 doi:10.20944/preprints202003.0344.v1

A Quantitative Genomic View of the Coronaviruses: SARS-COV2

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

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

Abstract

In 2020, the pandemic caused by the Coronaviruses (CoV) that are a large family of viruses that cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV2). The Coronavirus disease (COVID-19) is a new strain that was discovered in 2019 and has not been previously identified in humans. It is the high time to investigate the quantitative and/or qualitative genomic informations of the virus SARS-CoV2 in order to strengthen the healthcare facility to fight against this viral disease. In this article, a through quantitative understanding of the purine and pyrimidine spatial distribution/organization of all 89 complete sequences of SARS-CoV (available as on date in the *NCBI virus database*, is made using different parameters such as fractal dimension, Hurst exponent, Shannon entropy and *GC* content of the nucleotide sequences of the genome of SARS-CoV2. Also a cluster among all the the SARS-CoV sequences of nucleotide have been made based on their phylogeny made through their closeness (Hamming distance) based on respective purine-pyrimidine distribution.

Keywords: Fractal Dimension, Shannon Entropy, Hurst Exponent, GC Content & SARS-CoV2.

*Corresponding author
Email addresses: sarimif@gmail.com (Sk. Sarif Hassan), ranjeetkumarrout@nitsri.net (Ranjeet Kumar Rout), vipul_1phd17@nitsri.net (Vipul Sharma)

 doi:10.20944/preprints202003.0344.v1

1. Introduction

The Coronavirus disease (COVID-19) is caused by SARS-COV2 and represents the causative agent of a potentially fatal disease that is of great global public health concern [1], [2]. Based on the large number of infected people that were exposed to the wet animal market in Wuhan City, China, it is suggested that this is likely the zoonotic origin of COVID-19 [3, 4, 5]. Person-to-person transmission of COVID-19 infection led to the isolation of patients that were subsequently administered a variety of treatments [6, 7]. As of 11 February 2020, data from the World Health Organization (WHO) have shown that more than 43000 confirmed cases have been identified in 28 countries/regions, with $\geq$ 99% of cases being detected in China [8]. On 30 January 2020, the WHO declared COVID-19 as the sixth public health emergency of international concern [9]. SARS-CoV2 is closely related to two bat-derived severe acute respiratory syndrome-like coronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21 [10]. On 11 February 2020, the WHO formally named the disease triggered by $2019-nCoV$ as coronavirus disease 2019 (COVID-19). Also on that very day, the coronavirus study group of the International Committee on Taxonomy of Viruses named $2019-nCoV$ as severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) [11]. Complete genomic sequences have been released by the NCBI in the last few weeks to understand the evolutionary origin and molecular characteristics of this virus [12]. Ceraolo and Giorgi [13] have confirmed the high sequence similarity ($>$ 99%) between all sequenced 2019CoVs genomes available, with the closest BCoV sequence sharing 96.2% sequence identity, confirming the notion of a zoonotic origin of $2019-nCoV$. Coronaviruses are enveloped RNA viruses that are distributed broadly among humans, other mammals, and birds and that cause respiratory, enteric, hepatic, and neurologic diseases [10, 14, 15].

As on date 15th March, 2020, there are 89 nucleotide sequences of SARS-CoV2 available in the NCBI virus database [16, 17]. All these sequences are

 doi:10.20944/preprints202003.0344.v1

nearly about length 29 thousand and each of them are composed of four nucleotide bases viz. A, T, C and G. Importantly, they all are different from each other by means of spatial organizations of the nucleotide bases.

In this study, our aim is to attempt to discover the signatory imprint of this spatial organizations of the SARS-CoV2. The spatial distribution of the purine and pyrimidine bases over the nucleotide sequences of the SARS-CoV2 are being fetched out through some quantitative parameters such as fractal dimension, Hurst exponent and Shannon entropy. In addition, also density of each of the bases are also seen and density of GC content is also determined in order to understand the stability of the DNAs.

This discovery would aid in the diagnosis of SARS-CoV2 virus infection in humans and potential animal hosts (using polymerase chain reaction and immunological tests), in the development of antivirals (including neutralizing antibodies), and in the identification of putative epitopes for vaccine development.

1.1. Database used and Specifications

In this work we have taken all nucleotide sequences from the NCBI Virus Database ($https://www.ncbi.nlm.nih.gov/labs/virus/vssi/$) for experimental results and discussion purpose. This dataset contains 89 complete SARS-CoV2 nucleotide sequences as on date 15th March, 2020. We have transformed each DNA sequence to a binary sequence of $0's$ and $1's$ which is defined in equation 1. Here purines and pyrimidines nucleotide bases are represented as "1" and "0" respectively.

$$\begin{aligned} A/G &\rightarrow 0 \\ T/C &\rightarrow 1 \end{aligned} \tag{1}$$

Equation(1) represents purine and pyrimidine nucleotide bases which are encoded as 1 and 0 respectively into the transformed binary sequence.

 doi:10.20944/preprints202003.0344.v1

Before we proceed further, we have named all the 89 complete SARS-CoV2 nucleotide sequences based on their accession ID as listed below in the Table 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

Table 1: Naming the Nucleotide sequences of SARS-CoV2.

Sequence	Accession ID	Sequence	Accession ID	Sequence	Accession ID
S1	NC_045512	S31	MT121215	S61	MT039873
S2	MT188341	S32	MT159719	S62	MT039887
S3	MT188339	S33	MT159720	S63	MT039890
S4	MT188340	S34	MT159709	S64	MT027063
S5	MT184910	S35	MT159718	S65	MT027064
S6	MT184908	S36	MT012098	S66	MT027062
S7	MT184909	S37	MT050493	S67	MT019529
S8	MT184911	S38	MT152824	S68	MT020880
S9	MT184913	S39	MT135044	S69	MT019530
S10	MT184912	S40	MT135042	S70	MT019532
S11	MT184907	S41	MT135043	S71	MT019533
S12	MT163716	S42	MT135041	S72	MT020881
S13	MT163719	S43	MT126808	S73	MT019531
S14	MT163717	S44	MT123291	S74	MT007544
S15	MT163718	S45	MT123290	S75	MN996527
S16	MT159711	S46	MT123293	S76	MN996531
S17	MT159710	S47	MT123292	S77	MN996528
S18	MT159708	S48	MT118835	S78	MN996530
S19	MT159712	S49	MT106054	S79	MN996529
S20	MT159716	S50	MT106053	S80	MN988668
S21	MT159707	S51	MT106052	S81	MN997409
S22	MT159715	S52	MT093571	S82	MN994467
S23	MT159721	S53	MT093631	S83	MN988669
S24	MT159717	S54	MT072688	S84	MN994468
S25	MT159722	S55	MT066175	S85	MN988713
S26	MT159714	S56	MT066176	S86	MN975262
S27	MT159713	S57	MT044257	S87	MN938384
S28	MT159706	S58	MT049951	S88	MN985325
S29	MT066156	S59	MT044258	S89	MN908947
S30	MT159705	S60	MT039888		

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

Viruses of the family Coronaviridae possess a single-strand, positive-sense RNA genome ranging from 26 to 32 kilobases in length []. The length of these complete 89 sequences is varying from 29783 to 29981. So the range is 198 bp long. The smallest complete SARS-CoV sequence is S2 of length 29783 and the largest one is S47 having length 29981. There are two sequences S23 and S28 having length 29867 and few others having same lengths. There are 39 sequences having exactly same length which is 29882. Also there are 11 sequences having length 29903.

2. Proposed Methods

In this section, four different quantitative parameters have been defined to characterize the spatial distribution of the SARS-CoV2 sequences. Based on quantitative parameters (Shannon Entropy, Fractal Dimension, Hurst Exponent, distribution of purines- pyrimidines) ten different clusters have be generated. Following we present the methods in brief.

2.1. Fractal Dimension of Indicator Matrices

Let $D = \{0, 1\}$ be the set of two symbols characterizing the purine and pyrimidine bases of a nucleotide sequence and $S(l)$ be a binary sequence corresponding to a nucleotide sequence with the repetition of two characters from D of length l. Here, we convert each of the binary sequences into indicator matrices [18, 19, 20, 21]. In literature [22] there are several methods to find out the self organising structure of DNA sequences through indicator matrix. Then the indicator function for each sequence is defined as shown in equation 2:

$$\vartheta : \{0, 1\} \times \{0, 1\} \rightarrow \{0, 1\}, \tag{2}$$

such that the indicator matrix:

$$\vartheta_{hk} = \vartheta(x, y) = \begin{cases} 1, if x = y \\ 0, if x \neq y \end{cases} \quad \text{where } x, y \in \{0, 1\}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

Here ϑ_{hk} is a matrix with the distribution 0 and 1. A binary image can be obtained from the matrix through which we can visualise correlation between purines and pyrimidines and auto-correlation for the same sequence. It can be well understood by assigning a black dot to 1 and a white dot to 0. From the indicator matrix we can visualise the fractal like distribution of $0's$ and $1's$ (*purines and pyrimidines*). The fractal dimension of the indicator matrix can calculated as the average number of $\sigma(p)$ of 1, which can be taken from $P \times P$ indicator matrix with $p \times p$ randomly. Using $\sigma(p)$, the fractal dimension is defined in equation 3

$$D = -\frac{1}{P}\sum_{n=2}^{P}\frac{log\sigma(p)}{logp} \tag{3}$$

The self-organization of the purine and pyrimidine bases for all the SARS-CoV2 sequences can be obtained through the fractal dimension of the indicator matrix.

2.2. Hurst Exponent

The Hurst Exponent (HE) is used for time series analysis to interpret the autocorrelation [23, 24]. The value of HE is in between 0 to 1. The HE value $0 < HE < 0.5$ and $0.5 < HE < 1$ designates negative and positive auto-correlation of a time series respectively and 0.5 denotes a absolute randomness of a time series which indicates the equally likely value from a particular value either by increasing or by decreasing. The HE of a binary sequence s_n is defined as

$$\left(\frac{n}{2}\right)^{HE} = \frac{X(n)}{Y(n)} \tag{4}$$

where

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

and

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

The auto correlation of purine-pyrimidine bases for all the SARS-CoV2 sequences is obtained through the Hurst exponent.

2.3. 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 [25, 26].

The binary Shannon entropy is a measure of the uncertainty in a binary string. 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.

2.4. GC Content and Nucleotides Density

In molecular biology, The GC content is usually calculated as a percentage and sometimes called $G + C$ ratio or GC-ratio [27, 28]. GC-content percentage is calculated by the formula $\frac{Count(G+C)}{Count(A+T+G+C)} \times 100\%$ [29, 30]. A DNA with low GC-content is likely to be unstable than DNA with high GC-content; however, the hydrogen bonds themselves do not have a particularly significant impact on molecular stability, which is instead caused mainly by molecular interactions of base stacking. The GC-content percentages as well as GC-ratio can be measured by several means, but one of the simplest methods is to measure the melting temperature of the DNA double helix using spectrophotometry.

In addition to the GC content, the density of the nucleotides A, T, C and G also separately are obtained in the present study [31, 32].

 doi:10.20944/preprints202003.0344.v1

3. Results and Illustrations

It is well understood from their very frequency of number of nucleotides usages that the SARS-CoV2 sequences are not randomly chosen. So we explicitly trying to get the spatial distribution of the purine and pyrimidine organizations among the SARS-CoV2 sequences through the parameters as defined in the previous section. In addition to the investigation of the purine-pyrimidine distribution, we wish to explore the density of each of the nucleotides as well as *GC* content which has a significant role in stability.

3.1. Classification Based on Fractal dimension of Indicator Matrices

For each binary sequence (purine and pyrimidine) of SARS-CoV2, the fractal dimension (using Equation (3)) is calculated. Based on the fractal dimension, we have made classifications (clusters) for all the the SARS-CoV2 sequences. There are three distinct fractal dimensions (0.3, 0.4755 and 0.6) have been obtained and consequently only three clusters of the sequences are turned up. The following Table 2 demonstrate the sequences and their corresponding FDs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

Table 2: Sequences and their corresponding FDs

Seq	FD	Seq	FD	Seq	FD	Seq	FD	Seq	FD
S47	0.300	S7	0.6	S26	0.6	S45	0.6	S72	0.6
S13	0.300	S8	0.6	S27	0.6	S46	0.6	S73	0.6
S28	0.300	S9	0.6	S29	0.6	S52	0.6	S74	0.6
S79	0.300	S10	0.6	S30	0.6	S57	0.6	S75	0.6
S48	0.475	S11	0.6	S31	0.6	S58	0.6	S76	0.6
S49	0.475	S12	0.6	S32	0.6	S59	0.6	S77	0.6
S50	0.475	S14	0.6	S33	0.6	S60	0.6	S78	0.6
S51	0.475	S15	0.6	S34	0.6	S61	0.6	S80	0.6
S53	0.475	S16	0.6	S35	0.6	S62	0.6	S81	0.6
S54	0.475	S17	0.6	S36	0.6	S63	0.6	S82	0.6
S55	0.475	S18	0.6	S37	0.6	S64	0.6	S83	0.6
S56	0.475	S19	0.6	S38	0.6	S65	0.6	S84	0.6
S1	0.600	S20	0.6	S39	0.6	S66	0.6	S85	0.6
S2	0.600	S21	0.6	S40	0.6	S67	0.6	S86	0.6
S3	0.600	S22	0.6	S41	0.6	S68	0.6	S87	0.6
S4	0.600	S23	0.6	S42	0.6	S69	0.6	S88	0.6
S5	0.600	S24	0.6	S43	0.6	S70	0.6	S89	0.6
S6	0.600	S25	0.6	S44	0.6	S71	0.6		

The plot of the FD and corresponding histogram are figured in the Fig. 1.

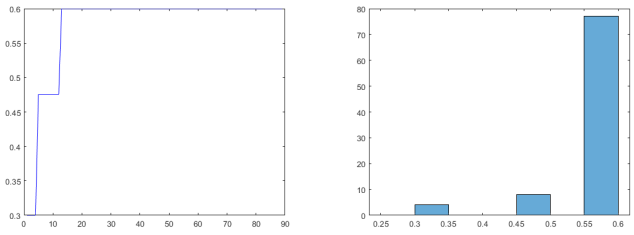


Figure 1: Plot of the Fractal dimension (FD) and corresponding histogram of all the purine-pyrimidine binary sequences corresponding to SARS-CoV2 sequences.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

The dimension of each of indicator matrix is above 29000×29000 and consequently we fail to demonstrate image of the indicator matrix here. The sequences S47, S13, S28 and S79 have the FD 0.3 which depicts that the amount of fractality (a kind of non-linearity) is small and so the purine and pyrimidine organization is rather well-organized and closely affine-type. There are eight sequences S48, S49, S50, S51, S53, S54, S55 and S56 having FD 0.4755 and FD of rest all the sequences of purine and pyrimidine of SARS-CoV2 have been found as 0.6 which is close to the FD of cantor set, which is coincidentally significant [33, 34].

3.2. Classification Based on Hurst exponent

For each of the binary sequences of SARS-CoV2, the Hurst exponent (HE) (using Equation (4)) is determined and then ten clusters are formed using k-means clustering technique for all the sequences. The Hurst exponents and the histograms of all the SARS-CoV2 sequences are plotted in the Fig. 2.

It has been observed that the HE is confined in the interval $(0.643, 0.655)$ of length 0.0123. This suggests that spatial distribution of the purine and pyrimidine bases of all the SARS-CoV2 sequence is positively autocorrelated. It is noted that there is a sequence S1 having HE 0.712 which can be seen the following Table 3. This sequence S1 (accession ID: $NC_0 4551$) has highest HE and clearly this sequence is having a significantly different spatial organization of purine and pyrimidine bases. The length of the sequence S1 is 29903. It is worth mentioning that there are other ten sequences (S1, S13, S14, S15, S39, S40, S41, S42, S57, S60 and S89) having same length 29903 but their HE is significantly differed from the HE of the sequence S1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

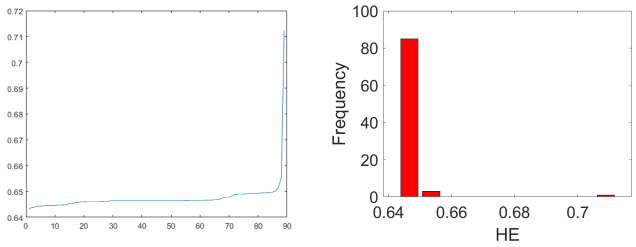


Figure 2: Plot of the Hurst exponent (HE) and corresponding histogram of all the purine-pyrimidine binary sequences corresponding to SARS-CoV2 sequences.

Based on the HE obtained from the binary sequences of SARS-CoV2, ten clusters have been formed. The clusters are formed using k-means clustering. The cluster-1 contains 41 sequences (S81, S82, S62, S7, S11, S16, S17, S18, S19, S20, S24, S25, S26, S27, S29, S31, S32, S34, S48, S49, S50, S51, S64, S65, S66, S72, S73, S87, S8, S23, S9, S10, S83, S85, S44, S70, S80, S58, S59, S53, S45) having centre at 0.6464. The cluster-2 contains 11 sequences (S39, S40, S41, S60, S74, S13, S14, S89, S57, S12, S3) having centre at 0.6494. The cluster 3 and 4 contain only one sequence each viz. S1, S47 respectively centres at $0.7125, 0.6431$ respectively. The cluster-5 contains 11 sequences (S37, S77, S55, S56, S86, S61, S54, S46, S52, S6, S36) having centre at 0.6448. The sequence S30 belongs to the cluster S6 whose centre is at 0.6554. The sequences (S68, S78, S88, S71, S67, S42, S15, S69) are contained in the cluster-7 whose centre is at 0.6483. The cluster-8 contains sequences (S43, S63, S2, S5, S21, S22, S33, S35, S38, S84) whose center is at 0.6460. The sequence S4 belongs to the cluster-9 whose centre is at 0.6516. The cluster 10 contains four sequences S79, S75, S76, S28 whose centre is at 0.6441. It is noted that the sequences S55 ands S66 have exactly same HE 0.6445500767, which confirms their identical long-range correlation though the length of these two sequence (S55 is of length 29870 and S66 is of length 29872) is differed by 2 bp. Also it is seen that the sequences S21, S22, S33, S35 belonging to the cluster-8 have same HE 0.6460659477. For all the sequences S81, S82, S62, S7, S11, S16, S17, S18, S19, S20, S24, S25, S26, S27, S29, S31, S32, S34, S48, S49, S50, S51, S64, S65, S66, S72, S73 and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

S87 belonging to the cluster-1 have the HE 0.6463681216. In the same cluster-1, there are three sequences S9, S10, S83 having the same HE 0.6464832466. There are four sequences S39, S40, S41 and S60 which are contained in the cluster-2, having same HE 0.6491763266.

3.3. Classification Based on Shannon Entropy

For all the 89 binary sequences (purine-pyrimidine) of SARS-CoV2, the Shannon entropy (SE) are determined and then ten different clusters are formed based on SE obtained for all the sequences. The Shannon entropy and the histograms of all the SARS-CoV2 sequences are plotted in the Fig. 3.

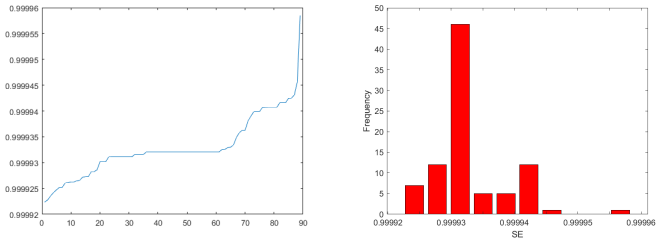


Figure 3: Plot of the Shannon entropy (SE) and corresponding histogram of all the purine-pyrimidine binary sequences corresponding to SARS-CoV2 sequences.

It is obtained that the SE is ranging from 0.9999 to 1, i.e. the length of the range is too small. The SE is precisely same for all the sequences which is 0.9999 except one sequence S30 which is of length 29945. This confirms that the amount of uncertainly is at maximum and the probability of occurrence of purine and pyrimidine bases across the sequence S30 is 0.5. Although the sequence is not randomly composed (positively autocorrelated with HE 0.65538) of nucleotide bases, the purine and pyrimidine bases are composed with equal probability.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

Table 3: Hurst exponent of all the 89 purine-pyrimidine binary sequences corresponding to SARS-CoV2 sequences

Sequence	HE	Sequence	HE	Sequence	HE
S47	0.6430890415	S11	0.6463681216	S44	0.6465117331
S79	0.6438083222	S16	0.6463681216	S70	0.6466330524
S75	0.6439455807	S17	0.6463681216	S80	0.6466330524
S76	0.644253158	S18	0.6463681216	S58	0.6466886489
S28	0.6443251468	S19	0.6463681216	S59	0.6467842124
S37	0.6444723241	S20	0.6463681216	S53	0.6468519632
S77	0.644522348	S24	0.6463681216	S45	0.6470843102
S55	0.6445500767	S25	0.6463681216	S68	0.6475613391
S56	0.6445500767	S26	0.6463681216	S78	0.6477094551
S86	0.6445721887	S27	0.6463681216	S88	0.6477094551
S61	0.6447294561	S29	0.6463681216	S71	0.6483049547
S54	0.6447448006	S31	0.6463681216	S67	0.6487736204
S46	0.6447796066	S32	0.6463681216	S42	0.6488825729
S52	0.6448639682	S34	0.6463681216	S15	0.6488825729
S6	0.6452720329	S48	0.6463681216	S69	0.6488862311
S36	0.6453293176	S49	0.6463681216	S39	0.6491763266
S43	0.6456954376	S50	0.6463681216	S40	0.6491763266
S63	0.6457628936	S51	0.6463681216	S41	0.6491763266
S2	0.6459898777	S64	0.6463681216	S60	0.6491763266
S5	0.6459943594	S65	0.6463681216	S74	0.6491856177
S21	0.6460659477	S66	0.6463681216	S13	0.6494006145
S22	0.6460659477	S72	0.6463681216	S14	0.6494692533
S33	0.6460659477	S73	0.6463681216	S89	0.6494692533
S35	0.6460659477	S87	0.6463681216	S57	0.6497613576
S38	0.6460690307	S8	0.6464119458	S12	0.6497671065
S84	0.6461814077	S23	0.6464158218	S3	0.6501090217
S81	0.6462179229	S9	0.6464832466	S4	0.651582672
S82	0.6462179229	S10	0.6464832466	S30	0.6553858343
S62	0.6463189347	S83	0.6464832466	S1	0.7124517615
S7	0.6463681216	S85	0.6464832466		

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

Table 4: Shannon entropy (SE) of all the purine-pyrimidine binary sequences corresponding to SARS-CoV2 sequences.

Sequence	SE	Sequence	SE	Sequence	SE
S47	0.9999223787	S85	0.9999311193	S87	0.9999320596
S28	0.9999227878	S81	0.9999315857	S62	0.9999325138
S79	0.9999235693	S82	0.9999315857	S44	0.999932568
S75	0.9999242037	S70	0.9999315948	S58	0.9999328857
S77	0.9999247235	S80	0.9999315948	S45	0.9999329935
S37	0.9999252013	S7	0.9999320596	S23	0.9999333912
S76	0.9999252013	S11	0.9999320596	S53	0.9999348593
S46	0.9999260645	S16	0.9999320596	S68	0.9999357908
S86	0.9999261031	S17	0.9999320596	S78	0.9999362496
S55	0.9999262613	S18	0.9999320596	S88	0.9999362496
S56	0.9999262613	S19	0.9999320596	S71	0.9999380434
S54	0.9999264586	S20	0.9999320596	S13	0.999938992
S61	0.999926567	S24	0.9999320596	S69	0.9999398601
S52	0.999927186	S25	0.9999320596	S15	0.9999398762
S43	0.999927264	S26	0.9999320596	S42	0.9999398762
S6	0.9999272835	S27	0.9999320596	S74	0.9999407143
S5	0.9999282594	S29	0.9999320596	S67	0.9999407381
S8	0.9999282594	S31	0.9999320596	S39	0.9999407539
S36	0.999928592	S32	0.9999320596	S40	0.9999407539
S63	0.9999301724	S34	0.9999320596	S41	0.9999407539
S84	0.9999301724	S48	0.9999320596	S60	0.9999407539
S2	0.9999301848	S49	0.9999320596	S1	0.9999416252
S38	0.9999311008	S50	0.9999320596	S14	0.9999416252
S9	0.9999311193	S51	0.9999320596	S89	0.9999416252
S10	0.9999311193	S59	0.9999320596	S12	0.9999424669
S21	0.9999311193	S64	0.9999320596	S57	0.99994249
S22	0.9999311193	S65	0.9999320596	S3	0.999943128
S33	0.9999311193	S66	0.9999320596	S4	0.9999456377
S35	0.9999311193	S72	0.9999320596	S30	0.9999585474
S83	0.9999311193	S73	0.9999320596		

 doi:10.20944/preprints202003.0344.v1

Having all the SE of the binary representation of purine and pyrimidine of the SARS-CoV sequences, only three culsters have been formed using k-means clustering technique. The cluster-1 contains 21 sequences S68, S78, S88, S71, S13, S69, S15, S42, S74, S67, S39, S40, S41, S60, S1, S14, S89, S12, S57, S3 and S4 having SE centred at 0.999940381147619. The other 67 sequences belong to the other cluster-2 whose centre is at 0.999930184068656. Though these two clusters can be considered same. There is only one cluster-3 which contains only one sequence S30 whose SE is 0.9999585474 (approximately 1) as already mentioned before.

It is worth mentioning that the SE is very much linear in trend for all these purine and pyrimidine distribution among the SARS-CoV2 sequences. This is something is crucial in Coronavirus (SARS-CoV2) unlike other sequences as obtained in previous studies made [35, 36, 37]. The amount of uncertainly is at maximum which says the equally likely occurrence of purine and pyrimidine bases across the sequences among all the SARS-CoV2.

4. GC, A, T, C and G Density in the SARS-CoV2

In this section, we shall try to investigate the density of each nucleotides and also the GC content in the SARS-CoV2 sequences. Based on density, the sequences are classified as follows. Here we present entire detail list of percentage of density obtained for the GC content among all the SARS-CoV2 sequences as shown in the Table 4, 5 and 6. It is found that the density of GC content is around 37.5% which says that the SARS-CoV2 sequences are essentially AT rich. That is one purine base nucleotide (A) and one pyrimidine base nucleotide (T) are rich (approximately 30% A and 32% T) in these sequences of SARS-CoV2 and as mentioned in the Shannon entropy subsection the occurrence of purine and pyrimidine bases are equally probable. This is what is significant speciality of the SARS-CoV2 sequences.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

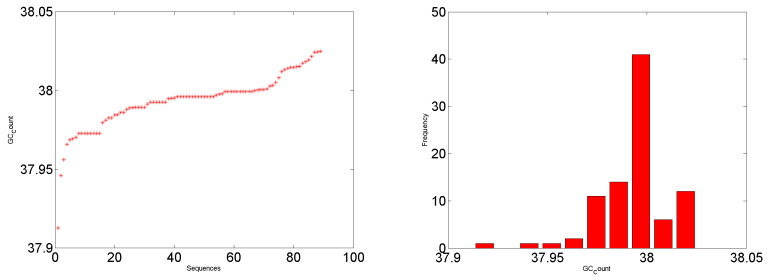


Figure 4: Plot of the GC content density and its corresponding histogram of the SARS-CoV2 sequences.

Based on the GC content density in the SARS-CoV2 sequences, ten different clusters are formed using k-means clustering technique and the following Table 7 describes the sequences and their corresponding clusters where they belong.

There ten clusters (C) having centres at 37.9460, 38.0143, 37.9826, 37.9952, 38.0002, 37.9714, 37.9888, 38.0230, 37.9561 and 37.9128. The density of all these sequences lies in the interval $(37.91284, 38.02505)$. The cluster-10, 9 and 1 contain only one sequence S30, S13 and S60 which has 37.91284%, 37.94602% and 37.95605% of GC contents respectively. It is noted that the sequence S30 does have the SE 1 as pointed earlier.

Following in the Figures 4, 5, 6 and 7, we have given the plot of the percentage of A, T, C and G with their respective histograms.

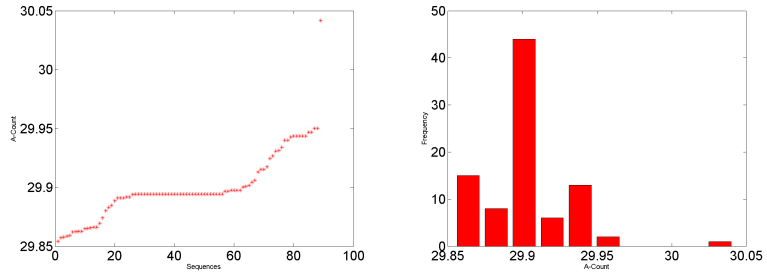


Figure 5: Plot of the A content density and its corresponding histogram of the SARS-CoV2 sequences.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

Table 5: Sequences and their respective percentage of the GC content.

Sequence	% of G,C	% of A	% of T	% of C	% of G
S1	37.972778651	29.9434839314	32.0837374177	18.3660502291	19.6067284219
S2	38.0082597455	29.8626733371	32.1290669174	18.36282443	19.6454353154
S3	37.9795610655	29.9313117775	32.089127157	18.3548333054	19.6247277601
S4	37.9889391654	29.9245852187	32.0864756159	18.3475783476	19.6413608178
S5	37.9860785757	29.8942507195	32.1163242086	18.3789572318	19.607121344
S6	37.9953145917	29.8828647925	32.1151271754	18.3801874163	19.6151271754
S7	37.9994645606	29.8942507195	32.1062847199	18.3789572318	19.6205073288
S8	37.9659995984	29.8942507195	32.1196707048	18.3588782545	19.607121344
S9	37.9827320795	29.8975972157	32.1062847199	18.3689177431	19.6138143364
S10	37.9961180644	29.8942507195	32.1096312161	18.3789572318	19.6171608326
S11	37.9994645606	29.8942507195	32.1062847199	18.3789572318	19.6205073288
S12	37.9703649196	29.9428036258	32.0868314547	18.3597016423	19.6106632773
S13	37.9460254824	29.9501722235	32.1038022941	18.3560177909	19.5900076915
S14	37.972778651	29.9401397853	32.0870815637	18.362706083	19.610072568
S15	37.972778651	29.926763201	32.0971140019	18.3560177909	19.6167608601
S16	37.9994645606	29.8942507195	32.1062847199	18.3789572318	19.6205073288
S17	37.9994645606	29.8942507195	32.1062847199	18.3789572318	19.6205073288
S18	37.9994645606	29.8942507195	32.1062847199	18.3789572318	19.6205073288
S19	37.9961180644	29.8942507195	32.1096312161	18.3756107356	19.6205073288
S20	37.9894250719	29.8942507195	32.1163242086	18.3689177431	19.6205073288
S21	37.9961180644	29.8942507195	32.1096312161	18.3789572318	19.6171608326
S22	37.9994645606	29.8909042233	32.1096312161	18.3789572318	19.6205073288
S23	37.9951116617	29.9059162286	32.0989721097	18.381491278	19.6136203837
S24	37.9927715682	29.8975972157	32.1096312161	18.3756107356	19.6171608326
S25	37.9994645606	29.8942507195	32.1062847199	18.3789572318	19.6205073288
S26	37.9994645606	29.8942507195	32.1062847199	18.3789572318	19.6205073288
S27	37.9961180644	29.8942507195	32.1096312161	18.3756107356	19.6205073288
S28	38.0051561925	29.8623899287	32.1291057019	18.3848394549	19.6203167375
S29	37.9894250719	29.8975972157	32.1129777123	18.3722642393	19.6171608326

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

Table 6: Sequences and their respective percentage of the GC content.

Sequence	% of G,C	% of A	% of T	% of C	% of G
S30	37.912840207	30.0417431959	32.0454165971	18.3336116213	19.5792285857
S31	37.9994645606	29.8942507195	32.1062847199	18.3789572318	19.6205073288
S32	37.9961180644	29.8942507195	32.1096312161	18.3756107356	19.6205073288
S33	37.9927715682	29.8942507195	32.1129777123	18.3756107356	19.6171608326
S34	37.9894250719	29.8942507195	32.1163242086	18.3689177431	19.6205073288
S35	37.9927715682	29.8942507195	32.1129777123	18.3756107356	19.6171608326
S36	38.0121268969	29.8649961475	32.1228769555	18.374593816	19.637533081
S37	38.0183559992	29.8619950425	32.1196489583	18.389495545	19.6288604542
S38	37.9978579557	29.8848651181	32.1172769262	18.3713769329	19.6264810228
S39	37.972778651	29.9434839314	32.0837374177	18.3693943751	19.6033842758
S40	37.972778651	29.9434839314	32.0837374177	18.3693943751	19.6033842758
S41	37.972778651	29.9434839314	32.0837374177	18.3693943751	19.6033842758
S42	37.972778651	29.9401397853	32.0870815637	18.3693943751	19.6033842758
S43	38.0004016602	29.880171375	32.1194269648	18.3826482796	19.6177533806
S44	37.9980596166	29.8885952293	32.1133451541	18.3700779499	19.6279816667
S45	37.9860785757	29.9042902082	32.1096312161	18.3722642393	19.6138143364
S46	38.0132981389	29.8740352167	32.1126666444	18.3935313575	19.6197667814
S47	38.0140755812	29.865581535	32.1203428838	18.3983189353	19.6157566459
S48	37.9961180644	29.8942507195	32.1096312161	18.3756107356	19.6205073288
S49	37.9927715682	29.9009437119	32.1062847199	18.3789572318	19.6138143364
S50	38.0028110568	29.8909042233	32.1062847199	18.3789572318	19.623853825
S51	37.9994645606	29.8942507195	32.1062847199	18.3789572318	19.6205073288
S52	38.0174146015	29.8693904889	32.1131949096	18.3891493637	19.6282652378
S53	38.0010707355	29.9002877602	32.0986415044	18.3764973566	19.6245733788
S54	38.0195229949	29.8648150012	32.115662004	18.3891852001	19.6303377948
S55	38.0147304988	29.8627385336	32.1225309675	18.3829929695	19.6317375293
S56	38.0147304988	29.8660863743	32.1191831269	18.3863408102	19.6283896887
S57	37.9694345049	29.9501722235	32.0803932716	18.3660502291	19.6033842758
S58	38.0032152187	29.9015339273	32.095250854	18.3870319512	19.6161832675
S59	37.9961180644	29.8975972157	32.1062847199	18.3789572318	19.6171608326
S60	37.9560579206	29.9468280775	32.0971140019	18.3560177909	19.6000401298

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

Table 7: Sequences and their respective percentage of the GC content.

Sequence	% of G,C	% of A	% of T	% of C	% of G
S61	38.0216538732	29.8662554889	32.1120906379	18.3923842725	19.6292696008
S62	37.9999330634	29.8939054185	32.1061615181	18.3774557381	19.6224773252
S63	37.9894250719	29.8909042233	32.1163242086	18.3722642393	19.6171608326
S64	37.9961180644	29.8942507195	32.1096312161	18.3756107356	19.6205073288
S65	37.9927715682	29.8942507195	32.1129777123	18.3722642393	19.6205073288
S66	37.9961180644	29.8942507195	32.1096312161	18.3756107356	19.6205073288
S67	37.9845479782	29.9307669153	32.0846851065	18.3685073079	19.6160406703
S68	37.9892940783	29.9130143861	32.0976915356	18.3740381398	19.6152559384
S69	37.9812033847	29.9341115087	32.0846851065	18.3718519014	19.6093514833
S70	37.9948465683	29.8965967272	32.1085567045	18.3783422013	19.616504367
S71	37.9972565158	29.917360902	32.0853825822	18.3779985948	19.619257921
S72	37.9961180644	29.8942507195	32.1096312161	18.3756107356	19.6205073288
S73	37.9961180644	29.8942507195	32.1096312161	18.3756107356	19.6205073288
S74	37.968755227	29.94681029	32.084434483	18.3688488944	19.5999063326
S75	38.0242529814	29.8572959936	32.1184510251	18.3940774487	19.6301755326
S76	38.025055269	29.8586454077	32.1162993234	18.3928451799	19.6322100891
S77	38.0245838497	29.8589945406	32.1164216097	18.3943463844	19.6302374653
S78	37.9880231508	29.9153591382	32.096617711	18.3734234385	19.6145997123
S79	38.0150880134	29.8541492037	32.1307627829	18.383906119	19.6311818944
S80	37.9915001841	29.8965967272	32.1119030887	18.374995817	19.616504367
S81	38.0007362538	29.8919045547	32.1073591915	18.3795723035	19.6211639503
S82	38.0007362538	29.8919045547	32.1073591915	18.3795723035	19.6211639503
S83	37.9927715682	29.8942507195	32.1129777123	18.3756107356	19.6171608326
S84	37.9961180644	29.8942507195	32.1096312161	18.382303728	19.6138143364
S85	37.9827320795	29.8942507195	32.0962452312	18.3655712469	19.6171608326
S86	38.0152825256	29.857899323	32.1268181514	18.3792479389	19.6360345868
S87	37.9961180644	29.8942507195	32.1096312161	18.3756107356	19.6205073288
S88	37.9846776622	29.9153591382	32.0999631996	18.3700779499	19.6145997123
S89	37.972778651	29.9434839314	32.0837374177	18.3660502291	19.6067284219

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

Table 8: Sequences and their respective percentage of the GC content and clusters where they belong.

Seq	% of GC	C	Seq	% of GC	C	Seq	% of GC	C
S30	37.912840207	10	S24	37.9927715682	4	S22	37.9994645606	5
S13	37.9460254824	1	S33	37.9927715682	4	S25	37.9994645606	5
S60	37.9560579206	9	S35	37.9927715682	4	S26	37.9994645606	5
S8	37.9659995984	6	S49	37.9927715682	4	S31	37.9994645606	5
S74	37.968755227	6	S65	37.9927715682	4	S51	37.9994645606	5
S57	37.9694345049	6	S83	37.9927715682	4	S62	37.9999330634	5
S12	37.9703649196	6	S70	37.9948465683	4	S43	38.0004016602	5
S1	37.972778651	6	S23	37.9951116617	4	S81	38.0007362538	5
S14	37.972778651	6	S6	37.9953145917	4	S82	38.0007362538	5
S15	37.972778651	6	S10	37.9961180644	4	S53	38.0010707355	5
S39	37.972778651	6	S19	37.9961180644	4	S50	38.0028110568	5
S40	37.972778651	6	S21	37.9961180644	4	S58	38.0032152187	5
S41	37.972778651	6	S27	37.9961180644	4	S28	38.0051561925	5
S42	37.972778651	6	S32	37.9961180644	4	S2	38.0082597455	2
S89	37.972778651	6	S48	37.9961180644	4	S36	38.0121268969	2
S3	37.9795610655	3	S59	37.9961180644	4	S46	38.0132981389	2
S69	37.9812033847	3	S64	37.9961180644	4	S47	38.0140755812	2
S9	37.9827320795	3	S66	37.9961180644	4	S55	38.0147304988	2
S85	37.9827320795	3	S72	37.9961180644	4	S56	38.0147304988	2
S67	37.9845479782	3	S73	37.9961180644	4	S79	38.0150880134	2
S88	37.9846776622	3	S84	37.9961180644	4	S86	38.0152825256	2
S5	37.9860785757	7	S87	37.9961180644	4	S52	38.0174146015	2
S45	37.9860785757	7	S71	37.9972565158	4	S37	38.0183559992	2
S78	37.9880231508	7	S38	37.9978579557	5	S54	38.0195229949	8
S4	37.9889391654	7	S44	37.9980596166	5	S61	38.0216538732	8
S68	37.9892940783	7	S7	37.9994645606	5	S75	38.0242529814	8
S20	37.9894250719	7	S11	37.9994645606	5	S77	38.0245838497	8
S29	37.9894250719	7	S16	37.9994645606	5	S76	38.025055269	8
S34	37.9894250719	7	S17	37.9994645606	5			
S63	37.9894250719	7	S18	37.9994645606	5			
S80	37.9915001841	7						

Table 9: Sequences and their respective percentage of the GC content and clusters where they belong.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

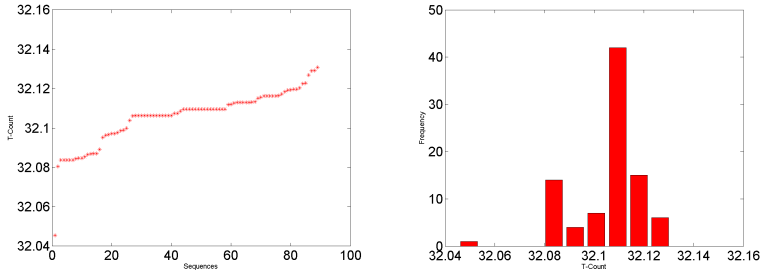


Figure 6: Plot of the T content density and its corresponding histogram of the SARS-CoV2 sequences.

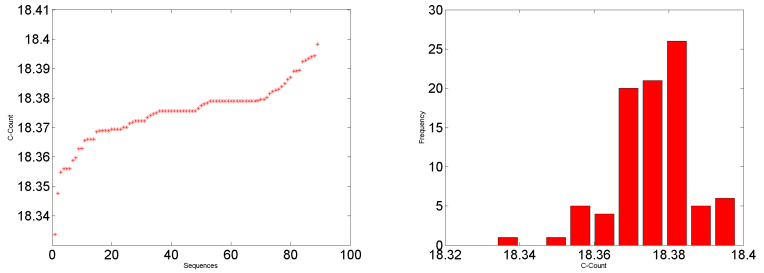


Figure 7: Plot of the C content density and its corresponding histogram of the SARS-CoV2 sequences.

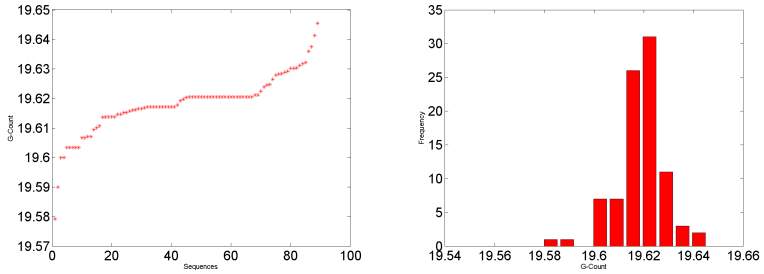


Figure 8: Plot of the G content density and its corresponding histogram of the SARS-CoV2 sequences.

In the following Table 8 the percentage of density in the SARS-CoV2 se-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

quences of A, T, C and G are given explicitly.

Table 10: Sequences and their respective percentage of the A, T, C and G contents.

Seq	% A	% T	% C	% G	Seq	% A	% T	% C	% G	Seq	% A	% T	% C	% G
S1	29.94	32.08	18.37	19.61	**S31**	29.89	32.11	18.38	19.62	**S61**	29.87	32.11	18.39	19.63
S2	29.86	32.13	18.36	19.65	**S32**	29.89	32.11	18.38	19.62	**S62**	29.89	32.11	18.38	19.62
S3	29.93	32.09	18.35	19.62	**S33**	29.89	32.11	18.38	19.62	**S63**	29.89	32.12	18.37	19.62
S4	29.92	32.09	18.35	19.64	**S34**	29.89	32.12	18.37	19.62	**S64**	29.89	32.11	18.38	19.62
S5	29.89	32.12	18.38	19.61	**S35**	29.89	32.11	18.38	19.62	**S65**	29.89	32.11	18.37	19.62
S6	29.88	32.12	18.38	19.62	**S36**	29.86	32.12	18.37	19.64	**S66**	29.89	32.11	18.38	19.62
S7	29.89	32.11	18.38	19.62	**S37**	29.86	32.12	18.39	19.63	**S67**	29.93	32.08	18.37	19.62
S8	29.89	32.12	18.36	19.61	**S38**	29.88	32.12	18.37	19.63	**S68**	29.91	32.10	18.37	19.62
S9	29.90	32.11	18.37	19.61	**S39**	29.94	32.08	18.37	19.60	**S69**	29.93	32.08	18.37	19.61
S10	29.89	32.11	18.38	19.62	**S40**	29.94	32.08	18.37	19.60	**S70**	29.90	32.11	18.38	19.62
S11	29.89	32.11	18.38	19.62	**S41**	29.94	32.08	18.37	19.60	**S71**	29.92	32.09	18.38	19.62
S12	29.94	32.09	18.36	19.61	**S42**	29.94	32.09	18.37	19.60	**S72**	29.89	32.11	18.38	19.62
S13	29.95	32.10	18.36	19.59	**S43**	29.88	32.12	18.38	19.62	**S73**	29.89	32.11	18.38	19.62
S14	29.94	32.09	18.36	19.61	**S44**	29.89	32.11	18.37	19.63	**S74**	29.95	32.08	18.37	19.60
S15	29.93	32.10	18.36	19.62	**S45**	29.90	32.11	18.37	19.61	**S75**	29.86	32.12	18.39	19.63
S16	29.89	32.11	18.38	19.62	**S46**	29.87	32.11	18.39	19.62	**S76**	29.86	32.12	18.39	19.63
S17	29.89	32.11	18.38	19.62	**S47**	29.87	32.12	18.40	19.62	**S77**	29.86	32.12	18.39	19.63
S18	29.89	32.11	18.38	19.62	**S48**	29.89	32.11	18.38	19.62	**S78**	29.92	32.10	18.37	19.61
S19	29.89	32.11	18.38	19.62	**S49**	29.90	32.11	18.38	19.61	**S79**	29.85	32.13	18.38	19.63
S20	29.89	32.12	18.37	19.62	**S50**	29.89	32.11	18.38	19.62	**S80**	29.90	32.11	18.37	19.62
S21	29.89	32.11	18.38	19.62	**S51**	29.89	32.11	18.38	19.62	**S81**	29.89	32.11	18.38	19.62
S22	29.89	32.11	18.38	19.62	**S52**	29.87	32.11	18.39	19.63	**S82**	29.89	32.11	18.38	19.62
S23	29.91	32.10	18.38	19.61	**S53**	29.90	32.10	18.38	19.62	**S83**	29.89	32.11	18.38	19.62
S24	29.90	32.11	18.38	19.62	**S54**	29.86	32.12	18.39	19.63	**S84**	29.89	32.11	18.38	19.61
S25	29.89	32.11	18.38	19.62	**S55**	29.86	32.12	18.38	19.63	**S85**	29.89	32.10	18.37	19.62
S26	29.89	32.11	18.38	19.62	**S56**	29.87	32.12	18.39	19.63	**S86**	29.86	32.13	18.38	19.64
S27	29.89	32.11	18.38	19.62	**S57**	29.95	32.08	18.37	19.60	**S87**	29.89	32.11	18.38	19.62
S28	29.86	32.13	18.38	19.62	**S58**	29.90	32.10	18.39	19.62	**S88**	29.92	32.10	18.37	19.61
S29	29.90	32.11	18.37	19.62	**S59**	29.90	32.11	18.38	19.62	**S89**	29.94	32.08	18.37	19.61
S30	30.04	32.05	18.33	19.58	**S60**	29.95	32.10	18.36	19.60					

In the above Table 8, it is observed that the intervals where the density of A, T, C and G lie are (29.85, 30.04), (32.05, 32.13), (18.33, 18.40) and (19.58, 19.65) respectively. That is the approximately A, T, C and G are spread over these SARS-CoV2 sequences in 30%, 32%, 18% and 19% respectively. It is noted that the density of A and T are significantly rich as seen here. This illustrates the density of purine and pyrimidine bases are kept almost same as confirmed in SE previously.

All the sequences of SARS-CoV2 sequences are clustered into different clusters. The centre of each cluster in all the four cases (A, T, C and G) is differed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

by 0.01 distance. The sequence $S79$ has the least percentage (29.85%) of the nucleotide base A where as the sequence S30 has least percentage of T, C and G densities. It is also observed that S79, S47 and S2 have the highest percentages (32.13%) of T density, 18.40% of C density and 19.65% of G density respectively.

Following we are yet to discover the purine-pyrimidine closeness of the SARS-CoV2 genomes based on Hamming distances. This would enable to cluster the sequences into some clusters based on the closeness of purine-pyrimidine sequences similarity.

4.1. Hamming Distance of the SARS-COV2

The similarity analysis of the SARS-CoV2 sequences have been measured by calculating the distance between the vectors of binary strings encoded on the basis of purines and pyrimidines nucleotide bases as mentioned earlier. There are several computing methods for measuring the distance between multidimensional vectors, such as Hamming Distance, Euclidean distance, Elastic-matching distance, Jeffrey's and Matusita distance, Manhattan distance andMinkowski norm. Reportedly, these methods have little effect on the similarity of vectors [38]. The Hamming Distance (HD) between two binary strings is the number of bits in which they differ [39, 40, 41]. Since length of the different $SARS-CoV2$ genome usually differ by some bases and hence a special care has been taken into consideration. Suppose there are two $SARS-CoV2$ S_x^1 and S_y^2 of length x and y respectively $(x > y)$, then

$$HD(S_x^1, S_y^2) = hd(S_y^1, S_y^2)$$

For example, take two binary sequences $S_x = 101011$ and $S_m = 0010$, of minimum length 4, from left to right alignment of these two sequences, we find the hamming distances are $hd(101011, 0010) = 1$, Finding the minimum hamming distance of the two binary sequences says about the maximum similarity of two sequences over the distribution of purines and pyrimidines. The minimum value of $HD = 0$ when the pattern of length $min(x, y)$ of two binary

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

sequences of are exactly identical i.e. similar distribution of purines and pyrimidines over the $SARS-CoV2$ of the two sequences and the maximum value of $HD = min(n, m)$ when the pattern of length $min(n, m)$ of two binary sequences of $SARS-CoV2$ are exactly opposite i.e. completely dissimilar distribution of purines and pyrimidines over $SARS-CoV2$ two sequences.To get the nearness of the $SARS-CoV2$ based on their purine-pyrimidine distribution, minimum Hamming distance is deployed.

In order to demonstrate the methodology, the measure of distances (Hamming distance) among the 89 SARS-CoV2 sequences as depicted in the Table 10 are taken into consideration. It is noted that if two virus sequences are having large hamming distance between them then it inferences that these two sequences are unlikely related to each other. From the Figure 9, the following conclusions can be drawn, the SARS-COV2 virus sequences MT044258(S59), MN994468 (S84), NC_045512(S1), ... MN039888(S60) are grouped together as a single cluster as the distance between them is almost negligible and it indicates the closeness among them. Also it should be noted that the sequences MT152824(S38), MN996531(S76), MT012098(S36) and MT975262(S86) are closely related to each other and therefore is treated as a cluster. Similarly the sequences MT163719(S13), MT007544(S74), MT03988(S62), MT188341(S2), MT188339(S3), MT188340 (S4), MN123290(S45), MT039873(S61), MT159721(S23) and MN072688(S54) depicts similar Hamming distances and are grouped together as shown in Figure 9. This closeness(nearness) among the SARS-CoV2 genomes would enable future such genomes or other Blasted results to get into the clusters quantitatively instead of just by sequential similarity.

5. Conclusions and Summary

It is needless to mention that the novel coronavirus has led to a public health emergency of world concern according to WHO ($https://www.who.int/$). One of the major reasons for such a global threat is due to the lack of quantitative as well as qualitative knowledge about this novel virus including its genomic and

 doi:10.20944/preprints202003.0344.v1

proteomic levels.

In this article, an attempt has been made to clarify the quantitative nature of the SARS-CoV complete sequences. This present study also reveals the closeness among the 89 complete sequences in the purine-pyrimidine level descriptions through phylogenetic analysis. Also one of the major fact of the 89 SARS-CoV sequences have been exposed that the purine and pyrimidine distribution among all these genes are evenly-equally spatially placed though the GC content is significantly low as described in the result. We believe this quantitative piece of information would enable researcher to comprehend the genomic description of the SARS-CoV sequences better and would atleast help passively in ensuring proper healthcare facility against this massive global emergency. In our future endeavour, we wish to understand the proteins of the SARS-CoV2.

Authors Contributions and Conflicts of Interest:

The author SH has formulated and carried out the study with RKR and VS. The authors SH and RKR analyse the study and written the manuscript and finally all the three authors checked and approved the manuscript. The authors declare that there is no conflicts of interest.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2020 doi:10.20944/preprints202003.0344.v1

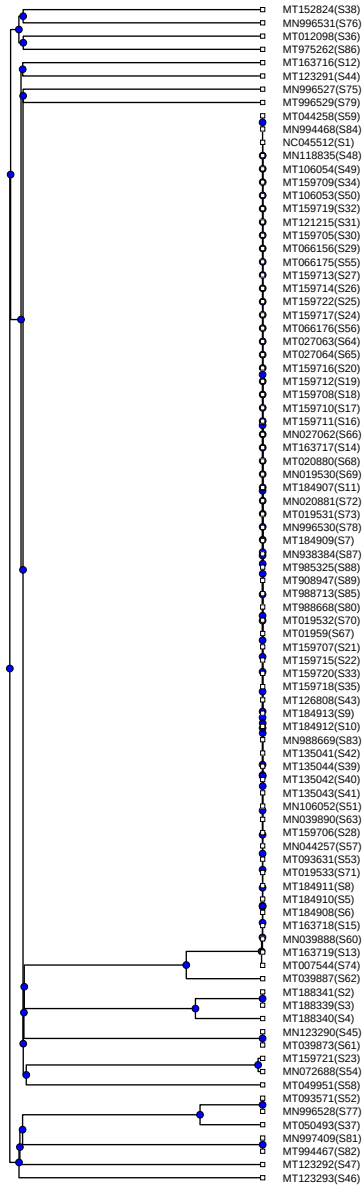
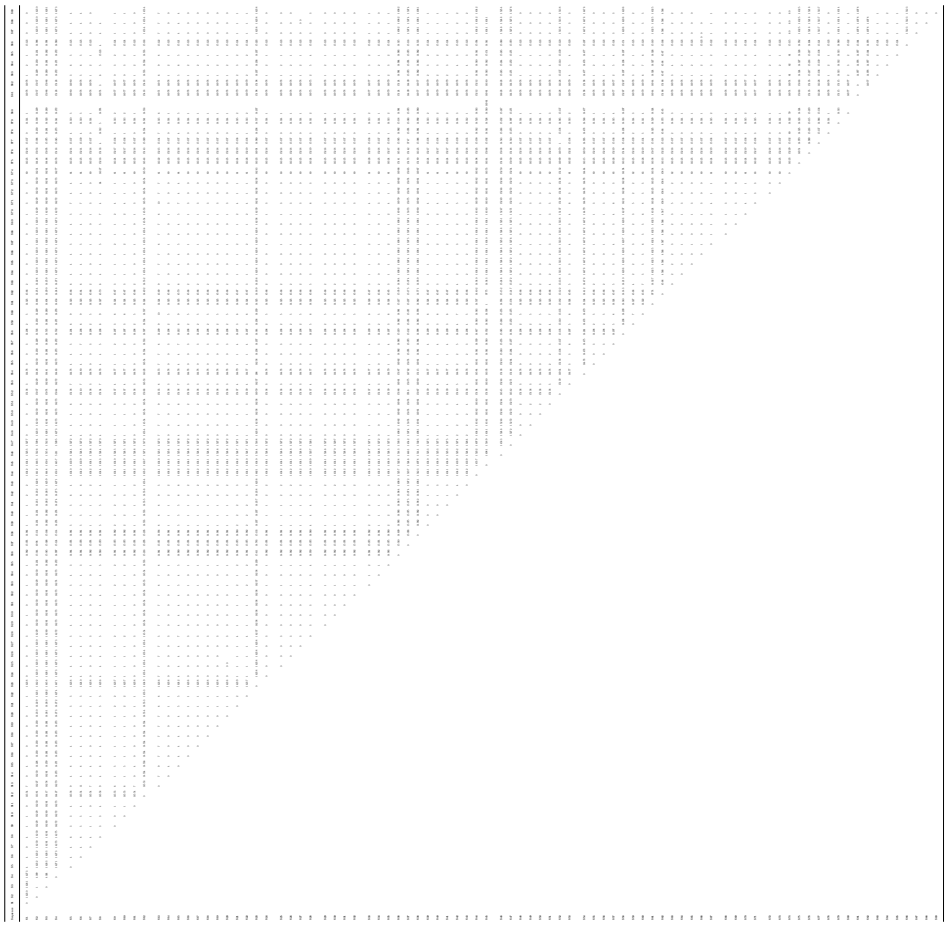


Figure 9: Phylogenetic tree of SARS-CoV2 sequences based on the distribution of purines and pyrimidines (cluster dendrogram using UPGMA distance method)

 doi:10.20944/preprints202003.0344.v1

Table 11: Hamming distance Matrix of the SARS-CoV2 sequences



References

[1] K. V. Holmes, Sars-associated coronavirus, New England Journal of Medicine 348 (20) (2003) 1948–1951.

[2] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, P. M. Wertheim-van Dillen, J. Kaandorp, J. Spaargaren, B. Berkhout, Identification of a new human coronavirus, Nature medicine 10 (4) (2004) 368–373.

[3] M. Lipsitch, D. L. Swerdlow, L. Finelli, Defining the epidemiology of covid-19—studies needed, New England Journal of Medicine (2020).

[4] A. S. Fauci, H. C. Lane, R. R. Redfield, Covid-19—navigating the uncharted (2020).

[5] W. Liu, Q. Zhang, J. Chen, R. Xiang, H. Song, S. Shu, L. Chen, L. Liang, J. Zhou, L. You, et al., Detection of covid-19 in children in early january 2020 in wuhan, china, New England Journal of Medicine (2020).

[6] F. Jiang, L. Deng, L. Zhang, Y. Cai, C. W. Cheung, Z. Xia, Review of the clinical characteristics of coronavirus disease 2019 (covid-19), Journal of General Internal Medicine (2020) 1–5.

[7] J. Stebbing, A. Phelan, I. Griffin, C. Tucker, O. Oechsle, D. Smith, P. Richardson, Covid-19: combining antiviral and anti-inflammatory treatments, The Lancet Infectious Diseases (2020).

[8] J. F.-W. Chan, C. C.-Y. Yip, K. K.-W. To, T. H.-C. Tang, S. C.-Y. Wong, K.-H. Leung, A. Y.-F. Fung, A. C.-K. Ng, Z. Zou, H.-W. Tsoi, et al., Improved molecular diagnosis of covid-19 by the novel, highly sensitive and specific covid-19-rdrp/hel real-time reverse transcription-polymerase chain reaction assay validated in vitro and with clinical specimens, Journal of Clinical Microbiology (2020).

[9] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, R. Agha, World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19), International Journal of Surgery (2020).

[10] M. A. Marra, S. J. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, et al., The genome sequence of the sars-associated coronavirus, Science 300 (5624) (2003) 1399–1404.

[11] P. Sun, X. Lu, C. Xu, W. Sun, B. Pan, Understanding of covid-19 based on current evidence, Journal of Medical Virology (2020).

[12] S. Zhang, M. Y. Diao, L. Duan, Z. Lin, D. Chen, The novel coronavirus (sars-cov-2) infections in china: prevention, control and challenges, Intensive Care Medicine (2020) 1–3.

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

[14] G. Kampf, D. Todt, S. Pfaender, E. Steinmann, Persistence of coronaviruses on inanimate surfaces and its inactivation with biocidal agents, Journal of Hospital Infection (2020).

[15] S. Khan, A. Ali, R. Siddique, G. Nabi, Novel coronavirus is putting the whole world on alert, Journal of Hospital Infection 104 (3) (2020) 252–253.

[16] J. Xu, S. Zhao, T. Teng, A. E. Abdalla, W. Zhu, L. Xie, Y. Wang, X. Guo, Systematic comparison of two animal-to-human transmitted human coronaviruses: Sars-cov-2 and sars-cov, Viruses 12 (2) (2020) 244.

[17] W.-B. Yu, G.-D. Tang, L. Zhang, R. T. Corlett, Decoding the evolution and transmissions of the novel pneumonia coronavirus (sars-cov-2) using whole genomic data, ChinaXiv 202002 (2020) v2.

[18] C. Cattani, G. Pierro, On the fractal geometry of dna by the binary image analysis, Bulletin of Mathematical Biology 75 (9) (2013) 1544–1570.

[19] C. Cattani, Fractals and hidden symmetries in dna, Mathematical problems in engineering 2010 (2010).

[20] S. S. Hassan, P. P. Choudhury, B. Daya Sagar, S. Chakraborty, R. Guha, A. Goswami, Quantitative description of genomic evolution of olfactory receptors, Asian-European Journal of Mathematics 8 (03) (2015) 1550043.

[21] R. K. Rout, P. Pal Choudhury, S. P. Maity, B. Daya Sagar, S. S. Hassan, Fractal and mathematical morphology in intricate comparison between tertiary protein structures, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 6 (2) (2018) 192–203.

[22] C. L. Berthelsen, J. A. Glazier, M. H. Skolnick, Global fractal dimension of human dna sequences treated as pseudorandom walks, Physical Review A 45 (12) (1992) 8902.

[23] A. Carbone, G. Castelli, H. E. Stanley, Time-dependent hurst exponent in financial time series, Physica A: Statistical Mechanics and its Applications 344 (1-2) (2004) 267–271.

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

[25] C. E. Shannon, Prediction and entropy of printed english, Bell system technical journal 30 (1) (1951) 50–64.

[26] S. Noorizadeh, E. Shakerzadeh, Shannon entropy as a new measure of aromaticity, shannon aromaticity, Physical Chemistry Chemical Physics 12 (18) (2010) 4742–4749.

[27] Y. Benjamini, T. Speed, Estimation and correction for gc-content bias in high throughput sequencing, Nucleic Acids Res 40 (10) (2011) e72.

[28] D. Risso, K. Schwartz, G. Sherlock, S. Dudoit, Gc-content normalization for rna-seq data, BMC bioinformatics 12 (1) (2011) 480.

[29] N. Galtier, G. Piganeau, D. Mouchiroud, L. Duret, Gc-content evolution in mammalian genomes: the biased gene conversion hypothesis, Genetics 159 (2) (2001) 907–911.

[30] F. Hildebrand, A. Meyer, A. Eyre-Walker, Evidence of selection upon genomic gc-content in bacteria, PLoS genetics 6 (9) (2010).

 doi:10.20944/preprints202003.0344.v1

[31] S. Dutta, M. Ojha, Relatedness between major taxonomic groups of fungi based on the measurement of dna nucleotide sequence homology, Molecular and General Genetics MGG 114 (3) (1972) 232–240.

[32] T. H. Jukes, Silent nucleotide substitutions and the molecular evolutionary clock, Science 210 (4473) (1980) 973–978.

[33] M. El Naschie, On dimensions of cantor set related systems, Chaos, Solitons & Fractals 3 (6) (1993) 675–685.

[34] I. S. Baek, Dimensions of the perturbed cantor set, Real Analysis Exchange 19 (1) (1993) 269–273.

[35] 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.

[36] R. K. Rout, S. S. Hassan, S. SINDHWANI, H. M. PANDEY, S. Umer, Intelligent classification and analysis of essential genes species using quantitative methods, ACM Transactions on Mulitmedia Computing, Communications, and Applications (TOMM) (2019) TOMM–2019.

[37] 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.

[38] S. Xu, Z. Li, S. Zhang, J. Hu, Primary structure similarity analysis of proteins sequences by a new graphical representation, SAR and QSAR in Environmental Research 25 (10) (2014) 791–803.

[39] Y. ZuGuo, C. GuoYi, Rescaled range and transition matrix analysis of dna sequences, Communications in Theoretical Physics 33 (4) (2000) 673.

[40] R. W. Hamming, Error detecting and error correcting codes, The Bell system technical journal 29 (2) (1950) 147–160.

 doi:10.20944/preprints202003.0344.v1

[41] M. Norouzi, D. J. Fleet, R. R. Salakhutdinov, Hamming distance metric learning, in: Advances in neural information processing systems, 2012, pp. 1061–1069.