A Vicenary Analysis of SARS-CoV-2 Genomes

Sk Sarif Hassan¹, Ranjeet Kumar Rout², Kshira Sagar Sahoo³, Nz Jhanjhi⁴, Saiyed Umer⁵, Thamer A. Tabbakh⁶*, and Zahrah A. Almusaylim⁷

¹Department of Mathematics, Pingla Thana Mahavidyalaya, Paschim Medinipur, 721140, India
²Computer Science & Engineering, National Institute of Technology Srinagar, Hazratbal, 190006, J&K, India
³Department of Computer Science and Engineering, SRM University, Amaravati, AP, 522502, India
⁴School of Computer Science and Engineering, Taylor’s University, Subang Jaya, 47500, Malaysia
⁵Department of Computer Science and Engineering, Aliah University, Kolkata, India
⁶Materials Science Research Institute, King Abdulaziz City for Science and Technology (KACST), Riyadh, 6086, Kingdom of Saudi Arabia
⁷General Administration of Research and Development Laboratories, King Abdulaziz City for Science and Technology (KACST), Riyadh, 6086, Kingdom of Saudi Arabia

*Corresponding Author: Thamer A. Tabbakh. E-mail: ttabbakh@kacst.edu.sa
Received: 24 January 2021; Accepted: 01 May 2021

Abstract: Coronaviruses are responsible for various diseases ranging from the common cold to severe infections like the Middle East syndromes and the severe acute respiratory syndrome. However, a new coronavirus strain known as COVID-19 developed into a pandemic resulting in an ongoing global public health crisis. Therefore, there is a need to understand the genomic transformations that occur within this family of viruses in order to limit disease spread and develop new therapeutic targets. The nucleotide sequences of SARS-CoV-2 are consist of several bases. These bases can be classified into purines and pyrimidines according to their chemical composition. Purines include adenine (A) and guanine (G), while pyrimidines include cytosine (C) and tyrosine (T). There is a need to understand the spatial distribution of these bases on the nucleotide sequence to facilitate the development of antivirals (including neutralizing antibodies) and epitopes necessary for vaccine development. This study aimed to evaluate all the purine and pyrimidine associations within the SARS-CoV-2 genome sequence by measuring mathematical parameters including; Shannon entropy, Hurst exponent, and the nucleotide guanine-cytosine content. The Shannon entropy is used to identify closely associated sequences. Whereas Hurst exponent is used to identifying the auto-correlation of purine-pyrimidine bases even if their organization differs. Different frequency patterns can be used to determine the distribution of all four proteins and the density of each base. The GC-content is used to understand the stability of the DNA. The relevant genome sequences were extracted from the National Center for Biotechnology Information (NCBI) virus database. Furthermore, the phylogenetic properties of the COVID-19
virus were characterized to compare the closeness of the COVID-19 virus with other coronaviruses by evaluating the purine and pyrimidine distribution.

Keywords: Fractal dimension; shannon entropy; hurst exponent; GC-content; SARS-CoV-2

1 Introduction

The coronavirus disease pandemic (COVID-19) is an ongoing global public health crisis caused by the severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) [1,2]. The disease originally started in Wuhan, China, and quickly spread to the rest of the world, infecting millions of people worldwide [3–5]. The rapid spread of the virus has overwhelmed the most advanced healthcare systems and has so far resulted in the death of over 2.5 million people worldwide. In January 2020, the World Health Organization (WHO) affirmed COVID-19 as a public health emergency of international concern [6–9]. In order to control the disease and hence give time for the healthcare systems to cope with the sudden demand, lockdowns were ordered in many countries worldwide, leading to major economic and social disruption. In view of this, there is an urgent need to understand the genomic of the virus so as to limit the spread, reduce mortality from the disease, and develop new effective treatments.

The SARS-CoV-2 was found to be connected to two bat-derived stern acute respiratory syndrome-like coronaviruses; bat-SL-CoVZC45 and bat-SLCoVZXC21 [10]. On the 11th February 2020, the WHO formally named the disease COVID-19. From that day onwards, the coronavirus research group of the International Committee on Taxonomy of Viruses called the virus SARS-CoV-2 [11]. The National Center for Biotechnology Information (NCBI) has a complete genomic sequence of the CoV evolutionary basis and molecular uniqueness [12]. Ceraoio et al. [13] identified a good sequence relationship (above 99%) between all sequenced 2019 CoVs genomes and the bat CoV genomes, with the closest bat CoV sequence sharing 96.2% of the sequence identity. This confirmed the zoonotic origin of the virus.

Coronaviruses are enclosed RNA viruses that circulate amongst humans, other mammals, and birds, causing respiratory, enteric, hepatic, and neurologic diseases [14,15]. A total of 89 nucleotide sequences of SARS-CoV-2 are accessible from the NCBI virus database [16,17]. All these sequences consist of nearly about 29000 bases. These bases can be classified into purines and pyrimidines according to their chemical composition. Purines include adenine (A) and guanine (G), while pyrimidines include cytosine (C) and tyrosine (T). There is a need to understand the spatial distribution of these bases on the nucleotide sequence to facilitate the development of antivirals (including neutralizing antibodies) and epitopes necessary for vaccine development. Various quantitative metrics can be used to understand the spatial distribution of purines and pyrimidines, including: Hurst exponent (HE), Shannon entropy (SE), and the nucleotide guanine-cytosine content (GC-content). The HE is used to identify the auto-correlation of purine-pyrimidine bases even if their organization differs. SE is used to identify closely associated sequences. Different frequency patterns can be used to determine the distribution of all four proteins and the density of each base. The GC-content is used to understand the stability of the DNA. Therefore, this study aimed to evaluate all the purine and pyrimidine associations within the SARS-CoV-2 genome sequence by measuring mathematical parameters including; SE, HE, and the nucleotide GC-content.
The rest part of this paper is organized as follows. In Section 2, the specification of the database is explained. Definition of different fundamental parameters and their effectiveness with respect to the database have been explained in Section 3. Experimental results and illustrations are demonstrated in Section 4. Section 5 concludes the article, emphasizing the critical factors of the entire analysis.

2 Materials and Methods

2.1 Specifications of the Used Database

All the CoV nucleotide sequences were acquired from the NCBI Virus Database (http://www.ncbi.nlm.nih.gov/labs/virus/vssi/). This dataset contains 89 complete SARS-CoV-2 nucleotide sequences from the 15th March 2020. For the purpose of the study, each DNA sequence has been converted into a binary sequence of “10s” and “00s” as per Eq. (1).

Eq. (1) corresponds to purine and pyrimidine nucleotide bases encoded as 1 and 0 correspondingly into the changed binary sequence. All the 89 complete SARS-CoV-2 nucleotide sequences were labeled according to their accession ID as listed below in Tab. 1.

\[
\begin{align*}
A/G & \rightarrow 0 \\
T/C & \rightarrow 1
\end{align*}
\] (1)

| Seq | Accession ID | Seq | Accession ID | Seq | Accession ID | Seq | Accession ID |
|-----|--------------|-----|--------------|-----|--------------|-----|--------------|
| S1  | NC_045512    | S19 | MT159712     | S37 | MT050493     | S55 | MT066175     | S73 | MT019531     |
| S2  | MT188341     | S20 | MT159716     | S38 | MT152824     | S56 | MT066176     | S74 | MT007544     |
| S3  | MT188339     | S21 | MT159707     | S39 | MT135044     | S57 | MT044257     | S75 | MN996527     |
| S4  | MT188340     | S22 | MT159715     | S40 | MT135042     | S58 | MT049951     | S76 | MN996531     |
| S5  | MT184910     | S23 | MT159721     | S41 | MT135043     | S59 | MT044258     | S78 | MN996530     |
| S6  | MT184908     | S24 | MT159717     | S42 | MT135041     | S60 | MT039888     | S79 | MN996529     |
| S7  | MT184909     | S25 | MT159722     | S43 | MT126808     | S61 | MT039873     | S80 | MN988668     |
| S8  | MT184911     | S26 | MT159714     | S44 | MT123291     | S62 | MT039887     | S81 | MN997409     |
| S9  | MT184913     | S27 | MT159713     | S45 | MT123290     | S63 | MT039890     | S82 | MN994467     |
| S10 | MT184912     | S28 | MT159706     | S46 | MT123293     | S64 | MT027063     | S83 | MN988669     |
| S11 | MT184907     | S29 | MT066156     | S47 | MT123292     | S65 | MT027064     | S84 | MN994468     |
| S12 | MT163716     | S30 | MT159705     | S48 | MT118835     | S66 | MT027062     | S85 | MN988713     |
| S13 | MT163719     | S31 | MT121215     | S49 | MT106054     | S67 | MT019529     | S86 | MN975262     |
| S14 | MT163717     | S32 | MT159719     | S50 | MT06053      | S68 | MT020880     | S87 | MN938384     |
| S15 | MT163718     | S33 | MT159720     | S51 | MT106052     | S69 | MT019530     | S88 | MN985325     |
| S16 | MT159711     | S34 | MT159709     | S52 | MT093571     | S70 | MT019532     | S89 | MN908947     |
| S17 | MT159710     | S35 | MT159718     | S53 | MT093631     | S71 | MT019533     |
| S18 | MT159708     | S36 | MT012098     | S54 | MT072688     | S72 | MT020881     |
The length of these complete 89 sequences varied between 29783 to 29981 nucleotides, and the range was 198 bp long. The smallest complete SARS-CoV sequence was S2 with a length of 29783, and the largest one was S47, with a length of 29981. Two sequences had a length of 29867, Thirty-nine sequences had a length of 29882, and 11 sequences were 29903 long.

2.2 Generation of Gene Clusters

Different quantitative parameters, including; SE, fractal dimension (FD), HE, and the distribution of purines-pyrimidines contents, were used to describe the spatial distribution of the bases of the SARS-CoV-2 sequences.

2.3 FD of the Indicator Matrices

FD is a key for characterizing fractal patterns or sets whereby \( D = \{0, 1\} \) is a set of two symbols characterizing the purine and pyrimidine bases of a nucleotide sequence, and \( S(l) \) is the binary sequence corresponding to a nucleotide sequence with the repetition of two characters from \( D \) to length \( l \). All the binary sequences in our study were transferred into the indicator matrices [18–21]. The patterns were demonstrating self-similarity in the fractal dimension point to which the fractal objects filled a particular Euclidean space in which it was entrenched. Several methods have been described in the literature to determine the self-organizing configuration of the DNA sequences throughout an indicator matrix. The indicator function for each sequence was then defined as shown in Eq. (2) [22]:

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

such that the indicator matrix:

\[
\vartheta_{hk} = \vartheta (x, y) = \begin{cases} 
1, & \text{if } x = y \\
0, & \text{if } x \neq y 
\end{cases}
\]

whereby \( \vartheta_{hk} \) is a matrix with the values 0 and 1. A binary image is generated through this matrix to understand the correlation between the sequences. Similarly, we can also depict the auto-correlation between the purine and pyrimidine for the same sequence. The image will be generated by assigning a ‘1’ to black dots and ’0’ to white dots. The purine and pyrimidines have been distributed like a fractal. The indicator matrix FD has been calculated as the average number of \( \sigma(p) \) of 1, acquired from the \( P \times P \) indicator matrix with \( p \times p \) randomly. Using \( \sigma(p) \), the FD is defined in Eq. (3):

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

The self-organization of purine and pyrimidine bases for all the SARS-CoV-2 sequences can be obtained through the indicator matrix FD. The box-counting method is the most commonly used to determine the FD.

2.4 HE

The autocorrelation of purine-pyrimidine bases for all the SARS-CoV-2 sequences was obtained through the HE. The HE was applied during the time series investigation to infer the autocorrelation [23,24]. The HE values range between 0 to 1. An HE value of 0.5 indicates the absolute randomness of the time series data, while a value below 0.5 indicates a negative
correlation, and a value above 0.5 indicates a positive correlation. The HE of a binary sequence $s_n$ is defined as below.

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

where

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

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

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

and

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

### 2.5 SE

The SE was used to measure the uncertainty of the binary sequence. Primary protein sequences were generated through different combinations of amino acids ranging from 30 to 3000. Some protein sequences were kept as a substring like AAAAAAAG and AAAAAAAATTTTTTTT, which resulted from coding of one or an assortment of amino acids. Such proteins are less likely to encode functional proteins. Therefore, the amount of information or the sequence uncertainty concerning a base pair was measured using the SE. The SE was used to measure the Bernoulli process entropy with the probability ($p$) of the two outcomes (0/1) defined as below;

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

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].

If the probability $p = 0$, the event will never occur; otherwise, if $p = 1$, a certain result will be generated with entropy 0. When $p = 0.5$, the uncertainty is at a maximum, and consequently, the SE is 1.

### 2.6 GC Content and Nucleotides Density

In molecular biology, the GC-content is usually calculated as a percentage and is sometimes called $G + C$ ratio or GC-ratio [25–28]. The percentage of GC-content and GC-ratio of the DNA sequences $s$ used to measure several resources. One of the simplest procedures is to measure the melting point of the DNA sequences using spectrophotometry. A higher GC-content indicates a more stable DNA structure.
The GC-content percentage was calculated by the formula \( \frac{\text{Count}(G+C)}{\text{Count}(A+T+G+C)} \times 100\% \) [29,30].

In addition to the GC-content, the density of the nucleotides \( A, T, C, \) and \( G \) were acquired separately in the present study [31,32].

3 Results and Illustrations

The frequencies of several nucleotides in the SARS-CoV-2 sequences were not selected randomly. In this study, we, therefore, tried to evaluate the purine and pyrimidine spatial distribution organizations among the SARS-CoV-2 sequences through the parameters as defined in the previous section. In addition to the investigation of the purine-pyrimidine distribution, we also explored the density of each of the nucleotides and GC—content, which has a significant role on the stability of the sequence.

3.1 Classification based on the FD of the Indicator Matrices

Three distinct FDs (0.3, 0.4755, and 0.6) were identified, indicating that only three clusters within the sequences are turned up. Tab. 2 demonstrates the sequences and their corresponding FD. The histograms of all the SARS-CoV-2 sequences that were plotted according to the FD are illustrated in Fig. 1.

The dimension of each indicator matrix was above 29000 \( \times \) 29000, and therefore it was not possible to generate an image of the indicator matrix. The sequences S47, S13, S28, and S79, had an FD of 0.3, which depicts that the amount of fractality (a kind of non-linearity) is small, indicating that the purine and pyrimidine organization is relatively well-organized and closely resembling the affine type. There were eight sequences, S48, S49, S50, S51, S53, S54, S55, and S56, having an FD 0.4755, and the rest of the purine and pyrimidine sequences had an FD of 0.6, indicating a significant closeness to the FD of the cantor set [33,34].

### Table 2: Sequences and their corresponding FDs

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

3.2 Classification Based on the HE

For each of the binary SARS-CoV-2 sequences, the HE was calculated using Eq. (4), and then ten clusters were formed using the k-means clustering technique for all the sequences. The
histograms of all the SARS-CoV-2 sequences that were plotted according to the HE are illustrated in Fig. 2.

Figure 1: Plot of the fractal dimension (FD) and corresponding histogram of all the purine-pyrimidine binary sequences corresponding to the SARS-CoV-2 sequences

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

The HE was confined to the interval 0.643, 0.655 of length 0.0123. This suggests a positive autocorrelation in the spatial distribution of the purine and pyrimidine bases for all the SARS-CoV-2 sequences. The sequence S1 (accession ID: NC04551) had the highest HE (0.712) as shown in Tab. 3. Furthermore the sequence also had a significantly different spatial organization of purine and pyrimidine bases. The length of the sequence S1 was 29903. Although there were ten other sequences (S1, S13, S14, S15, S39, S40, S41, S42, S57, S60, and S89) with the same length as S1, their HE value differed significantly from S1.
Table 3: Hurst exponent of all the 89 purine-pyrimidine binary sequences corresponding to SARS-CoV-2 sequences

| Seq. | HE   | Seq. | HE   | Seq. | HE   | Seq. | HE   | Seq. | HE   | Seq. | HE   |
|------|------|------|------|------|------|------|------|------|------|------|------|
| S47  | 0.6430890415 | S11  | 0.6463681216 | S44  | 0.6456954376 | S50  | 0.6463681216 | S40  | 0.6491763266 |
| S79  | 0.6438083222 | S16  | 0.6463681216 | S70  | 0.6466330524 | S63  | 0.6457628936 | S51  | 0.6463681216 | S41  | 0.6491763266 |
| S75  | 0.6439455807 | S17  | 0.6463681216 | S80  | 0.6466330524 | S2   | 0.6459898777 | S64  | 0.6463681216 | S60  | 0.6491763266 |
| S76  | 0.644253158  | S18  | 0.6463681216 | S58  | 0.6466886489 | S5   | 0.6459943594 | S65  | 0.6463681216 | S74  | 0.6491856177 |
| S28  | 0.6443251468 | S19  | 0.6463681216 | S59  | 0.6467842124 | S53  | 0.6459898777 | S66  | 0.6463681216 | S13  | 0.6494006145 |
| S37  | 0.6444723241 | S20  | 0.6463681216 | S58  | 0.6466886489 | S21  | 0.6460659477 | S72  | 0.6463681216 | S14  | 0.6494692533 |
| S77  | 0.644522348  | S24  | 0.6463681216 | S45  | 0.6470843102 | S33  | 0.6460659477 | S62  | 0.6463681216 | S57  | 0.6497613576 |
| S55  | 0.6445500767 | S25  | 0.6463681216 | S68  | 0.6475613391 | S38  | 0.6460690307 | S5   | 0.6459943594 | S30  | 0.6553858343 |
| S56  | 0.6445721887 | S26  | 0.6463681216 | S78  | 0.6477090451 | S87  | 0.6464119458 | S12  | 0.6497671065 |  |  |
| S86  | 0.6447294561 | S29  | 0.6463681216 | S71  | 0.6483049547 | S8   | 0.6464832466 |  |  |
| S44  | 0.644748006  | S31  | 0.6463681216 | S67  | 0.6487736204 | S42  | 0.648825729  |  |  |
| S46  | 0.644776066  | S32  | 0.6463681216 | S47  | 0.648825729  | S71  | 0.6483049547 |  |  |
| S2   | 0.644839682  | S33  | 0.6463681216 | S69  | 0.6488862311 | S88  | 0.6482531468 |  |  |
| S83  | 0.644839682  | S34  | 0.6463681216 | S76  | 0.6487736204 | S67  | 0.6483049547 |  |  |
| S61  | 0.6448639682 | S35  | 0.6463681216 | S84  | 0.648825729  | S76  | 0.6487736204 |  |  |
| S68  | 0.6449176326 | S36  | 0.6463681216 | S81  | 0.648825729  | S76  | 0.6487736204 |  |  |

Based on the HE obtained from the binary SARS-CoV-2 sequences, ten clusters were formed by using the k-means clustering. Cluster-1 contained 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, S78, S10, S39, S40, S41, S60, S74, S13, S14, S89, S57, S12, S3) all having their center at 0.6464. Cluster-2 contained 11 sequences (S39, S40, S41, S60, S74, S13, S14, S89, S57, S12, S3) with their center at 0.6494. Cluster-3 contained only one sequence (S1) centered at 0.7125. Similarly, cluster-4 contained only one sequence (S47), centered at 0.6431. Cluster-5 contained 11 sequences (S37, S77, S55, S56, S57, S61, S54, S46, S52, S6, S36), all having their centers at 0.6448. The sequence S30 was located in cluster S6, with its center at 0.6554. The sequences (S68, S78, S88, S71, S67, S42, S15, S69) are in cluster-7, whose center is at 0.6483. The cluster-8 contained sequences (S43, S63, S2, S5, S21, S22, S33, S35, S38, S84) whose center was at 0.6460. The sequence S4 belonged to cluster-9, centered at 0.6516. Cluster 10 contained four sequences, S79, S75, S76, S28, all with their center at 0.6441. The sequences S55 and S66 had the same HE of 0.6445500767. This confirmed their identical long-range correlation even though the length of these two sequences differed by 2 bp (S55: 29870 and S66: 29872). Furthermore, the sequences S21, S22, S33, S35 belonging to the cluster-8 also had the same HE of 0.646059477. The cluster-1 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 S87 belonging had an HE of 0.6463681216. Three sequences S9, S10, S83, in cluster-1 were found to have the same HE of 0.6464832466. Four sequences in cluster-2 S39, S40, S41 and S60, also had the same HE of 0.6491763266.

3.3 Classification Based on SE

For all the 89-binary purine-pyrimidine sequences of the SARS-CoV-2, the SE was first determined, and then ten different clusters were formed based on the SE obtained for all the sequences, as shown in Tab. 4. The SE and the histograms of all the SARS-CoV-2 sequences are illustrated in Fig. 3.

An SE ranging from 0.9999 to 1 indicates that the length of the range is too small, and therefore the SE is precisely the same for all the sequences. The SE for all sequences was 0.9999.
except for sequence S30, which was 29945. This indicates the maximum level of uncertainty for the S30 sequence with a probability of a purine-pyrimidine occurrence of 0.5. This means that although this sequence was not randomly composed of nucleotide bases as they are positively autocorrelated with an HE 0.6553, the purine and pyrimidine bases are composed with equal probability.

Table 4: Shannon entropy (SE) of all the purine-pyrimidine binaries sequences corresponding to SARS-CoV-2

| Seq  | SE      | Seq  | SE      | Seq  | SE      | Seq  | SE      | Seq  | SE      | Seq  | SE      |
|------|---------|------|---------|------|---------|------|---------|------|---------|------|---------|
| S47  | 0.9999223787 | S36  | 0.999928592  | S87  | 0.9999320596 | S88  | 0.9999362496 | S22  | 0.9999362496 | S89  | 0.9999416252 |
| S28  | 0.999927878  | S63  | 0.9999301724 | S62  | 0.9999325138 | S71  | 0.9999380434 | S33  | 0.9999311193 | S12  | 0.9999424669 |
| S79  | 0.9999235693 | S84  | 0.9999301724 | S44  | 0.999932568  | S13  | 0.999938992  | S35  | 0.9999311193 | S57  | 0.99994249 |
| S75  | 0.9999242037 | S2   | 0.9999301848 | S58  | 0.9999328857 | S69  | 0.9999398601 | S83  | 0.9999311193 | S38  | 0.9999311008 |
| S77  | 0.9999247235 | S64  | 0.9999320596 | S45  | 0.9999329935 | S15  | 0.9999398762 | S65  | 0.9999320596 | S9   | 0.9999311193 |
| S37  | 0.999925013  | S7   | 0.9999320596 | S23  | 0.999932568  | S14  | 0.9999398762 | S66  | 0.9999320596 | S10  | 0.9999311193 |
| S76  | 0.999925013  | S11  | 0.9999320596 | S53  | 0.9999348593 | S74  | 0.9999407143 | S72  | 0.9999320596 | S21  | 0.9999311193 |
| S46  | 0.9999260645 | S16  | 0.9999320596 | S68  | 0.9999357908 | S67  | 0.9999407381 | S73  | 0.9999320596 | S50  | 0.9999320596 |
| S86  | 0.9999261031 | S17  | 0.9999320596 | S78  | 0.9999362496 | S39  | 0.9999407381 | S70  | 0.9999320596 | S51  | 0.9999320596 |
| S55  | 0.9999262613 | S18  | 0.9999320596 | S25  | 0.9999320596 | S40  | 0.9999407539 | S39  | 0.9999407381 | S89  | 0.9999311193 |
| S56  | 0.9999262613 | S19  | 0.9999320596 | S26  | 0.9999320596 | S41  | 0.9999407539 | S30  | 0.9999407539 | S6   | 0.9999272835 |
| S54  | 0.9999264586 | S20  | 0.9999320596 | S27  | 0.9999320596 | S60  | 0.9999407539 | S1   | 0.9999416252 | S82  | 0.9999315857 |
| S61  | 0.999926567  | S24  | 0.9999320596 | S29  | 0.9999320596 | S59  | 0.9999320596 | S14  | 0.9999416252 | S34  | 0.9999320596 |
| S52  | 0.999927186  | S85  | 0.9999311193 | S31  | 0.9999320596 | S5   | 0.9999282594 | S70  | 0.9999315948 | S34  | 0.9999320596 |
| S43  | 0.999927264  | S81  | 0.9999315857 | S32  | 0.9999320596 | S8   | 0.9999282594 | S80  | 0.9999315948 | S82  | 0.9999315857 |

Figure 3: Plot of the Shannon entropy (SE) and matching histogram of all the purine-pyrimidine binary sequences corresponding with the SARS-CoV-2 sequences

After evaluating all the SE of the binary purine and pyrimidine represented for the SARS-CoV sequences, only three clusters were formed using the k-means clustering technique. The cluster-1 contained 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 centered at 0.999940381147619. The other 67
sequences belonged to cluster-2 and were all centered at 0.999930184068656. Therefore, these two clusters can be considered the same. Cluster-3 contained only one sequence (S30) with an SE of 0.9999585474 (approximately 1), as already mentioned before.

The distribution of SE for all the purine and pyrimidine distributions among the SARS-CoV-2 sequences was mostly linear. This is crucial for the SARS-CoV-2, unlike other sequences obtained in previous studies made [35–37]. The uncertainty level reached the maximum, which means that the probability of purine and pyrimidine bases occurring across the sequences among all the SARS-CoV-2 is equal.

### 3.4 GC, A, T, C, and G Density in the SARS-CoV-2

The sequences were classified according to the GC, A, T, and G densities, as follows. Tab. 5 shows the percentage density of the GC-content among all the SARS-CoV-2 sequences. The histograms of all the SARS-CoV-2 sequences that were plotted according to the GC content as shown in Fig. 4. The percentage density of the GC-content was around 37.5% meaning that the SARS-CoV-2 sequences is A and T rich. This means that A (30) was the most common purine base nucleotides, and T (32) was the most common pyrimidine base nucleotide(T). The occurrence of purine and pyrimidine bases was equally probable based on their SE. This is an important specialty of the SARS-CoV-2 sequences.

**Table 5: Sequences and their respective percentage GC Content and their corresponding clusters**

| Seq | % of GC | C | Seq | % of GC | C | Seq | % of GC | C | Seq | % of GC | C |
|----|--------|---|----|--------|---|----|--------|---|----|--------|---|
| S30 | 37.912840207 | 10 | S24 | 37.9927715682 | 4 | S22 | 37.9994645606 | 5 | S89 | 37.972778651 | 6 |
| S13 | 37.9460254824 | 1 | S33 | 37.9927715682 | 4 | S25 | 37.9994645606 | 5 | S3 | 37.9795610655 | 3 |
| S60 | 37.9560579206 | 9 | S35 | 37.9927715682 | 4 | S26 | 37.9994645606 | 5 | S69 | 37.9812033847 | 3 |
| S8 | 37.9659995984 | 6 | S49 | 37.9927715682 | 4 | S31 | 37.9994645606 | 5 | S5 | 37.9860785757 | 7 |
| S74 | 37.968755227 | 6 | S65 | 37.9927715682 | 4 | S51 | 37.9994645606 | 5 | S7 | 37.9991500184 | 7 |
| S57 | 37.9694345049 | 6 | S83 | 37.9927715682 | 4 | S62 | 37.9993303634 | 5 | S6 | 37.9953145917 | 4 |
| S12 | 37.9703649196 | 6 | S89 | 37.9927715682 | 4 | S65 | 37.9927715682 | 4 | S43 | 38.0004016602 | 5 |
| S1 | 37.972778651 | 6 | S23 | 37.9951116617 | 4 | S81 | 38.0007362538 | 5 | S53 | 38.0010707355 | 5 |
| S14 | 37.972778651 | 6 | S6 | 37.9953145917 | 4 | S82 | 38.0007362538 | 5 | S78 | 37.9860785757 | 7 |
| S15 | 37.972778651 | 6 | S10 | 37.9961806444 | 4 | S53 | 38.0010707355 | 5 | S78 | 37.9880231508 | 7 |
| S39 | 37.972778651 | 6 | S19 | 37.9961806444 | 4 | S50 | 38.0028110568 | 5 | S8 | 37.9892450719 | 7 |
| S40 | 37.972778651 | 6 | S21 | 37.9961806444 | 4 | S58 | 38.0032152187 | 5 | S86 | 37.9892450719 | 7 |
| S41 | 37.972778651 | 6 | S27 | 37.9961806444 | 4 | S28 | 38.0051561925 | 5 | S20 | 37.9892450719 | 7 |
| S42 | 37.972778651 | 6 | S32 | 37.9961806444 | 4 | S28 | 38.0051561925 | 5 | S20 | 37.9892450719 | 7 |
| S48 | 37.972778651 | 6 | S36 | 38.012168969 | 2 | S34 | 37.9894250719 | 7 | S9 | 38.0150880134 | 2 |
| S59 | 37.972778651 | 6 | S46 | 38.0132918139 | 2 | S38 | 37.9978579557 | 5 | S86 | 38.0152825526 | 2 |
| S64 | 37.972778651 | 6 | S47 | 38.0140755812 | 2 | S44 | 37.9980596166 | 5 | S52 | 38.0174146015 | 2 |
| S66 | 37.972778651 | 6 | S55 | 38.0147304988 | 2 | S7 | 37.9994645606 | 5 | S80 | 37.9915001841 | 7 |

Based on the GC-content density in the SARS-CoV-2 sequences, ten different clusters are formed using the k-means clustering technique as shown in Tab. 5. These ten clusters (C) had their centers 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 was located in the 37.91284, 38.02505 interval. Cluster-10 and cluster-9 contained only one sequence. The GC-content of the S30, S13, and S60 were 37.91284%, 37.94602%, and 37.95605%, respectively. As also previously explained, S30 also had an SE of 1.
The A, T, C, and G intervals and their corresponding densities are summarized in Tab. 6. The histograms of all the SARS-CoV-2 sequences that were plotted according to the density of A, T, G and C are illustrated in Figs. 5–8 respectively. The spread of A, T, C, and G over the SARS-CoV-2 sequences were approximately 30%, 32%, 18%, and 19%, respectively. These findings further confirm that SAR-CoV2 is significantly \(AT\) rich and the density of the purine and pyrimidine bases is similar as shown by the SE.

| Seq | % A | % T | % C | % G | 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 |

(Continued)
Table 6: Continued

| Seq | % A   | % T   | % C   | % G   | Seq | % A   | % T   | % C   | % G   | Seq | % A   | % T   | % C   | % G   |
|-----|-------|-------|-------|-------|-----|-------|-------|-------|-------|-----|-------|-------|-------|-------|
| 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 |     |     |     |     |     |

Figure 5: Plot of the A content density and its corresponding histogram of the SARS-CoV-2 sequences

All the sequences of SARS-CoV-2 sequences are clustered into different clusters. The position of each cluster center for all four bases differed by 0.01. The sequence S79 has the least percentage (29.85%) of the nucleotide base A, whereas sequence S30 has the lowest 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, followed by G (19.65%) of C (18.40%).
The similarity analysis of the SARS-CoV-2 sequences was measured by calculating the distance between the binary vectors of the binary strings encoded based on purines and pyrimidines nucleotide bases, as mentioned earlier. Several computing methods measure the distance between multidimensional vectors, such as Hamming distance (HD), Euclidean distance, Elastic-matching distance, Jeffrey and Matusita distance, Manhattan distance, and Minkowski norm. Reportedly, these methods have little effect on the vector similarity [38]. The HD between two binary strings is defined by the number of bits in which they vary [39,40]. However, here we had to take into consideration that the length of the different SARS-CoV-2 genome usually varies by some bases.
Suppose there are two \( SARS-CoV-2 S_{1}^{x} \) and \( S_{2}^{y} \) with a length of \( x \) and \( y \) respectively (\( x > y \)), then
\[
HD(S_{1}^{x}, S_{2}^{y}) = hd(S_{1}^{y}, S_{2}^{y})
\]
if the two binary sequences \( S_{x} = 101011 \) and \( S_{m} = 0010 \), have a minimum length of 4, from left to right the HDs are \((101011,0010) = 1\). The two binary sequences, \( x \), and \( y \) are identical if the HD= 0, which indicates a similar distribution of purines and pyrimidines over the \( SARS-CoV-2 \) sequences. Similarly, the distribution of purines and pyrimidines over the \( SARS-CoV-2 \) sequences are completely different when the HD = \( \min(x, y) \). To measure the distance of the \( SARS-CoV-2 \) based on their purine-pyrimidine distribution, minimum HD was used. The larger the HD between the sequences the lower the probability that these two sequences are related to each other.

**Figure 8:** Plot of the \( G \) content density and its corresponding histogram of the SARS-CoV-2 sequences

The \( SARS-COV-2 \) virus sequences MT044258(S59), MN994468 (S84), NC_045512(S1), and MN039888(S60) were grouped together as a single cluster as the distance between them was almost negligible, indicating that they are closely related. Furthermore, the sequences MT152824(S38), MN996531(S76), MT012098(S36), and MT975262(S86) were closely related to each other and therefore treated as a single 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) also had similar HD and were therefore grouped together. After taking into consideration the HD between these sequences, it was observed that they were very closely related. This closeness (nearness) among the \( SARS-CoV-2 \) genomes makes it possible for future such genomes or other blasted results to analyze clusters quantitatively instead of only relying on sequential similarity.

**4 Conclusions and Summary**

The novel coronavirus has led to a worldwide public health emergency. One of the major reasons for such a global threat is the lack of quantitative and qualitative knowledge about this novel virus, including its genomic and proteomic levels. In this article, we evaluated the quantitative nature of the \( SARS-COV-2 \) complete sequences. This present study revealed the
closeness amongst the 89 complete sequences in the purine-pyrimidine level descriptions through phylogenetic analysis. Based on this quantitative investigation, very interesting observations were made. The purine and pyrimidine were found to be evenly and equally spaced throughout all 89 SARS-CoV sequences. The GC− content was also significantly low. This quantitative data helps us to better understand the genomic sequences of the SARS-CoV-2 sequences and could potentially be used to reduce disease spread and to identify new therapeutic targets. However, these observations could be further strengthened by evaluating the SARS-CoV-2 proteins.

Acknowledgement: We appreciate the linguistic assistance provided by TopEdit (www.topeditsci.com) during the preparation of this manuscript.

Funding Statement: We are thankful to King Addulaziz City for Science and Technology (KACST) Saudi Arabia for providing support. We are thankful to the Center of Smart Society 5.0 [CSS5] for the support to complete this research.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References
[1] K. V. Holmes, “SARS-associated coronavirus,” New England Journal of Medicine, vol. 348, no. 20, pp. 1948–1951, 2003.
[2] L. Van Der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. M. Berkhout et al., “Identification of a new human coronavirus,” Nature Medicine, vol. 10, no. 4, pp. 368–373, 2004.
[3] M. Lipsitch, D. L. Swerdlow and L. Finelli, “Defining the epidemiology of Covid-19—Studies needed,” New England Journal of Medicine, vol. 382, no. 13, pp. 1194–1196, 2020.
[4] A. S. Fauci, H. C. Lane and R. R. Redfield, Covid-19—Navigating the uncharted. Waltham, Massachusetts, United States: Mass Medical Soc., 2020.
[5] W. Liu, Q. Zhang, J. Chen, R. Xiang and H. Song, “Detection of Covid-19 in children in early January 2020 in Wuhan, China,” New England Journal of Medicine, vol. 382, no. 14, pp. 1370–1371, 2020.
[6] F. Jiang, L. Deng, L. Zhang, Y. Cai, C. W. Cheung et al., “Review of the clinical characteristics of coronavirus disease 2019 (covid-19),” Journal of General Internal Medicine, vol. 35, no. 5, pp. 1–5, 2020.
[7] J. Stebbing, A. Phelan, I. Griffin, C. Tucker, O. Oechsle et al., “Covid-19: Combining antiviral and anti-inflammatory treatments,” The Lancet Infectious Diseases, vol. 20, no. 4, pp. 400–402, 2020.
[8] J. F. W. Chan, C. C. Y. Yip, K. K. W. To, T. H. C. Tang, S. Y. C. Wong 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, vol. 58, no. 5, pp. e00310-20, 2020.
[9] C. Sohrabi, Z. Alsafi, N. O’Neill, M. Khan, A. Kerwan et al., “World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19),” International Journal of Surgery, vol. 76, pp. 71–76, 2020.
[10] M. A. Marra, S. J. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson et al., “The genome sequence of the sars-associated coronavirus,” Science, American Association for the Advancement of Science, vol. 300, pp. 1399–1404, 2003.
[11] P. Sun, X. Lu, C. Xu, W. Sun and B. Pan, “Understanding of COVID-19 based on current evidence,” Journal of Medical Virology, vol. 92, no. 6, pp. 548–551, 2020.
[12] S. Zhang, M. Y. Diao, L. Duan, Z. Lin and D. Chen, “The novel coronavirus (Sars-cov-2) infections in china: Prevention, control and challenges,” Intensive Care Medicine, vol. 46, no. 4, pp. 1–3, 2020.
[13] C. Ceraolo and F. M. Giorgi, “Genomic variance of the 2019-ncov coronavirus,” Journal of Medical Virology, vol. 92, no. 5, pp. 522–528, 2020.
[14] G. Kampf, D. Todt, S. Pfaender and E. Steinmann, “Persistence of coronaviruses on inanimate surfaces and its inactivation with biocidal agents,” Journal of Hospital Infection, vol. 104, no. 3, pp. 246–251, 2020.

[15] S. Khan, A. Ali, R. Siddique and G. Nabi, “Novel coronavirus is putting the whole world on alert,” Journal of Hospital Infection, vol. 104, no. 3, pp. 252–253, 2020.

[16] J. Xu, S. Zhao, T. Teng, A. E. Abdalla, W. Zhu et al., “Systematic comparison of two animal-to-human transmitted human coronaviruses: Sars-cov-2 andsars-cov,” Viruses, vol. 12, no. 2, pp. 244, 2020.

[17] W. B. Yu, G. D. Tang, L. Zhang and R. T. Corlett, “Decoding the evolution and transmissions of the novel pneumonia coronavirus (Sars-cov-2) using whole genomic data,” Zoological Research, vol. 41, no. 3, pp. 247–257, 2020.

[18] C. Cattani and G. Pierro, “On the fractal geometry of DNA by the binary image analysis,” Bulletin of Mathematical Biology, vol. 75, no. 9, pp. 1544–1570, 2013.

[19] C. Cattani, Fractals and hidden symmetries in DNA. In: Mathematical Problems in Engineering. London, United Kingdom: Hindawi, 2010.

[20] S. S. Hassan, P. P. Choudhury, B. DayaSagar, S. Chakraborty, R. Guha et al., “Quantitative description of genomic evolution of olfactory receptors,” Asian-European Journal of Mathematics, vol. 8, no. 3, pp. 1550043, 2015.

[21] R. K. Rout, P. Pal Choudhury, S. P. Maity, B. DayaSagar and S. S. Hassan, “Fractal and mathematical morphology in intricate comparison between tertiary protein structures,” Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, vol. 6, no. 2, pp. 192–203, 2018.

[22] C. L. Berthelsen, J. A. Glazier and M. H. Skolnick, “Global fractal dimension of human dnasequences treated as pseudorandom walks,” Physical Review A, vol. 45, no. 12, pp. 8902–8913, 1992.

[23] A. Carbone, G. Castelli and H. E. Stanley, “Time-dependent Hurst exponent in financial time series,” Physica A: Statistical Mechanics and its Applications, vol. 344, no. 1–2, pp. 267–271, 2004.

[24] J. Miściniczuk and P. Wójcik, “Estimation of Hurst exponent revisited,” Computational Statistics & Data Analysis, vol. 51, no. 9, pp. 4510–4525, 2007.

[25] C. E. Shannon, “Prediction and entropy of printed English,” Bell System Technical Journal, vol. 30, no. 1, pp. 50–64, 1951.

[26] S. Noorizadeh and E. Shakerzadeh, “Shannon entropy as a new measure of aromaticity, shannonaromaticity,” Physical Chemistry Chemical Physics, vol. 12, no. 18, pp. 4742–4749, 2010.

[27] Y. Benjamini and T. Speed, “Estimation and correction for GC-content bias in high throughput sequencing,” Nucleic Acids Research, vol. 40, no. 10, pp. e72, 2011.

[28] D. Risso, K. Schwartz, G. Sherlock and S. Dudoit, “GC-content normalization for RNA-seqdata,” BMC Bioinformatics, vol. 12, no. 1, pp. 480, 2011.

[29] N. Galtier, G. Piganeau, D. Mouchiroud and L. Duret, “GC-content evolution in mammalian genomes: The biased gene conversion hypothesis,” Genetics, vol. 159, no. 2, pp. 907–911, 2001.

[30] F. Hildebrand, A. Meyer and A. Eyre-Walker, “Evidence of selection upon genomic GC-content in bacteria,” PLoS Genetics, vol. 6, no. 9, pp. e1001107, 2010.

[31] S. Dutta and M. Ojha, “Relatedness between major taxonomic groups of fungi based on the measurement of DNA nucleotide sequence homology,” Molecular and General Genetics MGG, vol. 114, no. 3, pp. 232–240, 1972.

[32] T. H. Jukes, “Silent nucleotide substitutions and the molecular evolutionary clock,” Science, vol. 210, no. 4473, pp. 973–978, 1980.

[33] M. El Naschie, “On dimensions of cantor set related systems,” Chaos, Solitons & Fractals, vol. 3, no. 6, pp. 675–685, 1993.

[34] I. S. Baek, “Dimensions of the perturbed cantor set,” Real Analysis Exchange, vol. 19, no. 1, pp. 269–273, 1993.

[35] J. K. Das, P. P. Choudhury, A. Chaudhuri, S. S. Hassan and P. Basu, “Analysis of purines andpyrimidines distribution over mirnas of human, gorilla, chimpanzee, mouse and rat,” Scientific Reports, vol. 8, no. 1, pp. 1–19, 2018.
[36] R. K. Rout, S. S. Hassan, S. Sindhwani, H. M. Pandey and S. Umer, “Intelligent classification and analysis of essential genes species using quantitative methods,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 16, pp. 38:1–38:21, 2019.

[37] J. P. Banerjee, J. K. Das, P. P. Choudhury, S. Mukherjee, S. S. Hassan et al., “The variations of human miRNAs and ising like base pairing models,” BioRxiv, pp. 319301, 2018.

[38] S. Xu, Z. Li, S. Zhang and J. Hu, “Primary structure similarity analysis of proteins sequences by a new graphical representation,” SAR and QSAR in Environmental Research, vol. 25, no. 10, pp. 791–803, 2014.

[39] Y. ZuGuo and C. GuoYi, “Rescaled range and transition matrix analysis of DNA sequences,” Communications in Theoretical Physics, vol. 33, no. 4, pp. 673–678, 2000.

[40] R. W. Hamming, “Error detecting and error-correcting codes,” The Bell System Technical Journal, vol. 29, no. 2, pp. 147–160, 1950.