93	0.188	3	N20	0.277	1	N54	0.277	1	N68	0.409	1
N94	0.188	3	N21	0.277	1	N55	0.277	1	N6	0.418	1
N95	0.188	3	N22	0.277	1	N56	0.277	1	N15	0.469	1
N2	0.188	3	N23	0.277	1	N57	0.277	1			
N102	0.191	3	N24	0.277	1	N59	0.277	1			
N7	0.194	3	N25	0.277	1	N60	0.277	1			
N1	0.196	3	N27	0.277	1	N61	0.277	1			
N8	0.202	3	N28	0.277	1	N62	0.277	1			
N100	0.210	5	N29	0.277	1	N63	0.277	1			
N69	0.232	5	N30	0.277	1	N64	0.277	1			
N71	0.232	5	N31	0.277	1	N65	0.277	1			
N72	0.232	5	N33	0.277	1	N67	0.277	1			
N73	0.232	5	N34	0.277	1	N44	0.277	1			
N74	0.232	5	N35	0.277	1	N13	0.278	1			
N75	0.232	5	N37	0.277	1	N12	0.278	1			
N76	0.232	5	N38	0.277	1	N14	0.278	1			
N77	0.232	5	N39	0.277	1	N66	0.278	1			
N78	0.232	5	N40	0.277	1	N82	0.285	1			

The amino acid $A_{18}(E)$ is absent in the proteins N80 and N99 (belong to the cluster 2) with length 13 and 38 respectively consequently the SE of these two proteins for its binary spatial representations $B_{18_{80}}$ and $B_{18_{99}}$. The other four clusters containing different proteins according to the closeness of the SE obtained for their spatial representations of the amino acids over the proteins. It is noted that the amino acid A_{18} is present over all the 105 proteins with almost certainty ($HEs < 0.5$).

3.40. Shannon Entropy of B_{19_j} for $j = 1, 2, \ldots 105$ and Classification

For the amino acid $A_{19}(K)$ the Shannon Entropy (SE) for the 105 binary sequences B_{19_j} for $j = 1, 2, \ldots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 41. The SE of the binary representations of ordering of the amino acid A_{19} over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

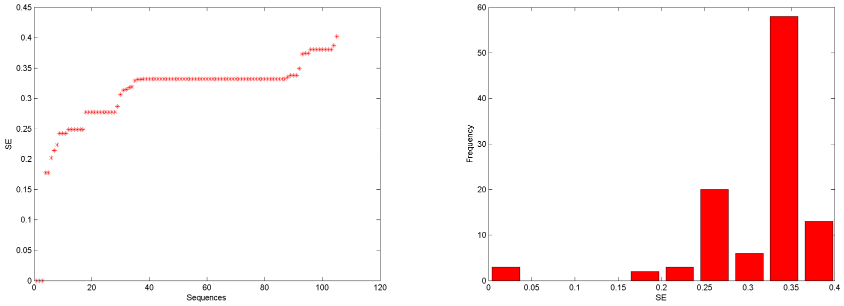


Figure 41: Plot of the SEs and corresponding histogram of all the binary sequences B_{19_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{19}(K)$.

Here the SE of the 105 binary representation of the amino acid A_{19} is ranging from 0 to 0.402. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clusters (C) are formed as presented in the Table 42.

Every term of the binary representations $B_{19_{80}}$, $B_{19_{81}}$ and $B_{19_{99}}$ of lengths 38, 43 and 13 respectively is zero and consequently the SE is turned out to be zero which implies the absence of the amino acid is without any uncertainty. The other proteins of the remaining clusters 1, 2, 4 and 5 have the presence of the amino acid A_{19} with almost certainty.

Table 42: SE of 105 B_{19_j} for $j = 1, 2, \dots 105$ corresponding to the amino acid $A_{19}(K)$.

Seq	SE	C	Seq	SE	C	Seq	SE	C	Seq	SE	C
N80	0.000	3	N5	0.314	1	N39	0.332	1	N58	0.338	1
N81	0.000	3	N1	0.315	1	N40	0.332	1	N68	0.349	1
N99	0.000	3	N2	0.318	1	N41	0.332	1	N4	0.373	4
N96	0.177	2	N87	0.319	1	N42	0.332	1	N104	0.374	4
N97	0.177	2	N105	0.329	1	N43	0.332	1	N103	0.374	4
N8	0.202	2	N54	0.331	1	N45	0.332	1	N88	0.381	4
N7	0.214	2	N57	0.331	1	N46	0.332	1	N89	0.381	4
N102	0.223	2	N12	0.332	1	N47	0.332	1	N90	0.381	4
N9	0.242	5	N26	0.332	1	N48	0.332	1	N91	0.381	4
N10	0.242	5	N14	0.332	1	N49	0.332	1	N92	0.381	4
N11	0.242	5	N16	0.332	1	N50	0.332	1	N93	0.381	4
N82	0.248	5	N17	0.332	1	N51	0.332	1	N94	0.381	4
N83	0.248	5	N18	0.332	1	N52	0.332	1	N95	0.381	4
N84	0.248	5	N19	0.332	1	N53	0.332	1	N3	0.387	4
N85	0.248	5	N20	0.332	1	N55	0.332	1	N6	0.402	4
N86	0.248	5	N21	0.332	1	N56	0.332	1			
N101	0.249	5	N22	0.332	1	N59	0.332	1			
N69	0.277	5	N23	0.332	1	N60	0.332	1			
N71	0.277	5	N24	0.332	1	N61	0.332	1			
N72	0.277	5	N25	0.332	1	N62	0.332	1			
N73	0.277	5	N27	0.332	1	N63	0.332	1			
N74	0.277	5	N28	0.332	1	N64	0.332	1			
N75	0.277	5	N29	0.332	1	N65	0.332	1			
N76	0.277	5	N30	0.332	1	N66	0.332	1			
N77	0.277	5	N31	0.332	1	N67	0.332	1			
N78	0.277	5	N33	0.332	1	N44	0.332	1			
N79	0.277	5	N34	0.332	1	N13	0.332	1			
N70	0.278	5	N35	0.332	1	N98	0.335	1			
N15	0.286	5	N37	0.332	1	N32	0.338	1			
N100	0.306	1	N38	0.332	1	N36	0.338	1			

3.41. Shannon Entropy of B_{20_j} for $j = 1, 2, \dots 105$ and Classification

For the amino acid $A_{20}(R)$ the Shannon Entropy (SE) for the 105 binary sequences B_{20_j} for $j = 1, 2, \dots 105$ have been determined. The plot of the SE for the binary sequences has been plotted and corresponding histogram is also given in the Fig 42. The SE of the binary representations of ordering of the amino acid A_{20} over all the primary protein sequences would reveal the amount of uncertainty of the presence or absence of the amino acid.

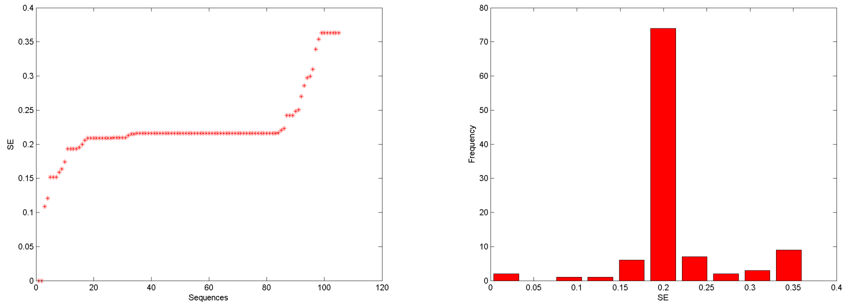


Figure 42: Plot of the SEs and corresponding histogram of all the binary sequences B_{20_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{20}(R)$.

Table 43: SE of 105 B_{20_j} for $j = 1, 2, \ldots 105$ corresponding to the amino acid $A_{20}(R)$.

Seq	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**	**Seq**	**SE**	**C**
N81	0.000	5	N86	0.210	1	N46	0.216	1	N1	0.250	1
N99	0.000	5	N71	0.213	1	N47	0.216	1	N100	0.270	1
N2	0.109	4	N31	0.215	1	N48	0.216	1	N5	0.286	3
N68	0.121	4	N39	0.215	1	N49	0.216	1	N80	0.297	3
N9	0.152	4	N26	0.216	1	N50	0.216	1	N98	0.299	3
N10	0.152	4	N14	0.216	1	N51	0.216	1	N15	0.310	3
N11	0.152	4	N16	0.216	1	N52	0.216	1	N8	0.339	3
N4	0.159	4	N17	0.216	1	N53	0.216	1	N95	0.354	3
N103	0.164	4	N18	0.216	1	N55	0.216	1	N88	0.363	3
N104	0.174	2	N19	0.216	1	N56	0.216	1	N89	0.363	3
N32	0.193	2	N20	0.216	1	N57	0.216	1	N90	0.363	3
N36	0.193	2	N21	0.216	1	N59	0.216	1	N91	0.363	3
N58	0.193	2	N22	0.216	1	N60	0.216	1	N92	0.363	3
N6	0.193	2	N23	0.216	1	N61	0.216	1	N93	0.363	3
N105	0.195	2	N24	0.216	1	N62	0.216	1	N94	0.363	3
N101	0.200	2	N25	0.216	1	N63	0.216	1			
N70	0.205	1	N27	0.216	1	N64	0.216	1			
N69	0.209	1	N28	0.216	1	N65	0.216	1			
N72	0.209	1	N29	0.216	1	N66	0.216	1			
N73	0.209	1	N30	0.216	1	N67	0.216	1			
N74	0.209	1	N33	0.216	1	N44	0.216	1			
N75	0.209	1	N34	0.216	1	N13	0.216	1			
N76	0.209	1	N35	0.216	1	N12	0.216	1			
N77	0.209	1	N37	0.216	1	N54	0.217	1			
N78	0.209	1	N38	0.216	1	N3	0.221	1			
N79	0.209	1	N40	0.216	1	N102	0.223	1			
N82	0.210	1	N41	0.216	1	N7	0.242	1			
N83	0.210	1	N42	0.216	1	N96	0.242	1			
N84	0.210	1	N43	0.216	1	N97	0.242	1			
N85	0.210	1	N45	0.216	1	N87	0.248	1			

Here the SE of the 105 binary representation of the amino acid A_{20} is ranging

from 0 to 0.404 with standard deviation 0.0638. Based on the SEs of the binary sequences all these 105 primary protein sequences of SARS-CoV2, five clusters (C) are formed as presented in the Table 43.

The amino acid $A_{20}(R)$ does not present in the protein sequence N81 and N99 and consequently the SE is zero for the binary representations $B_{20_{80}}$ and $B_{20_{99}}$. The other proteins of the remaining clusters 1, 2, 4 and 4 have the presence of the amino acid A_{20} with almost certainty.

3.42. A Collective Views of the SEs

It needless to mention that the SE is would be zero if the binary representations B_{i_j} such that the amino acid A_i is absent throughput the protein Nj. In fact that is what we have encountered throughout this study of SEs. It has been observed that the SE of the spatial distribution of the amino acids over some proteins is maximum for the smaller lengthy sequences such as N99, N80 etc. A surprising fact has been revealed that for some given amino acid A_i, the SE of some of the spatial distributions B_{i_j} for some protein sequence Nj irrespective of their lengths is identical for many values of j. This essentially reports that the probability of the presence of the amino acid A_i over those proteins is same.

Here we explore the correlation of amount uncertainty of the presence/absence of the amino acids over the proteins of SARS-CoV2 of the spatial representations. Following is 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.

Table 44: Correlation matrix of SEs of the presence of the amino acids over the proteins

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

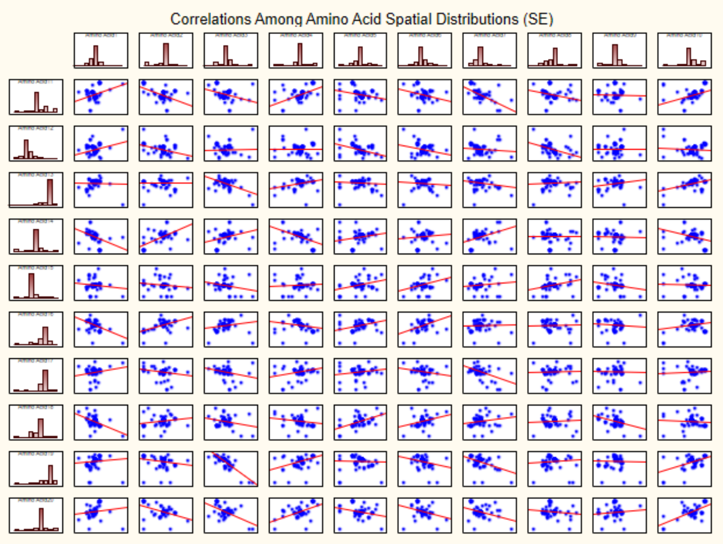


Figure 43: Plot of the correlation of the SEs of the amino acids distribution distinct pairwise.

The amount of information (based on SEs) in the spatial distribution of the amino acid A with the same of the amino acids Q, S, D, K and R are turned out to be positively correlated as shown in the Table 44. Likewise, the spatial distribution of the amino acid C is positively correlated with the same of the amino acids T, V, Y and E. Similarly the positive correlations of the spatial distributions of the amino acids F, G, H, I, L, M, N and P with the spatial distribution of the other amino acids are established in the correlation matrix, Table 23. The correlation based on SEs of the spatial distribution is also shown through graphs in the Fig 43. It is worthy mentioning that in the correlation matrix in the Table 45, the negative correlations of the spatial distribution of the proteins are also shown. As an example of the correlation based on SEs (the correlation coefficient r: 0.646) of the spatial distribution (autocorrelation) of the amino acid R with the spatial distribution of the amino acid P is given below in the Fig. 44.

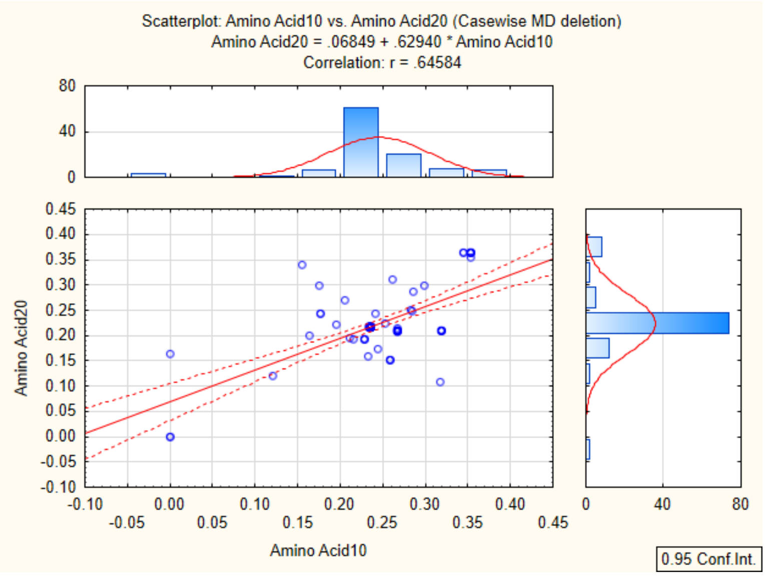


Figure 44: Plot of the correlation of the HEs of the amino acid R with that of P.

Next we are moving towards the entropy of conservation of amino acids over the 105 SARS-CoV2 proteins in the following subsection.

3.43. Amino Acid Conservation Shannon Entropy and Its Classification

For each of the 105 protein sequences the amino acid conservation information have been determined through HE as described earlier. In the following Table 45, the Shannon entropy(SE_T2) for each sequence and based on the SE, the formed clusters (C) are given. The plot of the SE over the 105 protein sequences with its histogram are given in the Fig. 45.

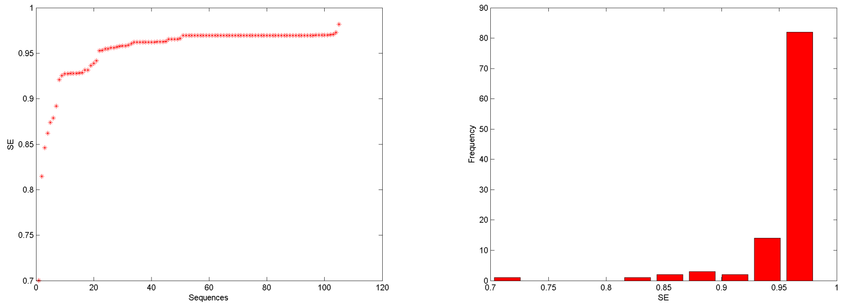


Figure 45: Plot of the SEs and corresponding histogram of all the protein sequences.

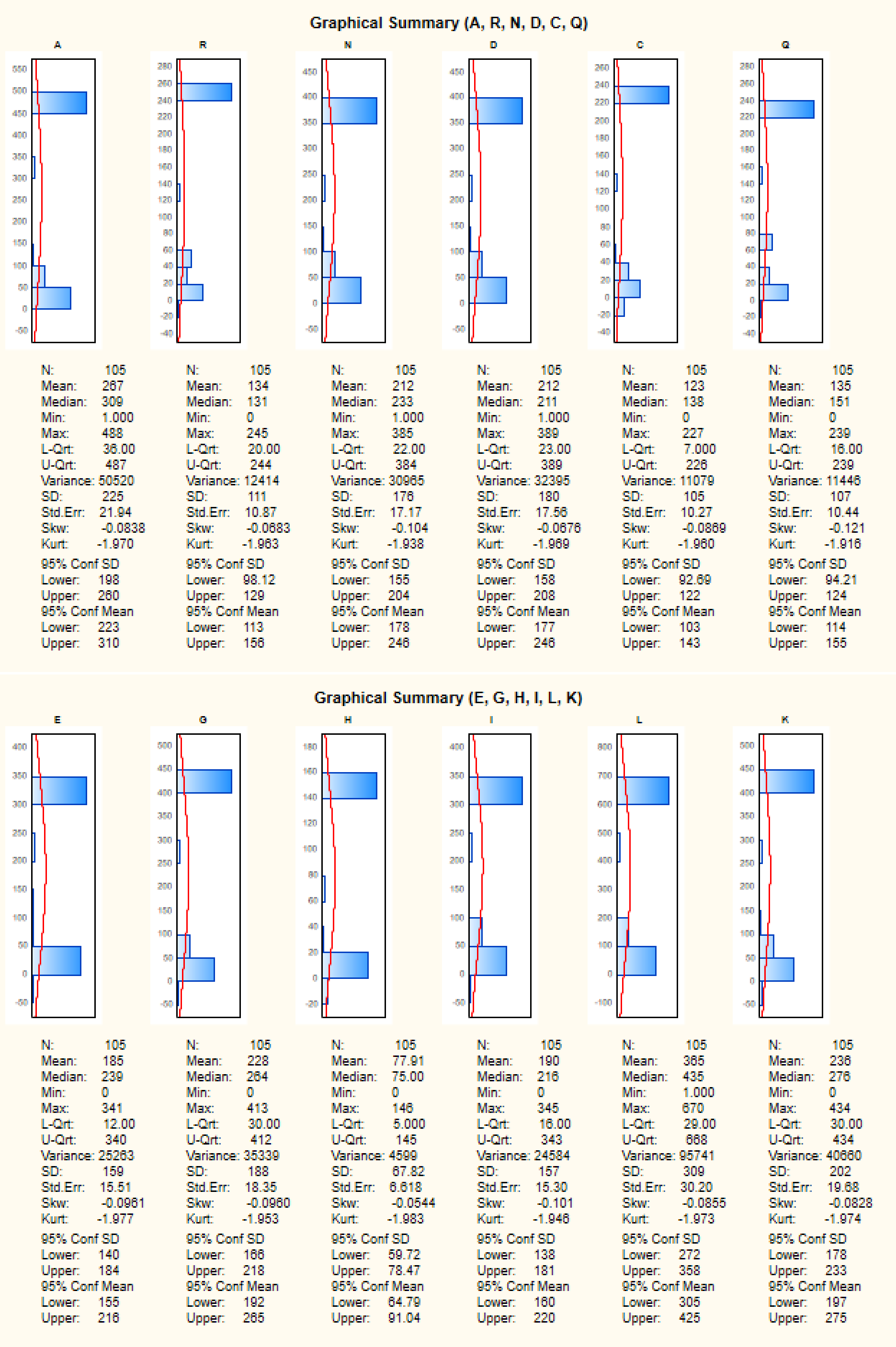
Table 45: Amino Acid Conservation Shannon Entropy

Seq	SE_T2	C	Seq	SE_T2	C	Seq	SE_T2	C	Seq	SE_T2	C
N99	0.700	4	N11	0.957	7	N13	0.970	2	N39	0.970	2
N81	0.815	6	N10	0.958	7	N23	0.970	2	N57	0.970	2
N97	0.846	6	N84	0.958	7	N37	0.970	2	N16	0.970	2
N96	0.862	5	N85	0.958	7	N49	0.970	2	N29	0.970	2
N103	0.874	5	N83	0.959	7	N64	0.970	2	N17	0.970	2
N80	0.879	5	N4	0.961	8	N66	0.970	2	N18	0.970	2
N68	0.892	5	N79	0.962	8	N60	0.970	2	N19	0.970	2
N15	0.921	9	N70	0.962	8	N12	0.970	2	N35	0.970	2
N3	0.925	9	N69	0.962	8	N65	0.970	2	N38	0.970	2
N91	0.928	9	N78	0.962	8	N56	0.970	2	N54	0.970	2
N94	0.928	9	N75	0.962	8	N41	0.970	2	N21	0.970	2
N90	0.928	9	N74	0.962	8	N55	0.970	2	N44	0.970	2
N88	0.928	9	N77	0.962	8	N30	0.970	2	N24	0.970	2
N98	0.928	9	N73	0.962	8	N53	0.970	2	N33	0.970	2
N89	0.928	9	N72	0.962	8	N59	0.970	2	N28	0.970	2
N92	0.929	9	N71	0.963	8	N40	0.970	2	N27	0.970	2
N95	0.931	1	N5	0.963	8	N43	0.970	2	N52	0.970	2
N93	0.931	1	N76	0.963	8	N48	0.970	2	N47	0.970	2
N87	0.936	1	N58	0.965	8	N50	0.970	2	N62	0.970	2
N8	0.939	3	N36	0.965	8	N51	0.970	2	N34	0.970	2
N101	0.942	3	N32	0.965	8	N25	0.970	2	N22	0.970	2
N2	0.953	7	N105	0.965	8	N26	0.970	2	N67	0.970	2
N104	0.953	7	N102	0.966	8	N45	0.970	2	N20	0.971	2
N9	0.955	7	N100	0.970	2	N46	0.970	2	N86	0.973	2
N7	0.955	7	N42	0.970	2	N14	0.970	2	N1	0.982	10
N82	0.956	7	N61	0.970	2	N31	0.970	2			
N6	0.956	7	N63	0.970	2						

It is observed that the Shannon entropy of amino acid conservations along protein sequences of SARS-CoV2 are ranging from 0.7 to 0.982 which is closed to 1. Clearly, all the amino acids are uniformly distributed over the protein sequences since the uncertainty is maximum as the SE is nearly 1. More than 50% of the proteins belong to the cluster 2 of SARS-CoV2 (54) having SE 0.970 which implies the amino acids are almost uniformly spread over the sequences. At last, the frequency analysis of the amino acids over the proteins is given in the following subsection.

3.44. Frequency Distribution of Amino Acids over the SARS-CoV2 Proteins

In this section, we shall look into the frequencies of each of the amino acids in the 105 SARS-CoV2 proteins. At first a complete statistical comparison among all the amino acid frequencies over the 105 proteins is reported in the following Fig. 46 and Fig. 47.



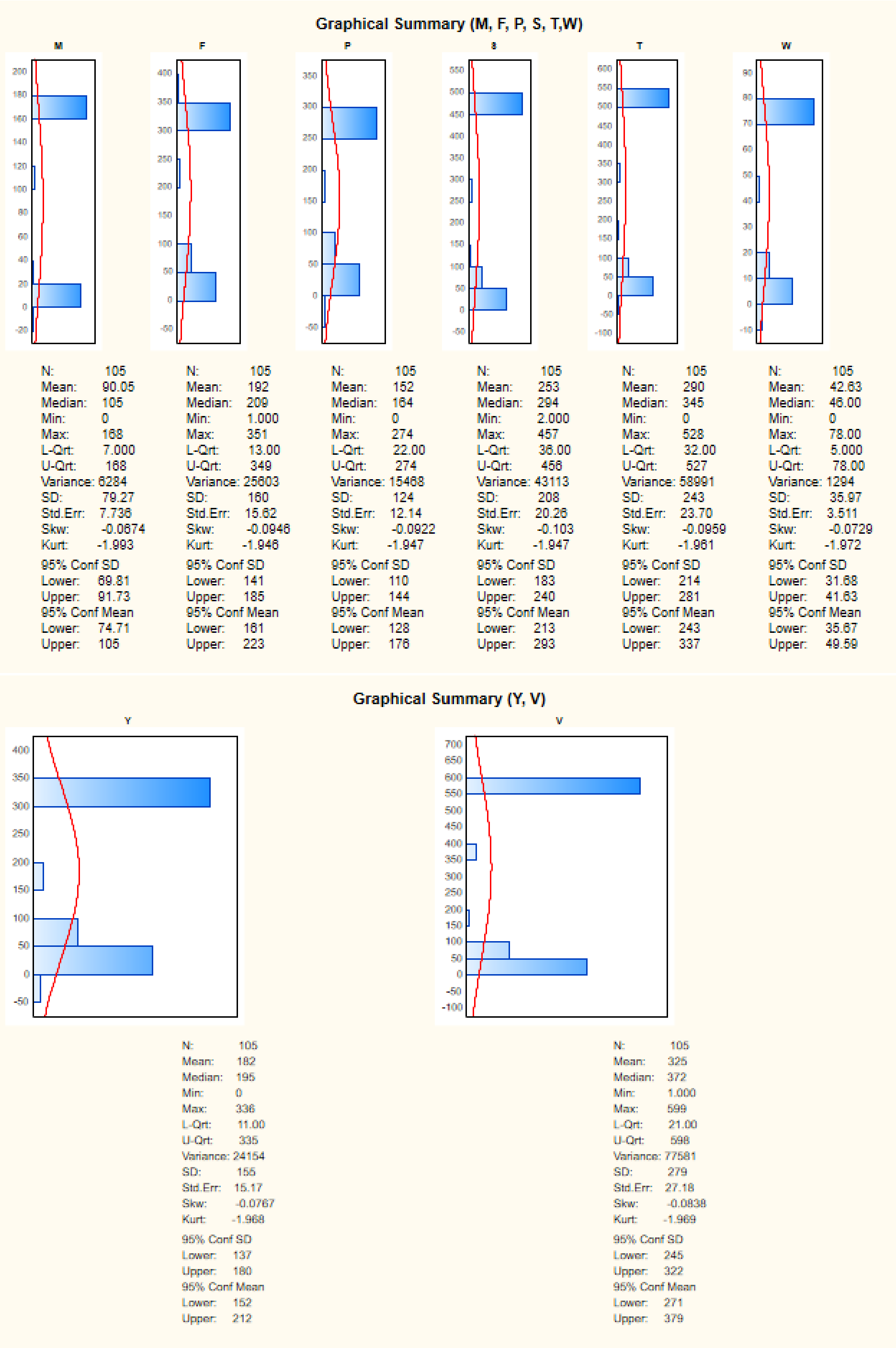
Graphical Summary (A, R, N, D, C, Q)

	A	R	N	D	C	Q
N:	105	105	105	105	105	105
Mean:	267	134	212	212	123	135
Median:	309	131	233	211	138	151
Min:	1.000	0	1.000	1.000	0	0
Max:	488	245	385	389	227	239
L-Qrt:	36.00	20.00	22.00	23.00	7.000	16.00
U-Qrt:	487	244	384	389	226	239
Variance:	50520	12414	30965	32395	11079	11446
SD:	225	111	176	180	105	107
Std.Err:	21.94	10.87	17.17	17.56	10.27	10.44
Skw:	-0.0838	-0.0683	-0.104	-0.0676	-0.0869	-0.121
Kurt:	-1.970	-1.963	-1.938	-1.969	-1.960	-1.916
95% Conf SD						
Lower:	198	98.12	155	158	92.69	94.21
Upper:	260	129	204	208	122	124
95% Conf Mean						
Lower:	223	113	178	177	103	114
Upper:	310	156	246	246	143	155

Graphical Summary (E, G, H, I, L, K)

	E	G	H	I	L	K
N:	105	105	105	105	105	105
Mean:	185	228	77.91	190	365	236
Median:	239	264	75.00	216	435	276
Min:	0	0	0	0	1.000	0
Max:	341	413	146	345	670	434
L-Qrt:	12.00	30.00	5.000	16.00	29.00	30.00
U-Qrt:	340	412	145	343	668	434
Variance:	25263	35339	4599	24584	95741	40660
SD:	159	188	67.82	157	309	202
Std.Err:	15.51	18.35	6.618	15.30	30.20	19.68
Skw:	-0.0961	-0.0960	-0.0544	-0.101	-0.0855	-0.0828
Kurt:	-1.977	-1.953	-1.983	-1.946	-1.973	-1.974
95% Conf SD						
Lower:	140	166	59.72	138	272	178
Upper:	184	218	78.47	181	358	233
95% Conf Mean						
Lower:	155	192	64.79	160	305	197
Upper:	216	265	91.04	220	425	275

Figure 46: Comparative statistical details about frequencies of the amino acids over the proteins.



```
N:         105
Mean:      90.05
Median:    105
Min:       0
Max:       168
L-Qrt:     7.000
U-Qrt:     168
Variance: 6284
SD:        79.27
Std.Err:   7.736
Skw:       -0.0674
Kurt:      -1.993
95% Conf SD
Lower:     69.81
Upper:     91.73
95% Conf Mean
Lower:     74.71
Upper:     105
```

```
N:         105
Mean:      192
Median:    209
Min:       1.000
Max:       351
L-Qrt:     13.00
U-Qrt:     349
Variance: 25603
SD:        160
Std.Err:   15.62
Skw:       -0.0946
Kurt:      -1.946
95% Conf SD
Lower:     141
Upper:     185
95% Conf Mean
Lower:     161
Upper:     223
```

```
N:         105
Mean:      152
Median:    164
Min:       0
Max:       274
L-Qrt:     22.00
U-Qrt:     274
Variance: 15468
SD:        124
Std.Err:   12.14
Skw:       -0.0922
Kurt:      -1.947
95% Conf SD
Lower:     110
Upper:     144
95% Conf Mean
Lower:     128
Upper:     176
```

```
N:         105
Mean:      253
Median:    294
Min:       2.000
Max:       457
L-Qrt:     36.00
U-Qrt:     456
Variance: 43113
SD:        208
Std.Err:   20.26
Skw:       -0.103
Kurt:      -1.947
95% Conf SD
Lower:     183
Upper:     240
95% Conf Mean
Lower:     213
Upper:     293
```

```
N:         105
Mean:      290
Median:    345
Min:       0
Max:       528
L-Qrt:     32.00
U-Qrt:     527
Variance: 58991
SD:        243
Std.Err:   23.70
Skw:       -0.0959
Kurt:      -1.961
95% Conf SD
Lower:     214
Upper:     281
95% Conf Mean
Lower:     243
Upper:     337
```

```
N:         105
Mean:      42.63
Median:    46.00
Min:       0
Max:       78.00
L-Qrt:     5.000
U-Qrt:     78.00
Variance: 1294
SD:        35.97
Std.Err:   3.511
Skw:       -0.0729
Kurt:      -1.972
95% Conf SD
Lower:     31.68
Upper:     41.63
95% Conf Mean
Lower:     35.67
Upper:     49.59
```

```
N:         105
Mean:      182
Median:    195
Min:       0
Max:       336
L-Qrt:     11.00
U-Qrt:     335
Variance: 24154
SD:        155
Std.Err:   15.17
Skw:       -0.0767
Kurt:      -1.968
95% Conf SD
Lower:     137
Upper:     180
95% Conf Mean
Lower:     152
Upper:     212
```

```
N:         105
Mean:      325
Median:    372
Min:       1.000
Max:       599
L-Qrt:     21.00
U-Qrt:     598
Variance: 77581
SD:        279
Std.Err:   27.18
Skw:       -0.0838
Kurt:      -1.969
95% Conf SD
Lower:     245
Upper:     322
95% Conf Mean
Lower:     271
Upper:     379
```

Figure 47: Comparative statistical details about frequencies of the amino acids over the proteins.

Following a correlation among the frequency distribution of each of the amino acids over the 105 proteins of SARS-CoV2. The correlation coefficient corresponding to frequency distribution over the proteins is given in the Table 46.

Table 46: Correlation matrix of the frequencies of the two amino acids pairwise

	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

The corresponding correlation are also given pairwise in a matrix form in the Fig. 48.

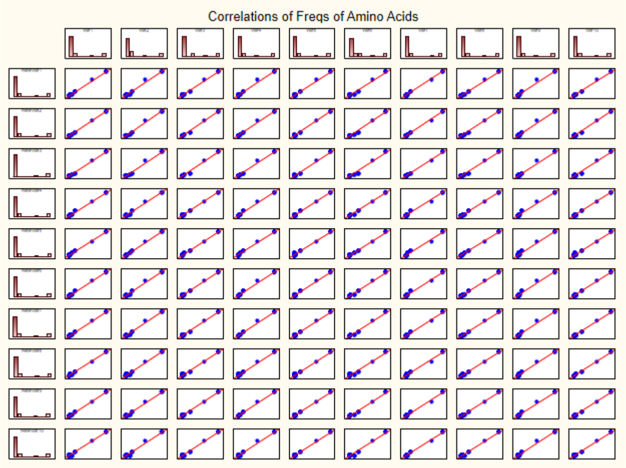


Figure 48: Correlation graphs for each of the amino acid frequencies.

It is observed that the correlation coefficient is very close to 1 which ensures the existence of significant correlations of frequencies of each of the amino acids over the proteins. In fact the correlation coefficient between the frequency distributions corresponding to the amino acids A (Aliphatic) and K (Basic) is 1.

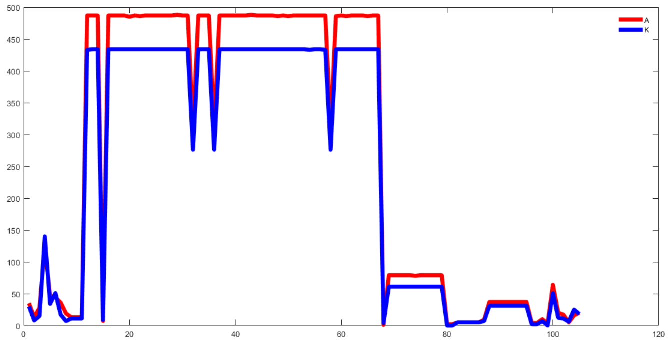


Figure 49: Frequency plots of the amino acids A and K over 105 proteins.

The frequency plots of the amino acids A and K is given in the Fig. 49. The plots show the the strong correlation between the frequency distribution over the proteins. Overall it is observed that the same length proteins have mostly similar frequency distribution of the twenty amino acids.

Next we are heading towards a comparative ...

4. Spatial Organization of Proteins of SARS-CoV

In 2003, the SARS coronavirus (SARS-CoV) had caused an epidemic in China including other 22 countries [42, 43]. There 14 protein sequences available in the NCBI database (taxid: 722424). The list of these protein (S1, S2, . . . S11) with their accessions is given here in the following Table 47.

Table 47: List of SARS-CoV proteins with their Accession and length

Accession	**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

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 its important function in the case of SARS-CoV [44, 45]. Among all other proteins of SARS-CoV, spike protein is the main antigenic component which is responsible for inducing host immune responses, neutralizing antibodies and/or protective immunity against virus infection [46]. We therefore illuminate here the spatial representations of the amino acids over the the spike protein including other 13 proteins as mentioned in the Table 47. 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

Table 48: HEs and SEs of 14 proteins of the SARS-CoV

HE																				
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

SE																				
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

trending) as well as with least amount of uncertainty of presence of the amino acids. It seems that the presence of all the amino acids are necessary to make a spike protein. It worth mentioning that yet there is 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 the 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 kep blank for the cases where the spatial distribution of an amino acid is completely a sequences of zeros i,e. absence of the amino acid over the protein.

Below in the Table 49, we derive the correlation coefficients of the HEs of the spatial representations of the amino acids over the 14 SARS-CoV proteins.

Table 49: 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

It is observed from the Table 49 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 over all the proteins the presence of the amino acid M and T are ensured. There are also other positive correlation exist as it can be seen in the Table 49.

It is noted that the SE is turned out to be zero for the cases where the spatial distribution corresponding to an amino acid which is absent over a protein. The spatial distribution of amino acids over the proteins of SARS-CoV are all without much of uncertainty except three cases where the SEs are greater than the 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 the Table 50.

Table 50: 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

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 the Table 50. 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. Conclusions and Summary

In this present study the spatial arrangement of amino acids over the SARS-CoV2 proteins have been looked into through its autocorrelation by Hurst exponent and the amount of information of presence of the amino acids over the proteins through Shannon entropy are analysed. Also the frequency distribution of amino acids are also taken into consideration for categorizing the protein sequences. At the end, the features of the spatial distribution of 14 proteins of the SARS-CoV are also have been determined and seen significant difference with the former proteins of the SARS-CoV2. The study is a collection of information about the spatial distribution of the amino acids over the proteins of SARS-CoV2 as well as SARS-CoV. The difference noted in the section 4 would enable reader to distinguish the proteins of two different types of the CoVs. We firmly believe that this spatial understanding of the presence and absence of each amino acids over the proteins would enable understanding the PPIs as well as identifying spike proteins in the case of SARS-CoV2.

References

[1] C. P. E. R. E. Novel, et al., The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19) in china, Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi 41 (2) (2020) 145.

[2] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al., A novel coronavirus from patients with pneumonia in china, 2019, New England Journal of Medicine (2020).

[3] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., Clinical features of patients infected with 2019 novel coronavirus in wuhan, china, The Lancet 395 (10223) (2020) 497–506.

[4] C. Wang, P. W. Horby, F. G. Hayden, G. F. Gao, A novel coronavirus outbreak of global health concern, The Lancet 395 (10223) (2020) 470–473.

[5] S. Perlman, Another decade, another coronavirus (2020).

[6] Z.-W. Ye, S. Yuan, K.-S. Yuen, S.-Y. Fung, C.-P. Chan, D.-Y. Jin, Zoonotic origins of human coronaviruses, International Journal of Biological Sciences 16 (10) (2020) 1686.

[7] C. Ceraolo, F. M. Giorgi, Genomic variance of the 2019-ncov coronavirus, Journal of Medical Virology (2020).

[8] C. S. G. of the International, et al., The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2, Nature Microbiology (2020) 1.

[9] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, The proximal origin of sars-cov-2, Nature Medicine (2020) 1–3.

[10] Y.-Z. Zhang, E. C. Holmes, A genomic perspective on the origin and emergence of sars-cov-2, Cell (2020).

[11] X. Tang, C. Wu, X. Li, Y. Song, X. Yao, X. Wu, Y. Duan, H. Zhang, Y. Wang, Z. Qian, et al., On the origin and continuing evolution of sars-cov-2, National Science Review (2020).

[12] E. L. Hatcher, S. A. Zhdanov, Y. Bao, O. Blinkova, E. P. Nawrocki, Y. Ostapchuck, A. A. Schäffer, J. R. Brister, Virus variation resource–improved response to emergent viral outbreaks, Nucleic acids research 45 (D1) (2017) D482–D490.

[13] E. W. Sayers, J. Beck, J. R. Brister, E. E. Bolton, K. Canese, D. C. Comeau, K. Funk, A. Ketter, S. Kim, A. Kimchi, et al., Database resources of the national center for biotechnology information, Nucleic acids research 48 (D1) (2020) D9.

[14] 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, Human Vaccines & Immunotherapeutics (2020) 1–7.

[15] 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, et al., Research and development on therapeutic agents and vaccines for covid-19 and related human coronavirus diseases (2020).

[16] 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, Journal of Medical Virology (2020).

[17] N. Goodacre, P. Devkota, E. Bae, S. Wuchty, P. Uetz, Protein-protein interactions of human viruses, in: Seminars in cell & developmental biology, Vol. 99, Elsevier, 2020, pp. 31–39.

[18] 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, Computational and structural biotechnology journal 18 (2020) 153–161.

[19] 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 (4) (2020) 360.

[20] 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. Huttenhain, et al., A sars-cov-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing, BioRxiv (2020).

[21] R. Kolodny, D. Petrey, B. Honig, Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction, Current opinion in structural biology 16 (3) (2006) 393–398.

[22] E. Krissinel, K. Henrick, Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions, Acta Crystallographica Section D: Biological Crystallography 60 (12) (2004) 2256–2268.

[23] X. Pennec, N. Ayache, A geometric algorithm to find small but highly similar 3d substructures in proteins., Bioinformatics (Oxford, England) 14 (6) (1998) 516–522.

[24] 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: Structure, Function, and Bioinformatics 68 (4) (2007) 915–921.

[25] M. I. GROMIHA, P. Ponnuswamy, Hydrophobic distribution and spatial arrangement of amino acid residues in membrane proteins, International journal of peptide and protein research 48 (5) (1996) 452–460.

doi:10.20944/preprints202004.0034.v1

[26] T. Kollár, I. Pálinkó, Z. Kónya, I. Kiricsi, Intercalating amino acid guests into montmorillonite host, Journal of Molecular Structure 651 (2003) 335–340.

[27] S. S. Hassan, R. K. Rout, V. Sharma, A quantitative genomic view of the coronaviruses: Sars-cov2 (2020).

[28] J. R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, Ncbi viral genomes resource, Nucleic acids research 43 (D1) (2015) D571–D577.

[29] K. L. Schierhorn, F. Jolmes, J. Bespalowa, S. Saenger, C. Peteranderl, J. Dzieciolowski, M. Mielke, M. Budt, S. Pleschka, A. Herrmann, et al., 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, Journal of virology 91 (10) (2017) e00198–17.

[30] U. A. Ashfaq, T. Javed, S. Rehman, Z. Nawaz, S. Riazuddin, An overview of hcv molecular biology, replication and immune responses, Virology journal 8 (1) (2011) 161.

[31] W. Luytjes, L. S. Sturman, P. J. Bredenbee, J. Charite, B. A. van der Zeijst, M. C. Horzinek, W. J. Spaan, Primary structure of the glycoprotein e2 of coronavirus mhv-a59 and identification of the trypsin cleavage site, Virology 161 (2) (1987) 479–487.

[32] J. P. Banerjee, J. K. Das, P. P. Choudhury, S. Mukherjee, S. S. Hassan, P. Basu, The variations of human mirnas and ising like base pairing models, BioRxiv (2018) 319301.

[33] 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 (4) (2019) 549–559.

[34] 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, Scientific reports 8 (1) (2018) 1–19.

[35] M. Sánchez-Granero, M. Fernández-Martínez, J. Trinidad-Segovia, Introducing fractal dimension algorithms to calculate the hurst exponent of financial time series, The European Physical Journal B 85 (3) (2012) 86.

[36] M. D. Kale, F. B. Butar, Fractal analysis of time series and distribution properties of hurst exponent, Ph.D. thesis, Sam Houston State University (2005).

[37] J. Mielniczuk, P. Wojdyłło, Estimation of hurst exponent revisited, Computational Statistics & Data Analysis 51 (9) (2007) 4510–4525.

[38] J. Lin, Divergence measures based on the shannon entropy, IEEE Transactions on Information theory 37 (1) (1991) 145–151.

[39] B. J. Strait, T. G. Dewey, The shannon information entropy of protein sequences, Biophysical journal 71 (1) (1996) 148–155.

[40] L. R. Nemzer, Shannon information entropy in the canonical genetic code, Journal of theoretical biology 415 (2017) 158–170.

[41] G. Dovbeshko, L. Berezhinsky, Low frequency vibrational spectra of some amino acids, Journal of molecular structure 450 (1-3) (1998) 121–128.

[42] X. Xiao, S. Chakraborti, A. S. Dimitrov, K. Gramatikoff, D. S. Dimitrov, The sars-cov s glycoprotein: expression and functional characterization, Biochemical and biophysical research communications 312 (4) (2003) 1159–1164.

[43] 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, Proceedings of the National Academy of Sciences 101 (12) (2004) 4240–4245.

[44] 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, Nature Reviews Microbiology 7 (3) (2009) 226–236.

doi:10.20944/preprints202004.0034.v1

[45] 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, Biochemical and biophysical research communications 324 (2) (2004) 773–781.

[46] J. Cinatl, B. Morgenstern, G. Bauer, P. Chandra, H. Rabenau, H. Doerr, Treatment of sars with human interferons, The Lancet 362 (9380) (2003) 293–294.