Quantitative Analysis of Genomic Sequences of Virus RNAs Using a Metric-Based Algorithm

Alex Belinsky[1] and Guennadi A. Kouzaev[2]

Abstract.The worldwide spread of SARS-CoV-2 virus increases interest in the research of virus genomics and the creation of more advanced study methods. This work aims to develop a new fast DNA walk algorithm for one-dimensional visualization of RNAs based on a big-data method and comparative examination of several viruses and their lines and strains. In this work, a new metric-based algorithm for quantitative and visual analyses of RNAs is proposed and considered. It allows finding any fragments of genomic sequences using the Hamming distance between the binary-expressed RNA characters and symbols of a fragment under the search and building one-dimensional trajectories of genomic walks convenient for quantitative and qualitative analyses of RNAs and DNAs. Similarly, human-language texts can be processed and compared with genomic sequences.This algorithm is used to investigate the complete genomic sequences of SARS CoV-2, MERS, Dengue, and Ebola viruses available from Genbank ® and GISAID databases. The distributions of *atg* codon-starting triplets along with these sequences are built and considered as their *atg*-schemes. Additionally to the *atg*-walks, single-symbols distributions are calculated to detect the codon-content mutations, which do not change the *atg*-triplet coordinates along with genomic sequences. The visual analyses of distributions consisting of several hundred triplets enable us to define the level of stability of RNAs towards essential mutations and perform their classing. Statistical studies are applied to distributions of the inter-*atg* and inter-symbol distances along with genomic sequences. The fractal dimension values of these distributions are calculated, enabling them to correspond to the mutations discovered by Hamming walks and fractal-dimension values of several ten virus samples investigated here. The developed metric-based-based algorithm allows building one-dimensional RNA schemes of different scale levels and effectively analyzing the virus mutations with their classing.

Keywords: RNA sequences, Hamming-distance metric measure, quantitative RNA, DNA walks, *atg*-walks, fractality, SARS Cov-2 virus, MERS virus, Dengue virus, Ebola virus

[1] MachH-3dP Inc., Bulrlington, Canada
[2] Department of Electronic Systems, Norwegian University of Science and Technology-NTNU, Trondheim, Norway. Corresponding author: guennadi.kouzaev@ntnu.no

1. Introduction

A virus is a tiny semi-life specimen carrying a genetic material (RNA or DNA - double-helix RNA structure) in a protein capsid and covered by a lipid coat. It penetrates the cell wall and urges this cell to 'manufacture' more viruses. Fortunately, only a few viruses are harmful, and one of them SARS CoV-2, has taken millions of lives, and it threatens all humankind. Although several vaccines have been announced to prevent the disease, the research should further understand the nature of this virus to develop new vaccines and drugs against current and future virus mutations. In this paper, some unique properties of virus linear RNAs have been discovered theoretically by analyzing complete genomic sequences registered in open-access genomic databases.

As it is known, some viruses belong to the RNA-based specimens, which transfer the genetic information by long chains of four organic acids, namely, Adenine (*a*), Cytosine (*c*), Guanine (*g*), and Uracil (*u*) [1]. In DNA-based viruses and double-stranded genetic polymers, Uracil (*u*) is substituted by Thymine (*t*).

In databases, RNA is often represented by a sequence of symbols '*a*', '*c*', '*g*', '*t*', where '*t*' substitutes '*u*' (Fig. 1) due to some gene sequencing peculiarities. It involves forming complementary DNA (cDNA) for stable measurements and biological materials storage [2]. However, direct RNA sequencing techniques are on the way [3],[4]. For many theoretical studies, the '*t*-modified' text is pertinent; otherwise, it should be transformed by substituting '*u*' instead of '*t*'.

ttaaaggtttataccttcccaggtaacaaaccaaccaactttcgatctcttgtagatctgttctctaaacgaactttaaa
atctgtgtggctgtcactcggctgcatgcttagtgcactcacgcagtataattaataactaattactgtcgttgacagg
acacgagtaactcgtctatcttctgcaggctgcttacggtttcgtccgtgttgcagccgatcatcagcacatctaggttt
cgtccgggtgtgaccgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaactcagtt
tgcctgttttacaggttcgcgacgtgctcgtacgtggctttggagactccgtggaggaggtcttatcagaggcacgtca
acatcttaaagatggcacttgtggcttagtagaagttgaaaaaggcgttttgcctcaacttgaacagccctatgtgttc
atcaaacgttcggatgctcgaactgcacctcatggtcatgttatggttgagctggtagcagaactcgaaggcattcag
tacggtcgtagtggtgagacacttggtgtccttgtccctcatgtgggcgaaataccagtggcttaccgcaaggttcttc
ttcgtaagaacggtaataaaggagctggtggccatagttacggcgccgatctaaagtcatttgacttaggcgacgag
cttggcactgatccttatgaagattttcaagaaaactggaacactaaacatagcagtggtgttacccgtgaactcatg
cgtgagcttaacggaggggcatacactcgctatgtcgataacaacttctgtggccctgatggctaccctcttgagtgc
at

Fig. 1. A fragment of the '*t*-modified' SARS CoV-2 complete genome (MN988668.1, GenBank) given in FASTA format.

According to today's knowledge, an RNA is a chain of codons (exons) to transfer genetic information and introns. Unfortunately, the role of the last ones is not well known [5]. The sequencing of RNA or DNA is the searching and identifying nucleotides by instrumental means. Codons in RNAs start with '*aug*' combination of nucleotides and end with one of the following three combinations: '*uaa*', '*uag*', or '*uga*'.

Considering that some DNAs may consist of several billions of nucleotides, the mathematical methods are widely used in genomics [6]. For instance, the RNA's symbols are substituted by number values, and DNA is considered a digital message. Then, the signal processing techniques are used for quantitative studies [7],[8]. They allow gene finding using digital filtering, searching for long-range correlation, intron/exon areas identification, and RNA and DNA visualization. Modern gene databases like GenBaNK [9], GISAID [10], etc. allow rather powerful graphical tools of visualization of DNA and RNA chains and codon distributions along these chains.

One of the many possible ways to digitalize and visualize an RNA strand would be to represent it by a walk on a plane [11]-[20]. For this purpose, the symbols *a, c, g, t* are coded by $\{(0,1);(1,0);(0,-1);(-1,0)\}$, for instance. An imaginable walker or a particle moving along an RNA chain performs a step in the direction of a vector corresponding to a nucleotide in the studied chain. In this way, the trajectories of these walks are computed and imaged on a plane. Additionally, the DNA walks can be plotted on a complex plane $\{a,c,g,t\} \Rightarrow \{(1);(-i);(-1);(i)\}$ with $i=\sqrt{-1}$ [14]. Similarly, the walks of 20 proteins can be built in 20-D space, and their projections are studied visually in 2-D or 3-D domains [15]. In Ref. [16], the vector components are the nucleotide triplets, and a long DNA consisting of 64 codon triplets can be represented as a curve in a 3-D space. Some

original and reviewed results on digitalization methods and imaging of RNA, DNA, and proteins can be found in [18], for instance.

The DNA walks allow the detection of codons and introns, hidden RNA periodicity, and RNA fragments' palindromes. Additionally, the RNA/DNA and protein molecules can be simulated using molecular modeling methods [19].

Many theoretical and experimental studies show fractality of RNA, DNA, and proteins [14],[20]-[27]. Fractals are self-similar or scale-invariant objects. It means that small 'sub-chains' geometry can be repeated on larger scales, although they can be randomly distorted. The exact reasons for this fractality are unknown, and it is arisen in complex nonlinear systems, as a rule. It is supposed that fractality allows a trade-off between different factors forming spatiotemporal structures. In simple cases, fractals are easily seen in images. A measure of fractality can be calculated numerically, and it is called the fractal dimension d_F. For instance, the fractal dimension of a line is $d_F = 1$. A geometrical distortion of self-similar shape increases d_F.

A biopolymer chain placed in a solvent is bent in a statistical manner. In the ideal case, a linear polymer has a fractal dimension $d_F = 2$. The fractal dimension of polymers depends on the solvent and its temperature [27]. Because polar solvents have frequency-dependent properties, they can be adjusted by applied microwave frequencies with the following influence on polymer fractal dimension. Then, some bioreactions can be controlled by this weak irradiation [28]. It is proposed to perform these reactions in high-gradient microwave fields [29].

In polymer theory and practice, fractality influences the chemical reaction rate, diffusion, surface absorption of long-chain and globular molecules, etc. [26],[27],[30]-[33]. For instance, DNA walks are recently applied to diagnose cancers following gene mutations [34],[35]. It was found that the mutated areas of DNA chains show a higher level of randomness, and it is measured by calculating the fractal dimension of walks.

As it has been already mentioned, a DNA or RNA can be treated as a text message. Many attempts to create a grammar theory of DNA are known. Initially, it was supposed that the DNA grammar is isomorphic to a human language [36],[37]. Codons in an RNA chain are considered as words, and introns are thought of as semantic information carriers. There is still no complete proof of this theory, although the formal grammar methods are used in genomics widely [38]-[42].

To find an equivalency of natural and DNA languages, they should be studied by the same methods, and comparable outcome data be obtained. For instance, in [41], a study is performed on the word frequency distribution using Sichel's compound Poisson distribution function [42]. After analyzing many DNA sequences and natural language texts, it is shown that both fit rather well the modified Poisson approximation if some definitions on DNA words and text sentences are agreed to be kept. A conclusion from Ref. [41] is that "DNA sequences have features that resemble natural language, and it may be modeled by linguistic methodology." To proceed further with this conclusion, the same algorithm should be applied to natural language texts and DNA sequences to obtain quantitative estimates.

The theory and mathematical modeling of RNAs and DNAs of viruses and cellular organisms is a highly complex matter because thousands of nucleotides compose these organic polymers partly in a random manner, unclear mechanisms of their synthesis, and pathological consequences of mutations of viruses for host organisms. Many mathematical tools have been developed, but new algorithms and studies can be fruitful, allowing new looks from different sides of this problem.

In this paper, a metric-based and text-independent algorithm for building distributions of the chosen fragments and the studied texts is developed. It is based on calculating the inter-symbol distance between a binary text sequence and an under-the-search fragment. The method is applied to calculate the distributions of *atg*-codon starting triplets and single-symbols of SARS CoV-2, Dengue, and Ebola virus sequences and build their geometrical schemes. Mutations of viruses are with the geometry variations of these schemes, and this algorithm allows easily tracking these

mutations and quantitively and visually estimating them. Statistical studies are with the calculation of fractal dimensions of the obtained distributions. They are used together with the visual analyses to class the mutations of several tens of samples of the mentioned viruses.

2. Metric-Based RNA Walk Algorithm

As it has been stated above, for both DNA and RNA descriptions by characters, their alphabet consists of four nucleotides. This designation is used for statistical studies of RNA and DNA if their physicochemical properties are outside the research scope.

Description of proteins requires 20 nucleotides with their symbolic designations. Non-hieroglyphic human language texts are with several dozens of characters. As modern programming languages allow the processing of multi-character strings, interesting results can be achieved in DNA research with this method.

Computers use characters coded by binary units. For instance, UTF-8 format allows encoding all 1,112,064 valid character code points, and it is widely used for the World Wide Web [43]. In this format, the first 128 characters (US-ASCII) require only one byte (eight binary numbers). If the DNA sequences are represented by binary units initially, then calculating the DNA sequences' numerical properties can reduce time.

Because the DNA/RNA chains are now written by binary sequences, they can be characterized quantitively using a suitable technique. One of them, proposed here, is to calculate a metric distance between the binary-represented symbols and a base "moving" along the characters' chains. Particularly, this distance can be Hamming one [44],[45], although other metric formulas used in the complex system and code theory are applicable as well [46]-[48].

This procedure gives quantitative numbers for each move, and then any pertinent signal-processing algorithm can be applied to these results. If needed, these DNA/RNA computed data are

compared quantitively with the study of human language texts, which have been processed in a similar way. The above approach could be helpful in understanding the DNA/RNA sequences better.

Consider an algorithm with the calculation of Hamming distance between a moving base and a genomic sequence of characters. This base can be *atg*-triplets or any given series of symbols, including a single one.

Hamming distance is a metric for comparing two binary data strings, and it is the number of bit positions in which the two bits are different. To calculate the Hamming distance $d_H(A,B)$ between two strings A and B, the XOR operation $(A \oplus B)$ is calculated and the total number of '1' s in the resultant string is counted.

In Matlab, a function $pdist2(A,B,\text{'hamming'})$ is responsible for calculating the mentioned distance between binary strings. Some other metrics are available in Matlab as well. It gives a percentage of '1' s in the resultant string of Hamming's distance calculation. If the compared strings are different in length, then Levenstein's distance [47] is used, for which several Matlab functions are found on the Internet [49].

Our algorithm is explained in Fig. 2 using the base '*atg,*' which is a replicating start triplet in DNA/RNA sequences.

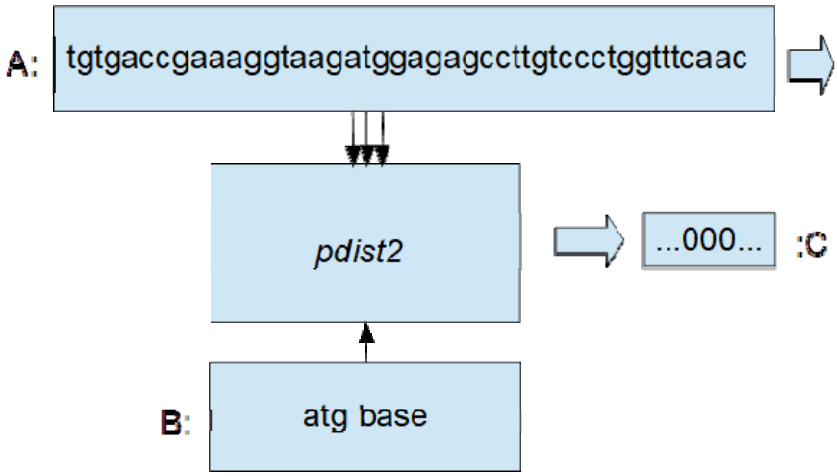

Fig. 2. Algorithm's flowchart.

Each symbol in an RNA sequence A of the length N (number of nucleotides) is compared with the ones in the triplet base $B:\{atg\}$. This number should be divisible by 3 because of using triplets. Otherwise, the studied sequence is extended by additional '*a*'s to make the number of nucleotides divisible by 3, which does not change our *atg*-walk results. Or one can use the mentioned Levenstein code [47]. Then, a $3N-$string C is created using a Matlab function $pdist2(A,B,\text{'hamming'})$. A numerical algorithm can process this new string to obtain helpful information.

Particularly, we calculate positions x_i of *atg*-triplets in an RNA sequence A and define the three neighbouring zeroes in the C string. After that, we calculate the word length $l_{i,i+1}^{(atg)}$. In our algorithm, a 'word' is a nucleotide sequence starting with an *atg*-triplet, and all symbols up to the next starting one (Fig. 3).

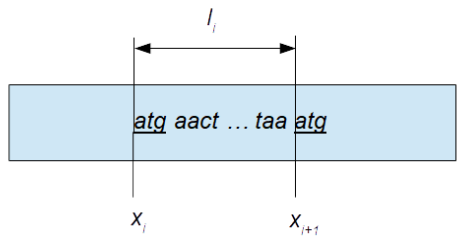

Fig. 3. A word and its length.

The number of these triplets is calculated by our code and verified by a Matlab function $count(C, 'atg')$. It is supposed that these triplets compose a 'scheme' of RNA, and therefore it can be used to characterize the genome chains.

To validate the algorithm, it is applied to many available virus genomes, and the calculated *atg*-positions are compared with those available from databases.

3. Main Results

3.1. Study of atg-Walks of SARS CoV-2 Virus Full Genome Sequences

In this paper, essential attention is paid to the study of SARS CoV-2 complete RNA genome sequences. The data used here and throughout this whole paper are from two genetic databases: GenBank [9] and GISAID [10]. A part of the studied genome sequences for this and other viruses is provided in Appendix 1.

Consider applying the developed approach to the complete genome of a Wuhan RNA sample MN988668.1 (GenBank). It consists of 30123 nucleotides, and it has 712 *atg*-triplets (See Table 1, first row, Appendix 1). Fig. 4a shows the distribution of *atg*-triplets for the first 5000-nucleotides of this complete genome by points.

The triplet positions x_i are represented by bars in Fig. 4b along a strip of an arbitrarily chosen width to make this distribution more understandable. It is seen that the x_i-distribution has repeating motifs on different scale levels, i.e., this distribution can have fractal properties. In Section 3, our initial assumption on fractality of *atg*-distributions is confirmed: we calculated the fractal dimensions of complete genomes of several tens of virus sequences. Presumably, the *atg*-triplets are distributed along with the RNA sequences of studied viruses according to the random Cantor multifractal law.

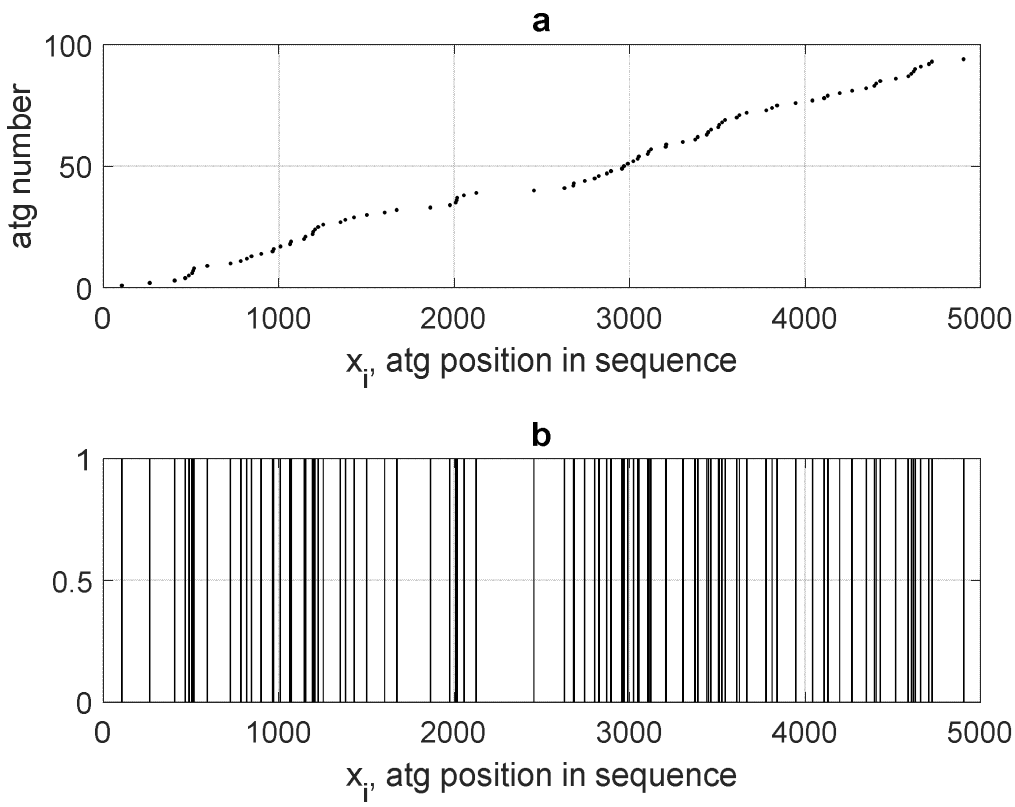

Fig. 4. Positions of *atg*-triplets along the genome sequence of SARS-CoV-2 virus MN988668.1 (GenBank) given for the first 5000 nucleotides.

Fig. 5 shows the distribution (in lines) of *atg*-triplets along with complete genome sequences for nineteen SARS CoV-2 viruses taken from GenBank and GISAID databases (see Table 1, Appendix). It is seen relatively compact localization of triplet curves despite the viruses of different clades and lines.

It confirms the conclusions of many specialists that no new strongly recombined strains have appeared to this moment despite many mutations found to date. Two inlets show the beginning and the tails of these curves to illustrate details. Although, in general, these curves are woven firmly, the tales are between the bat's SARS-CoV-2 light-blue curve (*hCoV-19/bat/Cambodia/RShSTT182/2010*, Table 1, row 6, Appendix 1) and the black trajectory obtained for a sequence from Brazil (*hCoV-19/Brazil/RS-00674HM_LMM52649/2020*, Table 1, row 14, Appendix 1).

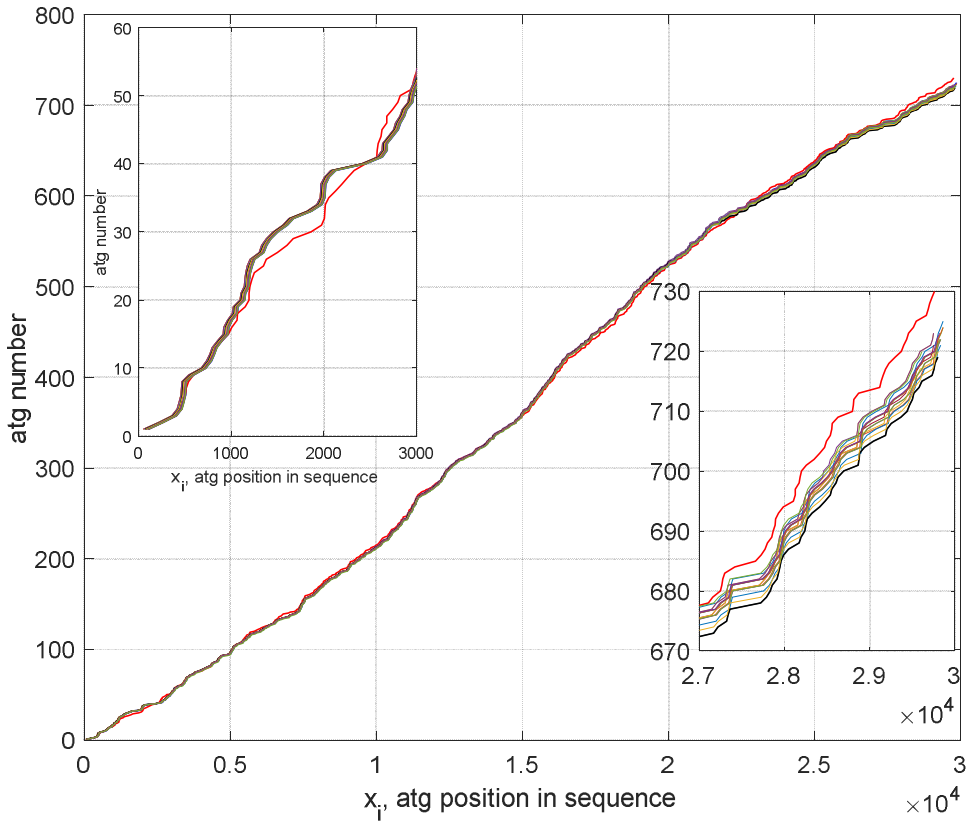

Fig. 5. Distributions of *atg*-triplets of nineteen SARS Cov-2 complete RNA sequences (Table 1, rows 1-19, Appendix 1). Inlets show the *atg*-distributions at the beginning and end of genome sequences.

A detailed study of each virus from Table 1, Appendix 1 shows that each considered sequence has an individual *atg*-distribution or 'fingertip.' It means that mutations are combined with the combined variations of word content, word length, and number. However, other mutations with variation only

word content exist. However, the *atg*-walks cannot see them, and the single-symbol distributions considered below will help us detecting these modifications of viruses (See Section 4).

Fig. 6 shows a detailed comparison of samples of four viruses causing increased troubles of the specialists with the one from Wuhan, China. It is seen that the tales of three curves are closed between the Wuhan's and Brazil's trajectories are with the evolution of viruses with time, probably. Although the difference between these curves is not significant, the mutations may have complicated consequences in the rate of contagiousness of viruses. The inlets show the details of these curves in their beginning and end.

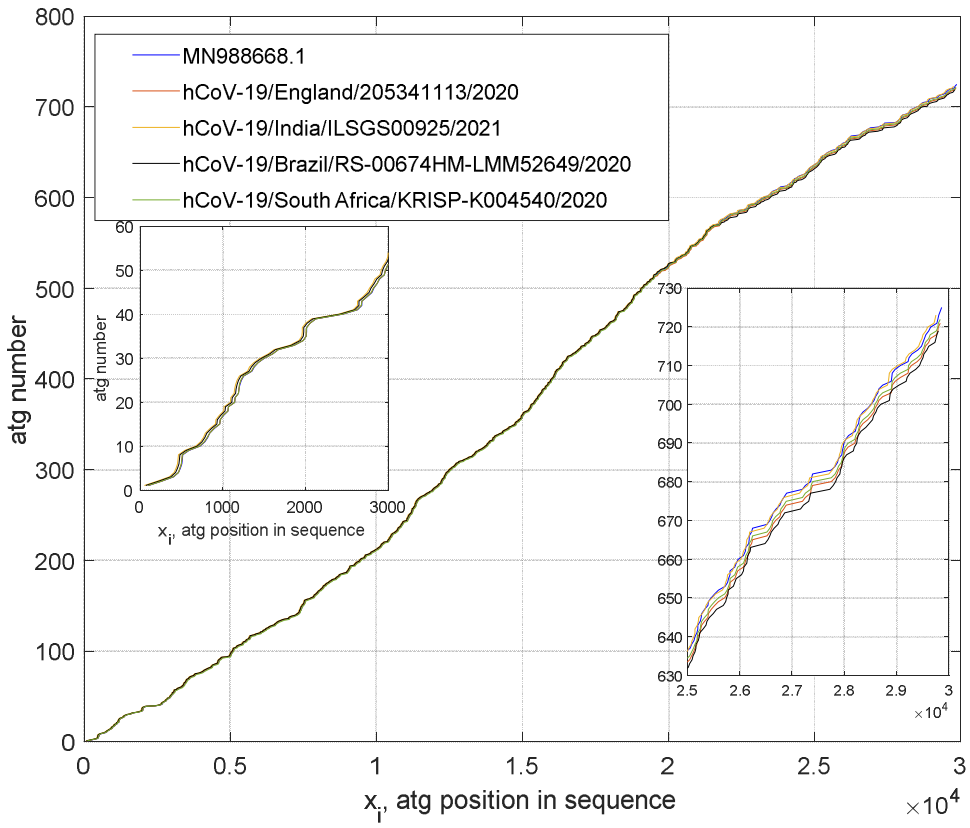

Fig. 6. Detailed distributions of *atg*-triplets for five trouble-making SARS Cov-2 complete RNA sequences (Table 1, rows 1,8,3,14,19, Appendix 1). Inlets show the *atg*-distributions at the beginning and end of genome sequences.

3.2. Study of atg-Walks of Full Genome Sequences of the Middle East Respiratory Syndrome-Related Coronavirus

Middle East respiratory syndrome-related (MERS) is a viral respiratory illness. Most people infected with MERS-CoV developed a severe respiratory disease which resulted in multiple human deaths. This virus originated from bats, but it initially spread through camels and was first registered in Saudi Arabia.

Our simulation of *atg*-distributions shows compactness of calculated curves, like the SARS CoV-2 characteristics (Fig. 7 and Table 2, Appendix 1). It follows that both viruses demonstrate relatively stable features towards the strong mutations connected with the recombination of the virus's parts. On average, MERS RNAs have a smaller number of *atg*-triplets and longer nucleotide words than SARS CoV-2 sequences.

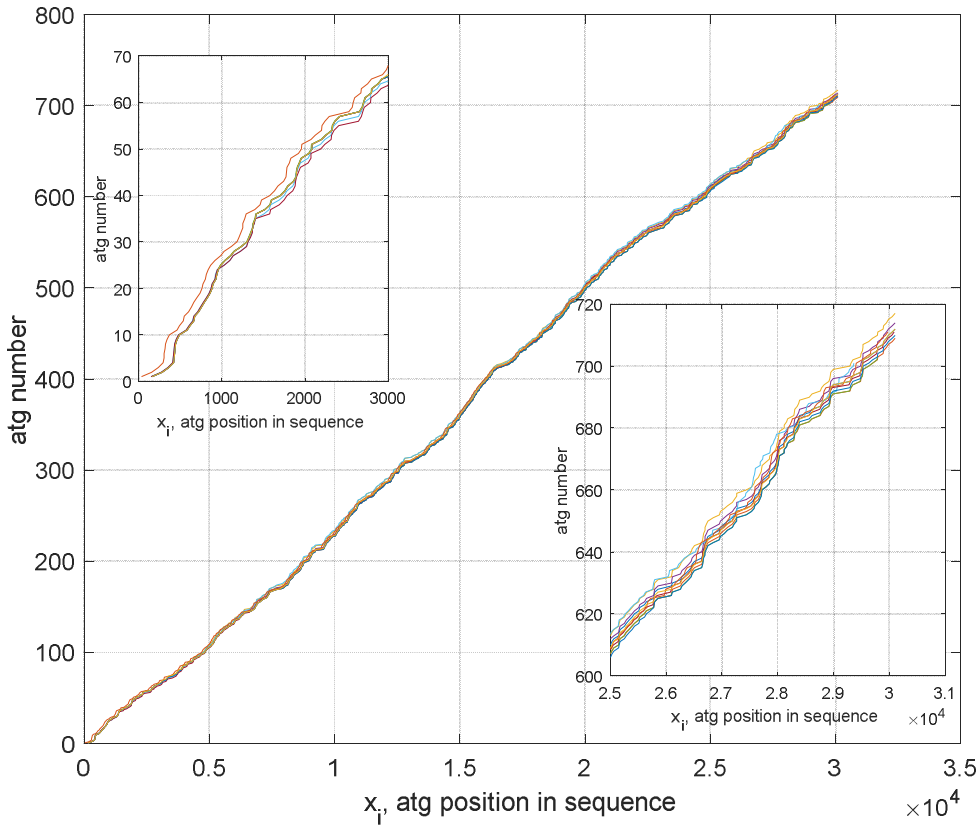

Fig. 7. Distributions of *atg*-triplets of ten samples of the MERS complete RNA sequences. Inlet shows the *atg*-distributions at the end of genome sequences (Table 2, rows 1-10, Appendix 1). Inlets show the *atg*-distributions at the beginning and end of genome sequences.

A comparison of RNAs of these viruses is shown in Fig. 8. The difference of *atg*-distribution is strong enough to distinguish these two viruses from each other.

In general, the two studied coronaviruses demonstrate relatively strong stability of their *atg*-distributions towards severe mutations leading to the variation of codons' positions, the word length, and number. It follows the conclusions of many scientists majoring in virology and virus genomics [50].

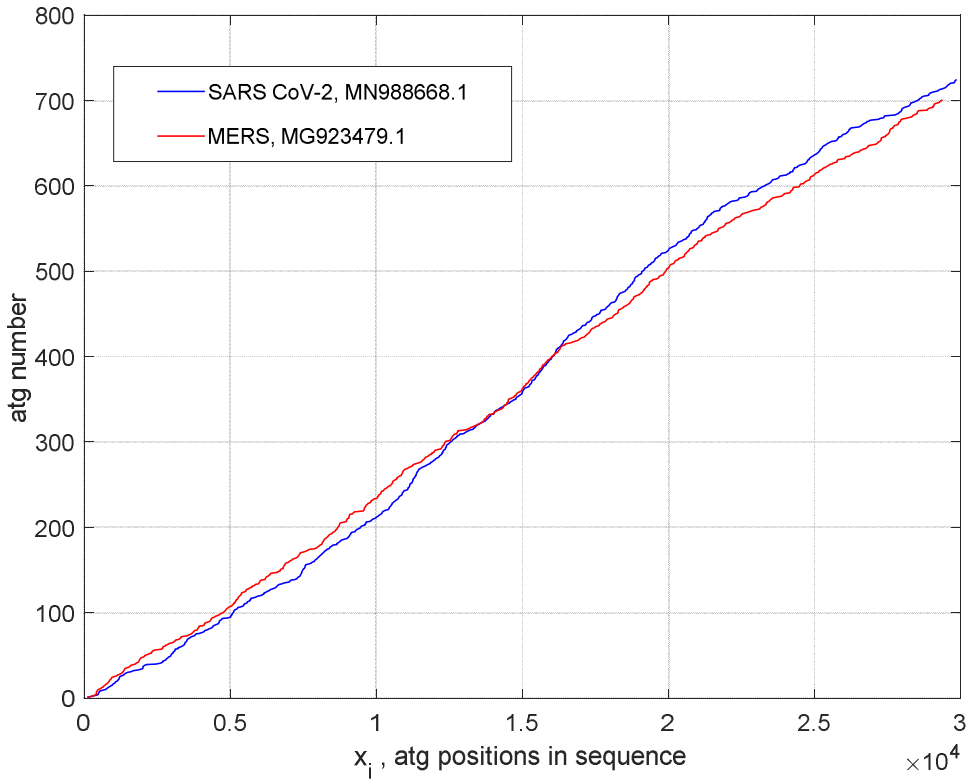

Fig. 8. Comparison of *atg*-distributions of complete RNA sequences of SARS CoV-2 and MERS viruses (Table 1, row 1 and Table 2, row 6, Appendix 1).

3.3. Study of atg-Walks of Full Genome Sequences of the Dengue Virus

Dengue virus is spread through mosquito bites. Unlike the coronaviruses, the Dengue virus (Table 3, Appendix 1) shows less stability. For instance, the Dengue virus has five genotypes DENV 1-5, and around 47 strains.

Only several of them are studied below, for which complete genome data are available. Fig. 9 (Table 3, rows 1.1-1.5, Appendix 1) shows the *atg*-distributions of five sequences of the Dengue virus-1 found in China. Rather large dispersion of these sequences is seen from these graphs.

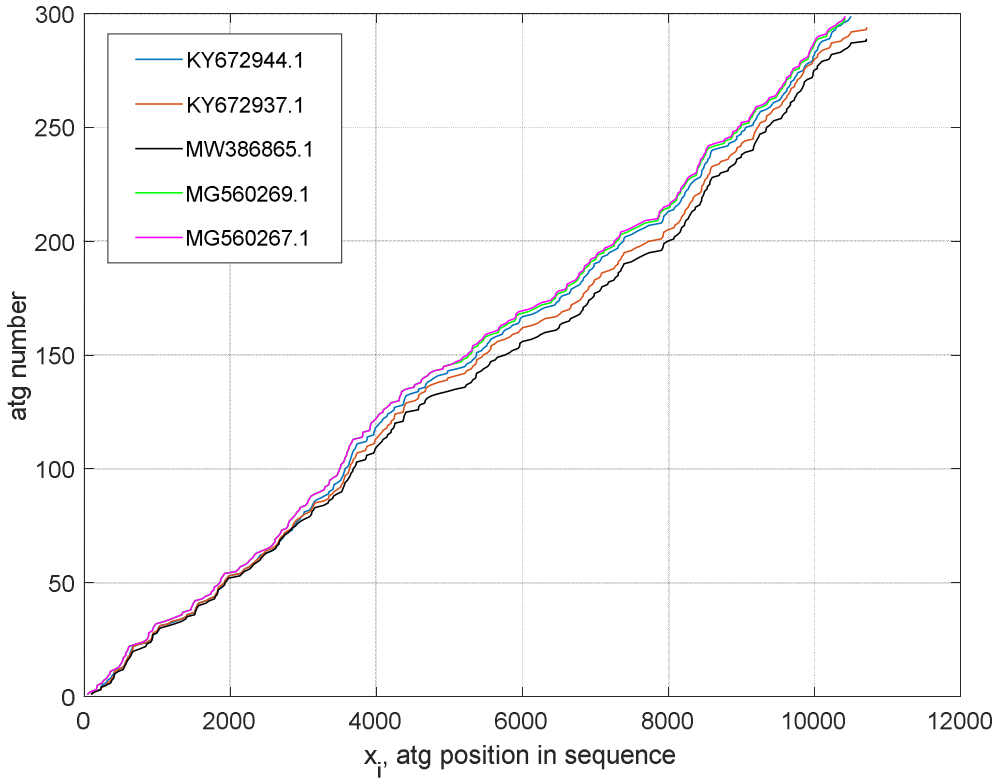

Fig. 9. Distributions of *atg*-triplets of five complete RNA sequences of the Dengue virus-1 (See Table 3, rows 1.1-1.5, Appendix 1).

Fig. 10 shows the *atg*-distributions of five complete sequences of the Dengue virus-2 (see Table 3, rows 2.1-2.5, Appendix 1). These distributions are more compactly localized, although their origin is from different parts of the world. In general, the observed Dengue virus-2 samples have decreased word length and increased number of words compared to the sequences of Dengue virus-1 (Table 3, Appendix 1).

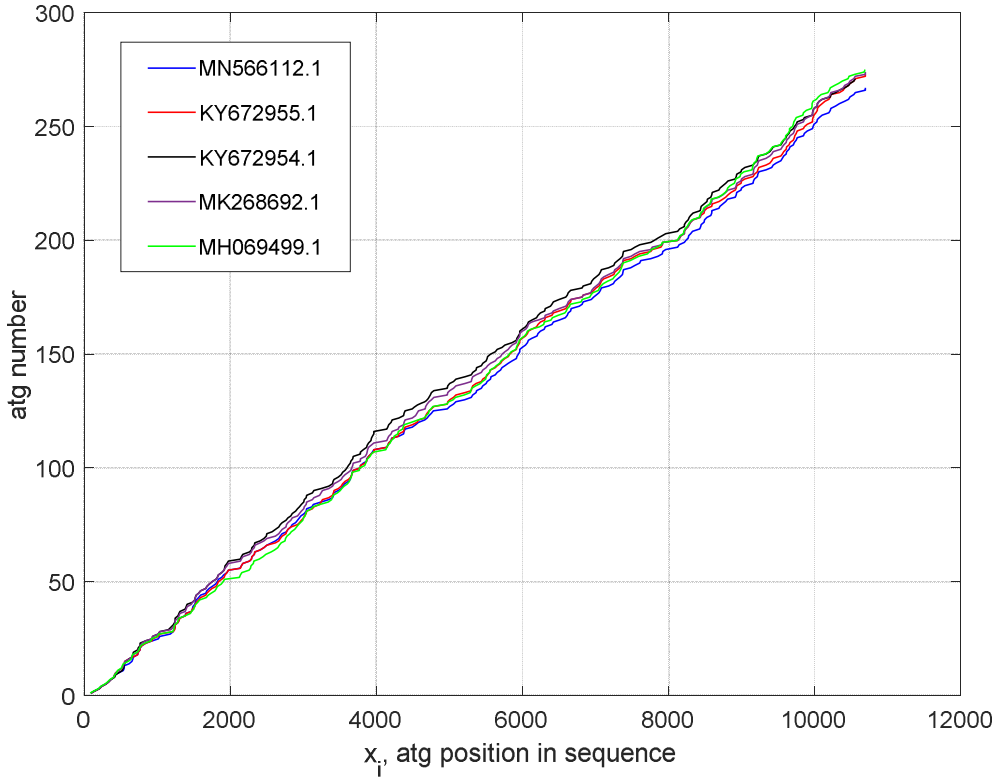

Fig. 10. Distributions of *atg*-triplets of five complete RNA sequences of the Dengue virus-2 (See Table 3, rows 2.1-2.5, Appendix 1).

In Fig. 11, five data sets for different strains of Dengue virus-3 are shown (Table 3, rows 3.1-3.5, Appendix 1). They have about the same number of nucleotides and comparable averaged lengths of words.

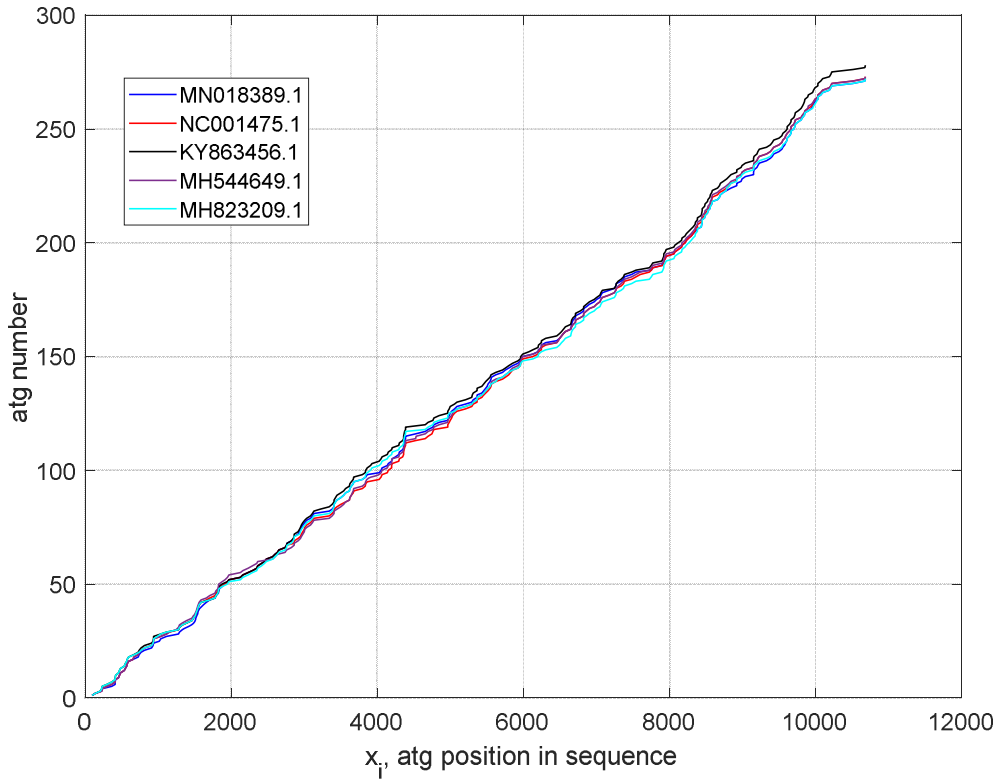

Fig. 11. Distributions of *atg*-triplets of five complete RNA sequences of the Dengue virus-3 (See Table 3, rows 3.1-3.5, Appendix 1).

In Fig. 12, three *atg*-distributions of a Gabon-strain [51] of Dengue virus-3 (Table 3, rows 3.6-3.8, Appendix 1) are given. It is supposed that this strain mutated from the earlier registered Gabon Dengue virus lines (Fig. 13, Table 3, rows 3.9, 3.10, Appendix 1). However, they are different in the length of complete genome sequences and their statistical characteristics, which are considered in Section 3 below.

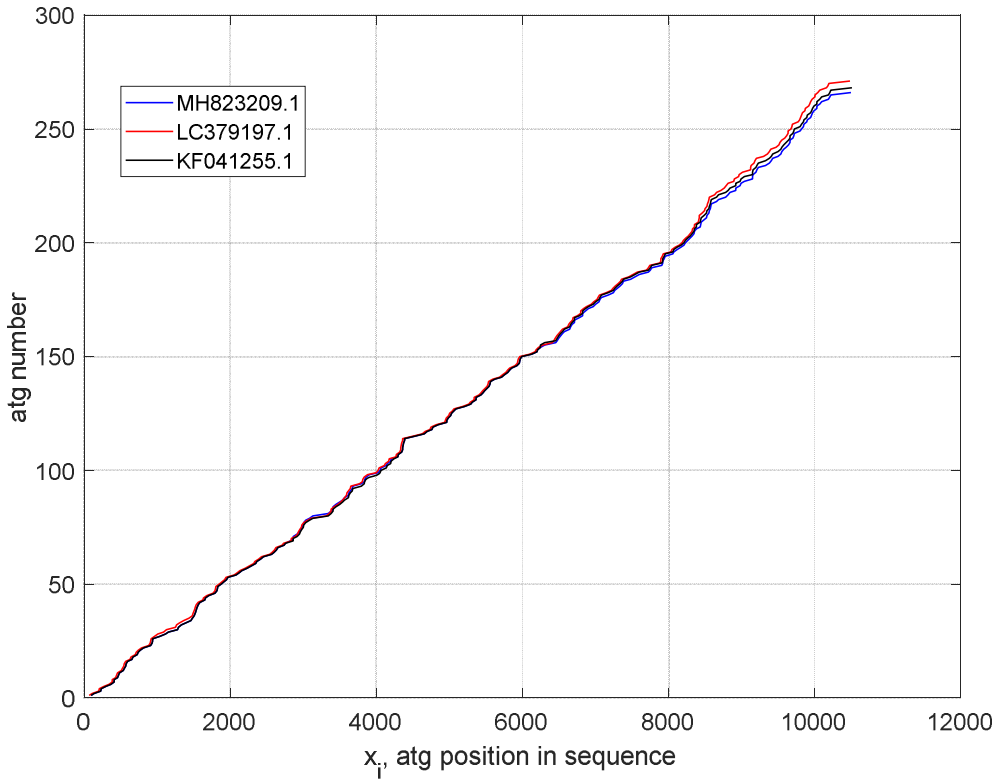

Fig. 12. Distributions of *atg*-triplets of three complete RNA sequences of Dengue virus-3 (Table 3, rows 3.6-3.8, Appendix 1).

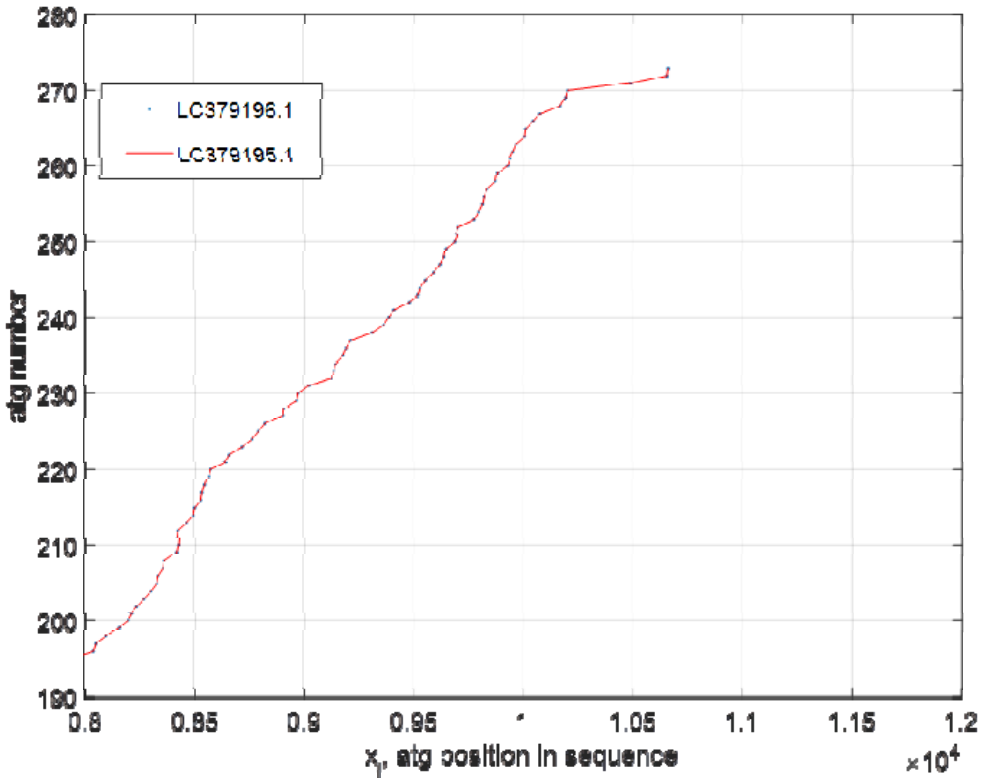

Fig. 13. Distributions of *atg*-triplets of two complete RNA sequences of Dengue virus-3 (Table 3, rows 3.9, 3.10, Appendix 1).

The following figure (Fig. 14, Table 3, rows 4.1-4.5, Appendix 1) presents the *atg*-distributions of complete RNA sequences for five Dengue virus-4 samples. They have individual and statistical differences with the above-considered Dengue viruses.

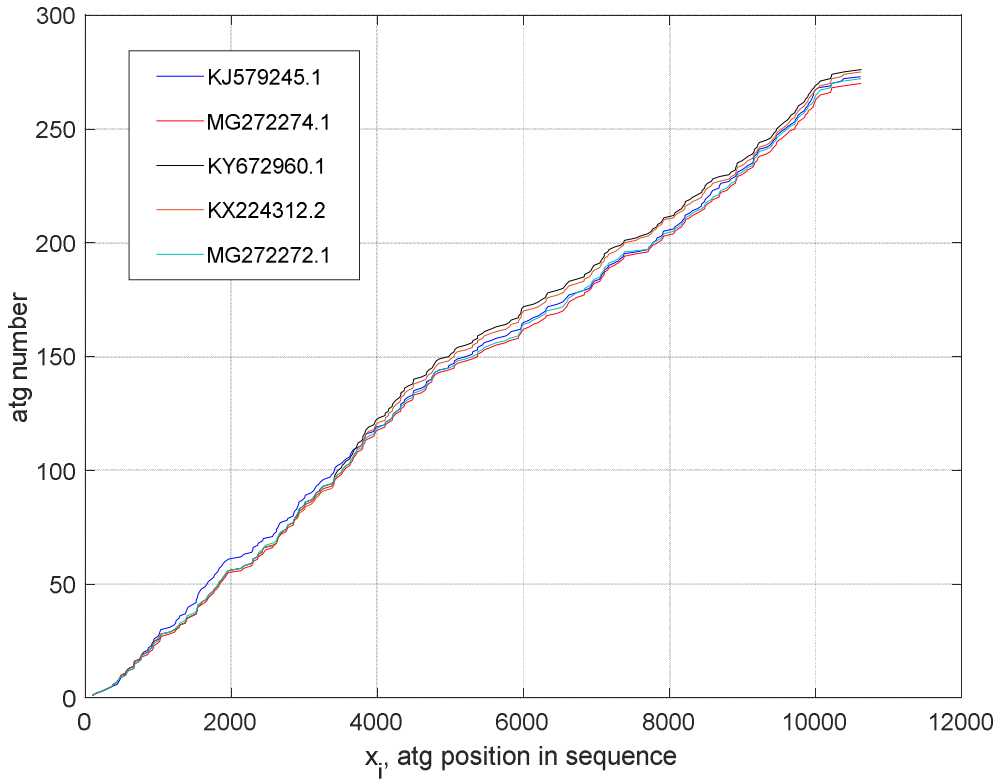

Fig. 14. Distributions *of atg*-triplets of five complete RNA sequences of Dengue virus 4 (See Table 3, rows 4.1-4.5, Appendix 1).

3.4. Study of *atg*-Walks of Full Genome Sequences of Ebola Virus

Like the Dengue virus, the Ebola one shows its instability and increased rate of mutations. There are four strains known in the world to this moment, although many other mutations can be found. Initially, the infection was registered in South Sudan and the Democratic Republic of the Congo, and it spreads due to contact with the body fluids of primates. This fever is distinguished with a high death rate (from 25% to 90% of the infected individuals).

The Ebola virus RNA consists of 19 000 nucleotides and more than three hundred *atg*-triplets. Fig. 15 shows four sequences of this virus belonging to the EBOV strain. Three of them are very close to each other, but the mutant Zaire virus (in red) has some differences from the three others.

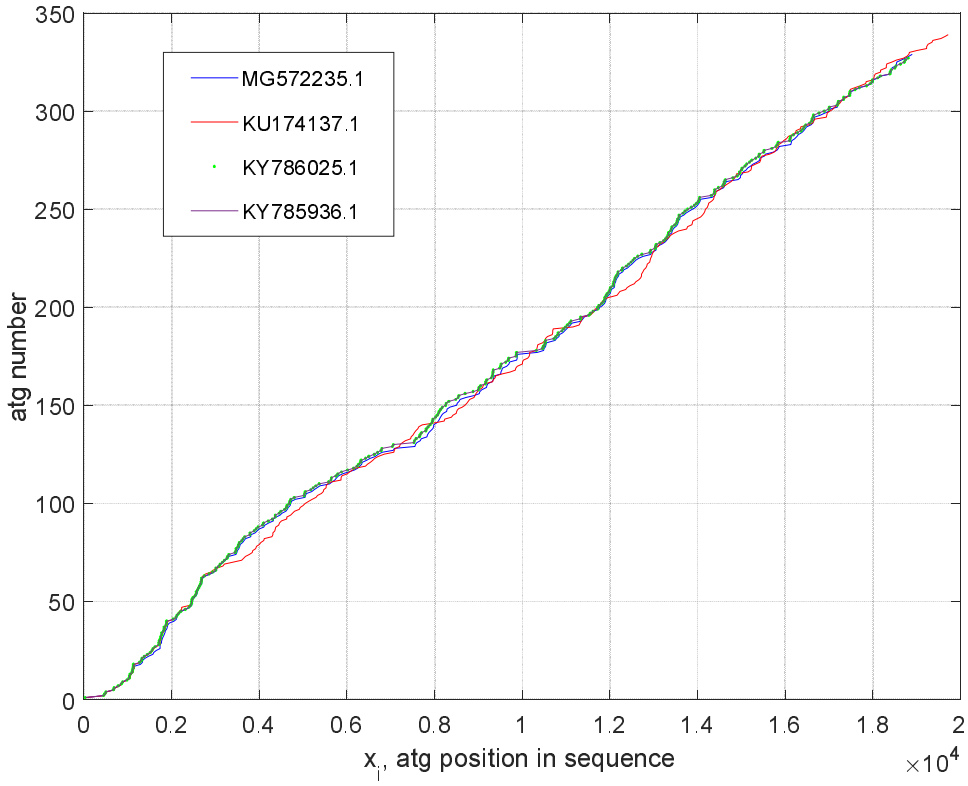

Fig. 15. Distributions of *atg*-triplets of four complete RNA sequences of Ebola (EBOV) virus (See Table 4, rows 1-4, Appendix 1).

As seen in Fig. 16 (Table 4, rows 5-7, Appendix 1), the Sudan-originated strain viruses have an increased number of *atg*-triplets and shorter words.

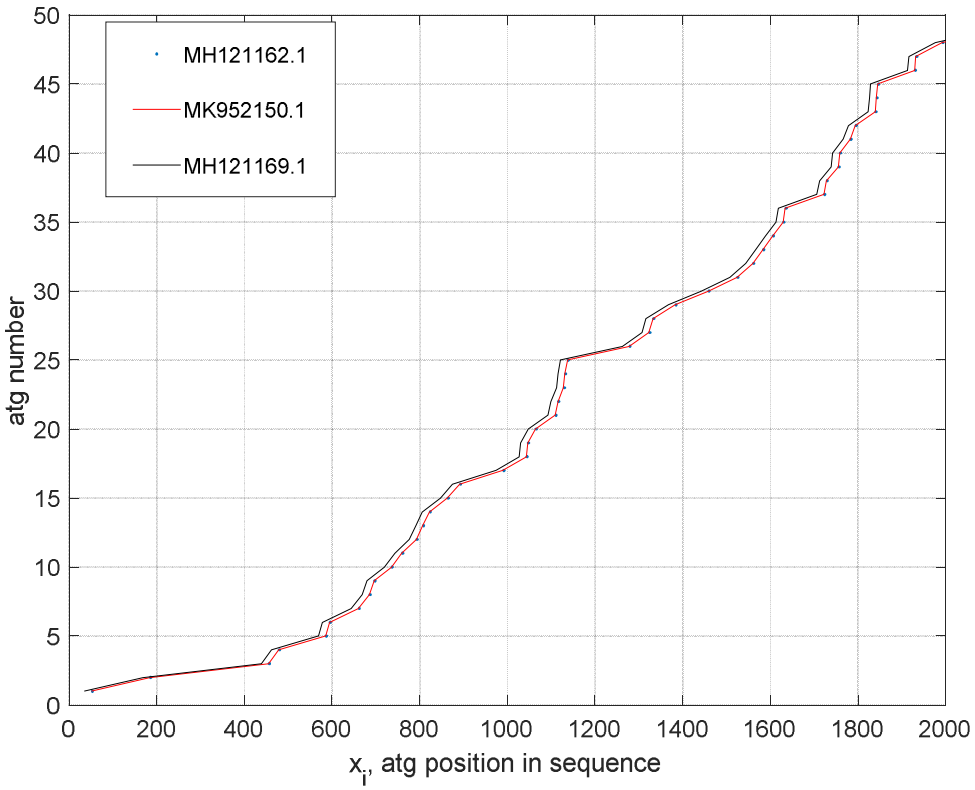

Fig. 16. Distributions of *atg*-triplets of three complete RNA sequences of Ebola (SUDV) virus (See Table 4, rows 5-7, Appendix 1).

The Bombali virus is considered a new strain of the Ebola virus registered in Sierra Leone, West Africa. The *atg*-distributions of the five studied here RNA sequences are different even visually from the two reviewed above, as seen in Fig. 17.

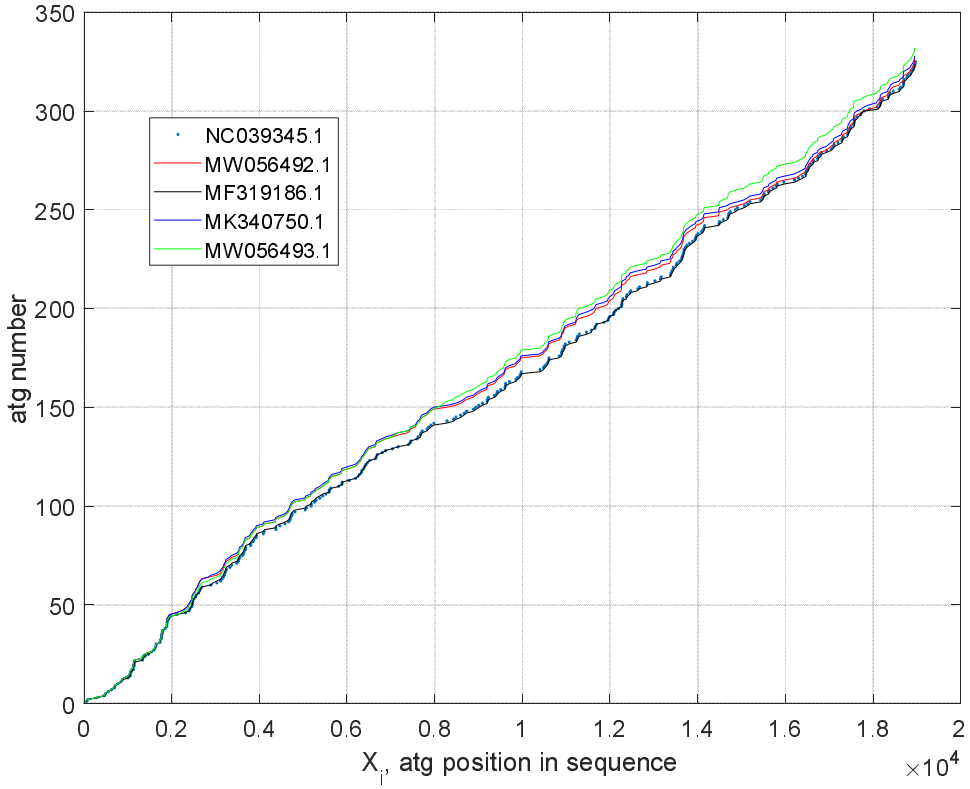

Fig. 17. Distributions of *atg*-triplets of three complete RNA sequences of Ebola (Bombali) virus (See Table 4, rows 8-12, Appendix 1).

Another Ebola virus strain that can be compared with the one studied above is the Bundibugyo virus, whose four *atg*-distributions are shown in Fig. 18.

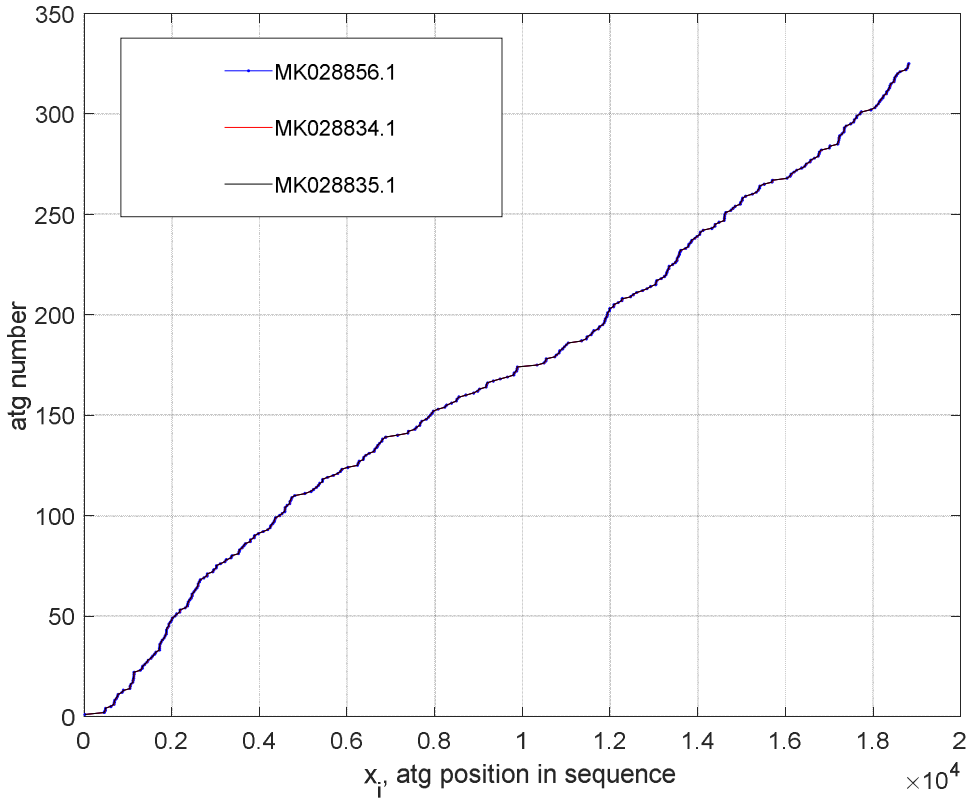

Fig. 18. Distributions of *atg*-triplets of three complete RNA sequences of Ebola (BDBV) virus (See Table 4, rows 13-15, Appendix 1).

The calculated distributions are consolidated in Fig. 19 to compare all the four strains, where, instead of points, the results are represented by thin curves to make distributions more visible. Here, the tendency of the *atg*-curve clustering is visible for each virus strain.

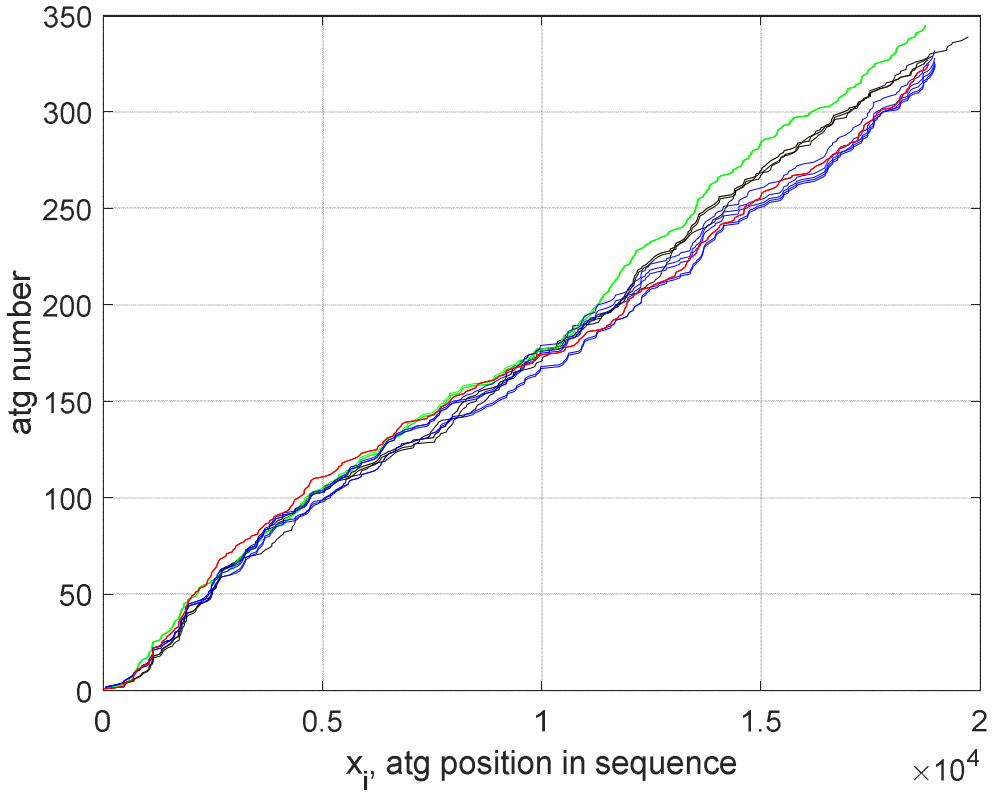

Fig. 19. Consolidated representation of *atg*-distributions of four strains of the Ebola virus (Table 4, Appendix 1). Black color- EBOV; Green color- SUDV; Violet color- Bombali; Red color- BDBV.

Reviewing all above-obtained results, the *atg*-walk is an effective visualization tool sensitive to the virus RNA's mutations connected with the number of codons' variation, width, and *atg*-coordinates. Additionally, these distributions supplemented by statistically calculated parameters can be 'fingertips' of mutated sequences, and these representations can help develop the phylogenetic trees of the virus strains and mutations.

3.5. Statistical Characterization of *atg*-Walks: Calculating, Mapping, and Processing of the Inter-*atg* Distance Values

The main unit called 'word' in this contribution is a nucleotide sequence starting with '*atg*' and symbols up to the next starting triplet (Fig. 3). The number of *atg*-s is calculated by our code and verified by a Matlab function $count(C, 'atg')$. These results are shown in Tables 1-4 in columns 3 (See Appendix). The Matlab functions $median(C)$ and $rms(C)$ calculate the median and root-mean-square (R.M.S) values of each sequence's word-length $l_{i,i+1}^{(atg)}$ distribution, correspondingly. The results are placed in columns 4 and 5 of the mentioned tables.

In many previous studies, the fractality of the distribution of nucleotides along with the DNA/RNA sequences is studied [14],[20]-[27]. It means that the motifs of small-size patterns are repeated on large-scale levels. Thus, the nucleotide distribution along a genome is not entirely random due to this long-range fractal correlation. The measure of self-similarity is its fractal dimension d_F that can be calculated using different approaches.

The large-size genomic data are often patterned, and each pattern can have its fractal dimension, i.e., the sequences can be multifractals [24]. This effect is typical in genomics, but it is also common in the theory of nonlinear dynamical systems, signal processing, brain tissue morphology, etc. [52]-[59].

Discovering fractality of genomic sequences is preceded by their numerical representation, for instance, by walks of different types [11],[12],[14]-[16],[21],[24]. Then, the each step value of a chosen walk is considered a sample of a continuous function, and the methods of signal processing theory are applied [7],[8].

In our case, a 'signal sample' is a word-length $l_{i,i+1}^{(atg)}$ value (see Fig. 3) arranged as a column signal sample vector. A particular distribution of the word lengths is shown in Fig. 20 by bars whose height

is equal to the word length. Then, the algorithms usually applied to the sampled signals can be used to compute the statistical properties of word-length distributions.

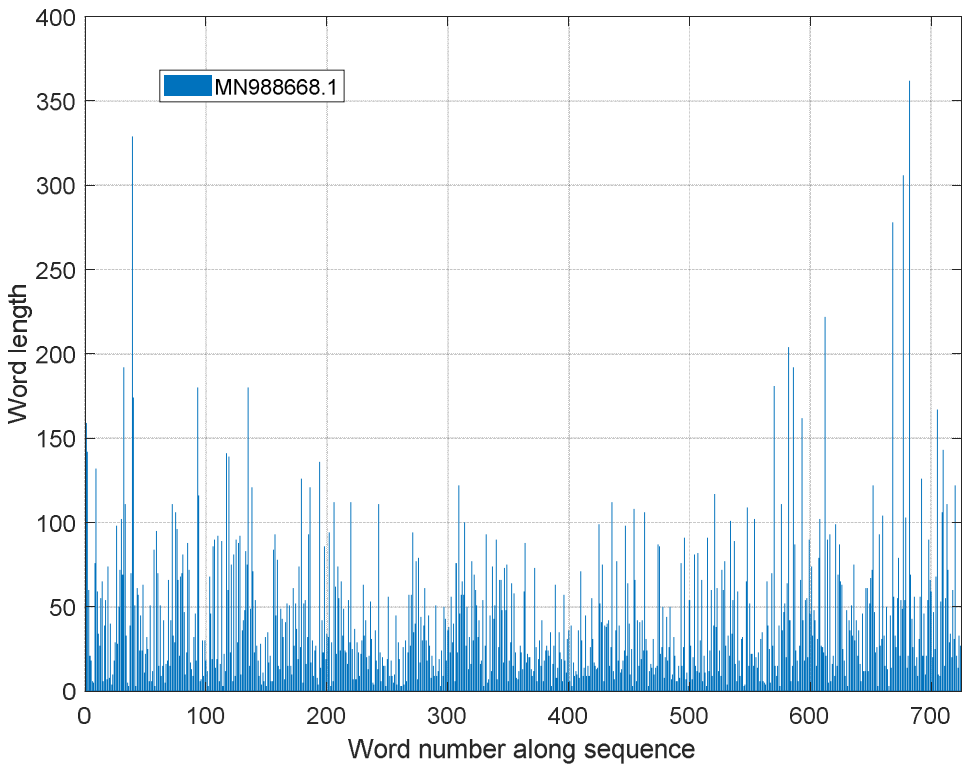

Fig. 20. Word-length $l_{i,i+1}^{(atg)}$ distribution in a SARS Cov-2 virus MN988668.1 sequence (GenBank, see Table 1, 1st row, Appendix 1).

In this contribution, the fractal dimension is calculated using a software package FracLab 2.2 [60] in order to compute the parameters of time-depending sequences of samples. Although many researchers tested this code, it is verified here to calculate the Weierstrass function, which is synthesized according to a given value of the fractal dimension [61]. It was found that this code provides results with reasonable accuracy if the default parameters of FracLab are used (Fig. 21).

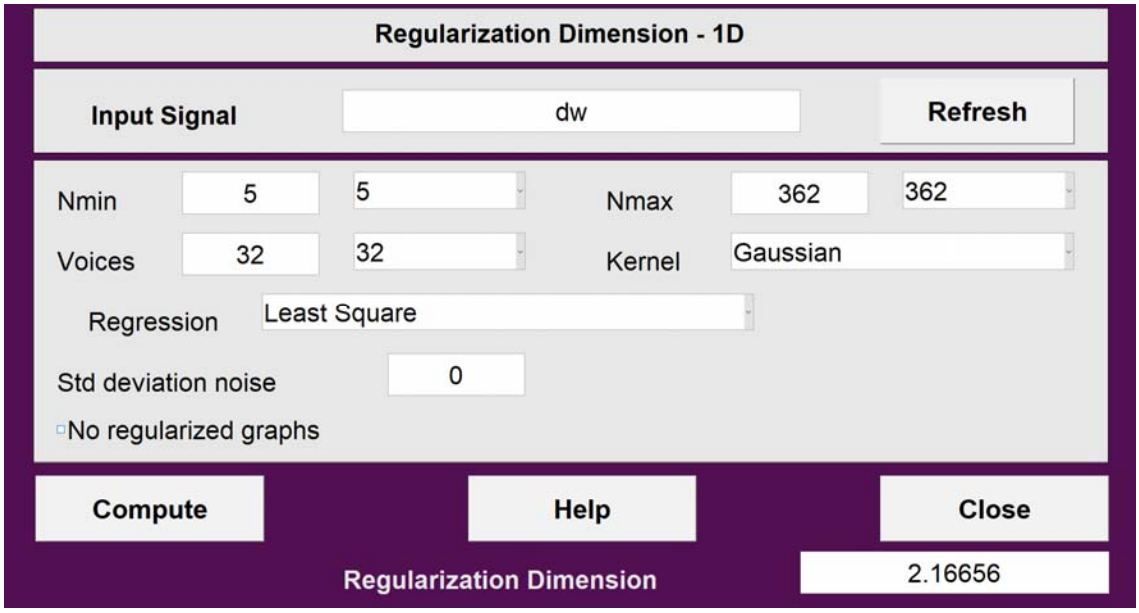

Fig. 21. FracLab window for calculation of fractal regularization dimension. The calculated dimension value $d_F = 2.16656$ is given for MN988668.1 (GenBank) virus RNA sequence.

In a strong sense, the fractal dimension is defined for the infinite sequences. In our case, the ones have only 730 -268 *atg*-triplets depending on a virus. Then, the fractal dimension values are estimated approximately. Comparing the Weierstrass function parameters' calculation, it is possible to conclude that the error in the worst case of the smallest *atg*-number can be in the order of several percent. It is acceptable for our analysis even of short-length RNA sequence viruses like the Ebola one.

In this research, applying the mentioned tool, it was discovered that all studied genomic sequences of the SARS Cov-2, MERS, Dengue, and Ebola viruses have fractality of their word-length distributions. The fractal regularization dimension values [56,57] are correlated with the corresponding dimension of uniform linear ideal polymers, which is $d_F = 2$ [26].

The fractal dimension calculation results are placed in columns 6 of Tables 1-4, Appendix. Fig. 22 shows the fractal dimension values of 17 genome sequences of SARS CoV-2 viruses in humans and two found in bats. Ten samples of genome sequences are shown in the same figure of MERS virus.

Although the *atg*-distributions of these viruses are visually close to each other (Fig. 8), the word-length $l_{i,i+1}^{(atg)}$ fractal dimension values are essentially different.

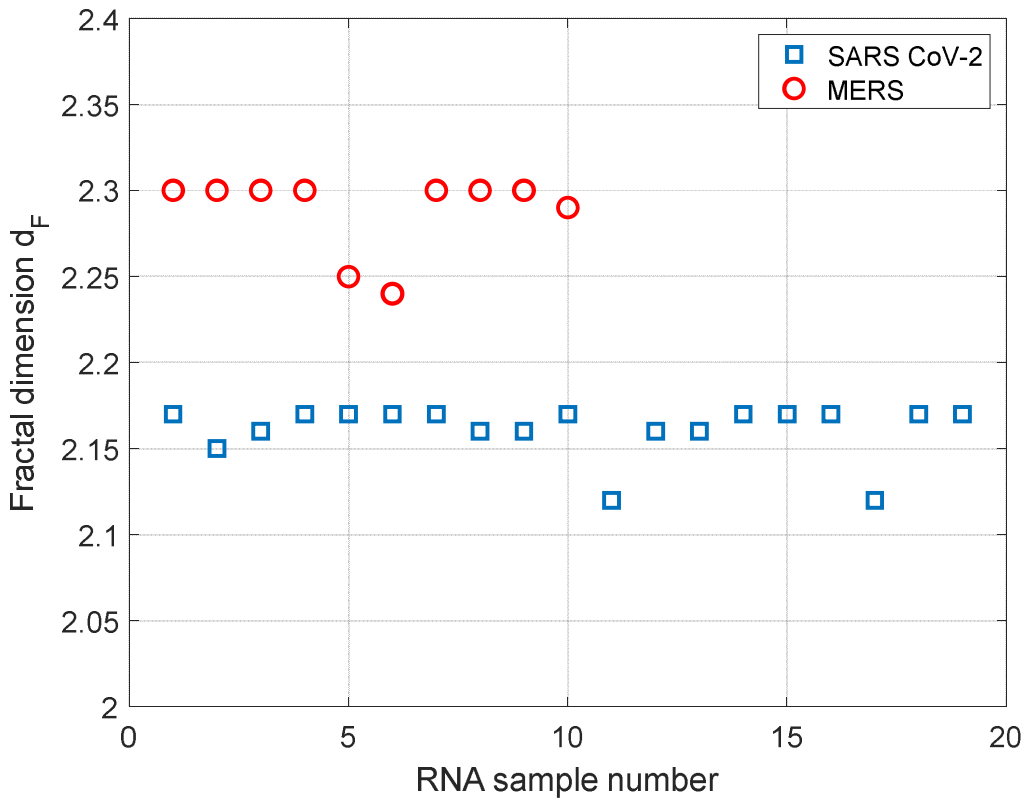

Fig. 22. Fractal dimensions d_F of word-length $l_{i,i+1}^{(atg)}$ distributions of complete genome sequences of SARS CoV-2 and MERS viruses.

The Dengue virus has five families and 47 strains; they have different *atg*-distributions and fractal dimensions. Some strains are close to each other according to the fractal calculations (Fig. 23). This gives the reason to conclude that the RNAs of the considered strains have similarities in the *atg*-distributions.

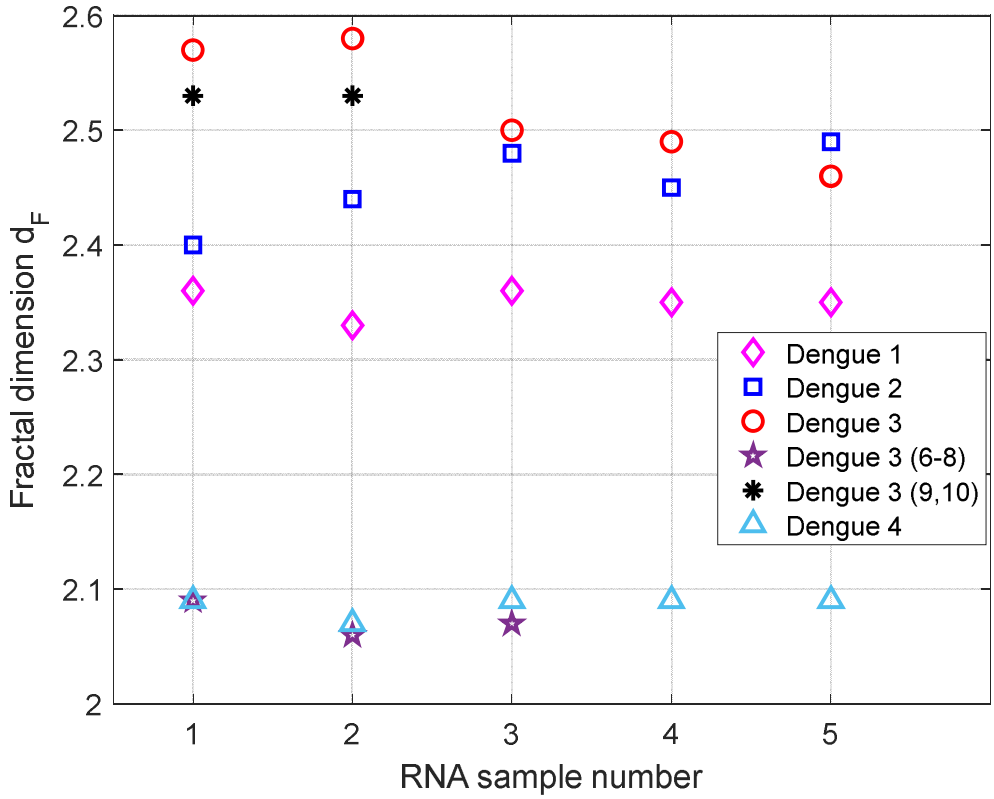

Fig. 23. Fractal dimensions d_F of word-length $l_{i,i+1}^{(atg)}$ distributions of complete genome sequences of Dengue 1-4 viruses and their strains. In round brackets, the numbers of rows are shown in Table 3, Appendix 1)

The same conclusion is evident in Fig. 24, where the fractal dimensions of several strains of the Ebola virus are given.

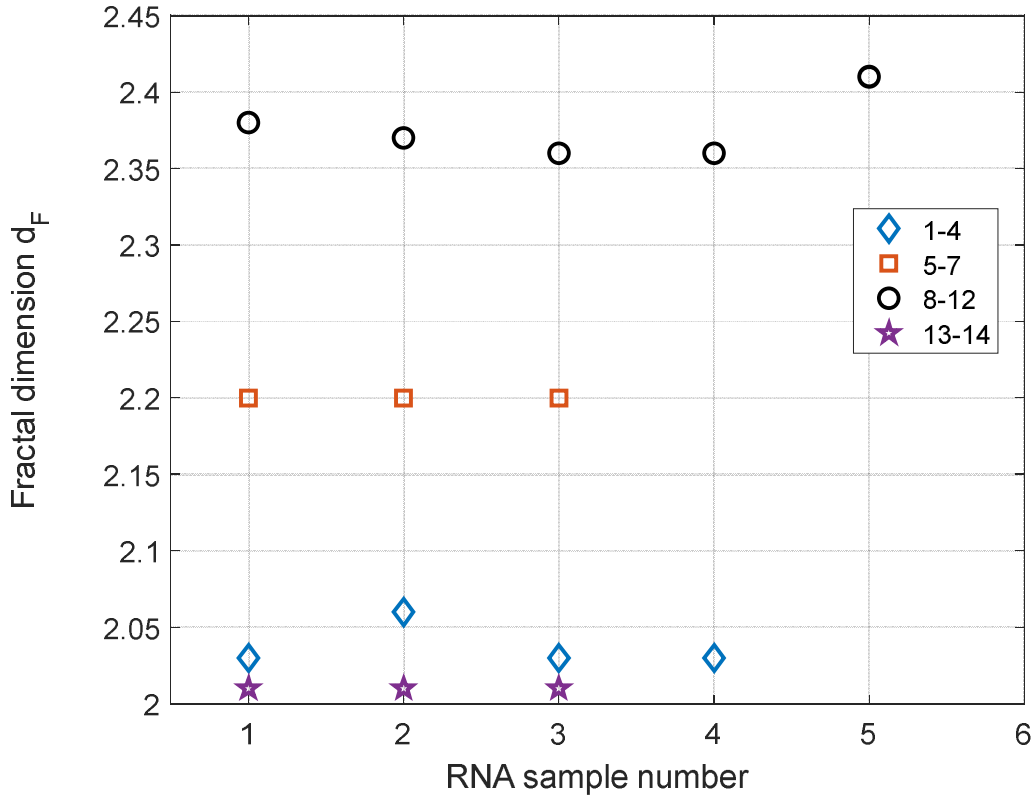

Fig. 24. Fractal dimensions d_F of word-length $l_{i,i+1}^{(atg)}$ distributions of complete genome sequences of Ebola virus strains. In round brackets, the numbers of rows are shown in Table 4, Appendix 1)

An analysis of the statistical characteristics of *atg*-distributions shows that the average word length $l_{i,i+1}^{(atg)}$ is coupled in a certain way with the fractal dimension. As a rule, word-length-decrease increases the fractal dimension, which means a more complicated distribution of *atg*-triplets.

3.6. Study of Single-symbol Metric-based Walks of SARS CoV-2 Virus Full Genome Sequences

Some viruses are prone to severe mutations coupled with re-modifications of *atg*-distributions, or their mutation rate is slow. Anyway, the variations of only codon content can lead to severe consequences, and these mutations should be studied carefully. The proposed algorithm allows the

creation of arbitrary-word-length metric-based walks, as mentioned in Section 2, including single-symbol one for the fine-level study of any viruses' mutations. The written with this purpose code was tested comparing calculated coordinates of single symbols and their positions found visually analyzing original data file. An example of these studies is considered below for several SARS CoV-2 viruses (Fig. 25). The *c*-curves are tightly woven with other although having individual shape. Analyzing Fig. 6 and 25, it seen that these virus sequences are products of mutations with shifting of *atg*-triplets and variation of symbols and their positions inside the coding words.

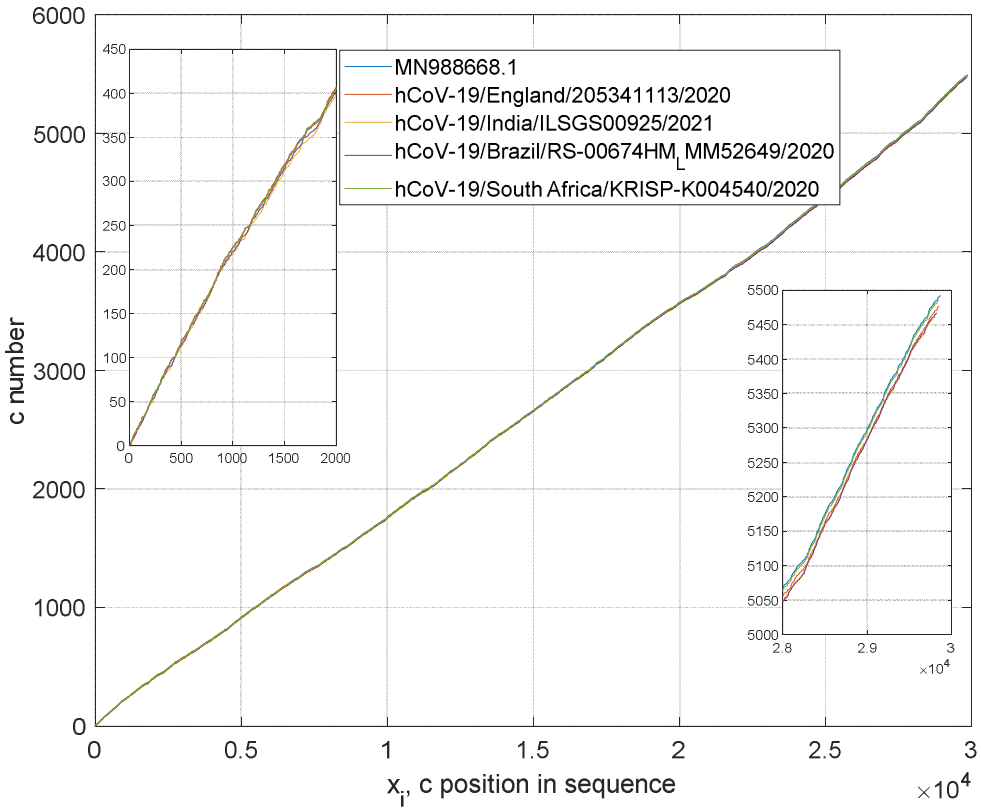

Fig. 25. Distributions of *c*-symbol along with three SARS CoV-2 complete RNA sequences (Table 1, row 1, row 11, and row 6, Appendix 1). Inlets show the *atg*-distributions at the beginning and end of genome sequences.

Statistical studies of *c*-distributions in SARS-CoV-2 RNAs chains are consolidated in Table 5 (Appendix 1) and Fig. 26 for all studied here 19 virus sequences. We examined the inter-*c* distances'

distributions, calculating the length $l_{i,i+1}^{(c)}$ between the neighboring *c*-symbols. The calculated fractal dimension values are recorded in Table 5, column 6, Appendix 1, and they are presented graphically in Fig. 26. It is seen that in comparison to Fig. 22 (*atg*-distributions), the fractal dimension, in this single-symbol case, shows more variability because of the word-content mutations in codons. Similarly, all three other single-symbol (*a*-, *t*-, and *g*-) distributions can be studied.

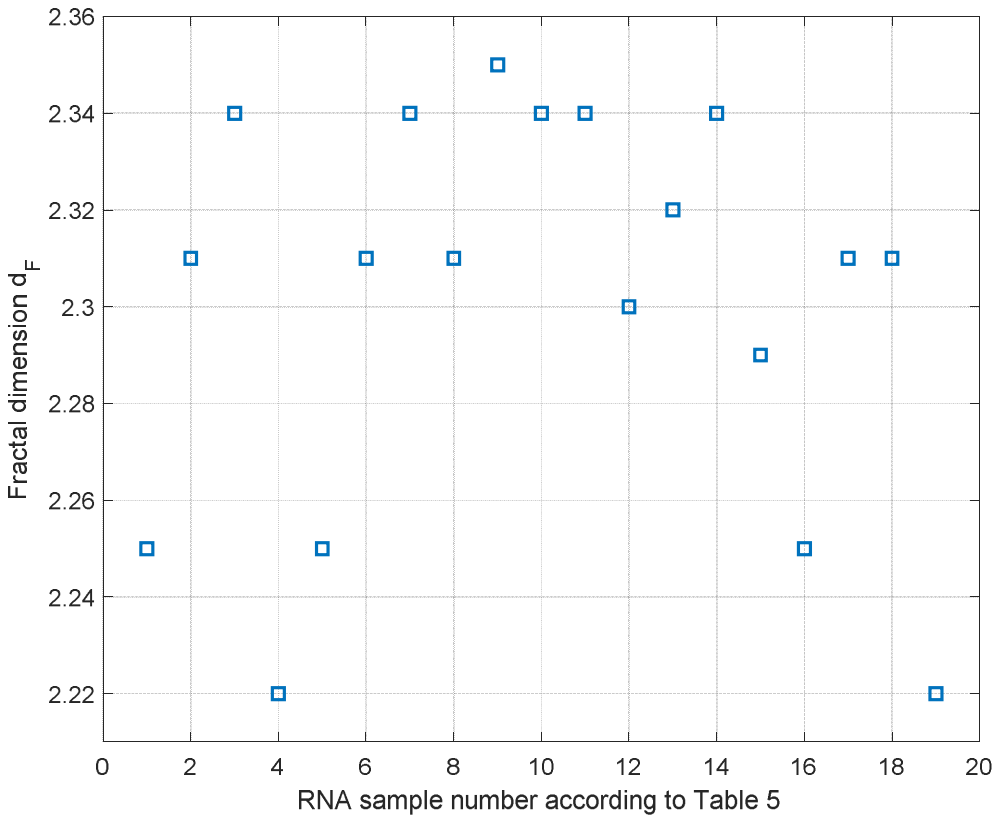

Fig. 26. Fractal dimensions d_F of inter-*c* distance length $l_{i,i+1}^{(c)}$ distributions of nineteen complete genome sequences of SARS CoV-2 (see Table 5, Appendix 1).

4. Discussion

The study of RNAs and DNAs of viruses and cellular organisms is a highly complex problem because of many nucleotides of these organic polymers, unclear mechanisms of their synthesis, and pathological mutation consequences for host organisms. Although many mathematical tools have been developed, new algorithms and views are exciting and can be fruitful.

The proposed algorithm is based on the calculation of Hamming distance between a binary expressed text of any origin and a fragment under the search. The position of this fragment in a studied genomic sequence is defined by analyzing the results of these calculations.

Primary attention is paid to calculating coordinates of *atg* triplets that start the codons in RNA sequences. The number of them in the studied RNAs is around several hundred and small enough compared to the number of nucleotides. At the same time, it gives a relatively strong level of detailing of a studied RNA. For each family of a virus, the number of *atg* triplets is not varied to a large extent.

It is proposed to name an *atg*-triplet set and coordinates of each its member as the *atg* scheme of an RNA. Studies on how these sets are transformed under mutations are performed by plotting one-dimensional trajectories of *atg* walks and their visual analyses for many tens of genomic sequences of several viruses found in GenBank® and GISAID databases. Additionally, statistical research is fulfilled by calculating the fractal dimensions of distributions of inter-*atg* triplet distances. It is shown that the *atg*-distributions and fractal dimension values tend to cluster in the boundaries of one virus family. Although more statistically meaningful studies are required to confirm this tendency, this effect can be significant in developing virus-family-oriented drugs and vaccines.

Additionally to the mentioned *atg*-scheme alteration, the mutations are accompanied by variations of codon content, and it is tracked by single-symbol walks built by the described algorithm. These trajectories and fractal dimension calculations allow following these mutations,

although more research is required in this way to understand better the geometrical issues of genomic sequences.

5. Conclusions

In this paper, a new approach has been developed for the analysis of RNA sequences. The proposed algorithm calculates the positions of any given fragments of nucleotides in RNA sequences using the Hamming metric distance between these fragment's symbols and characters in the studied sequences written both in binary form. It represents them by one-dimensional curves convenient for quantitative and visual analyses at the difference to known multi-dimensional genomic walks.

It has been proposed to consider the *atg* codon starting triplets as the elements of the genomic sequence schemes. Variations of these schemes represent essential mutations of viruses, followed by changing positions of these triplets in genomic sequences or even variations of their number. This idea has been studied by detailed simulations and quantitative and visual analyses of several pathogenic viruses, including SARS CoV-2, which genomes were taken from GenBank® and GISAID databases.

Studies of several tens of RNAs of SARS CoV-2 and MERS viruses show high compactness of their *atg*-distributions and stability of these viruses towards the mutations accompanied by essential variations of the number of codons, their lengths, and positions of coding words in genomic sequences. Unlike the studied SARS CoV-2 and MERS, the Dengue and Ebola viruses demonstrate stronger lability of *atg*-distributions expressed in the formation of several families of these viruses.

The visual analyses of these distributions have been followed by statistical studies of RNA 'words,' consisting of the symbols between the *atg*-triplets. The length of these words along the sequences is varied in a fractal random Cantor-set manner, presumably. The fractal dimension of these word lengths has been calculated using a genomic sequence 'signal' model. It has been found that the

various strains of viruses are different from each other by fractal dimensions of their word distributions in most cases.

The slow-mutation-rate viruses and their lines have been studied by single-symbol walks using the developed algorithm. Increased variability of fractal dimension values of the studied RNA single-symbol distributions of SARS CoV-2 virus samples has been shown while the *atg*-ones keep their relative stability. These results demonstrate unknown-earlier properties of virus RNAs important in understanding the viruses and their possibly dangerous mutations.

The proposed new calculation techniques and discovered peculiarities of RNA sequences are interesting in analyzing the virus RNAs and phylogenetic classification of viruses.

Abbreviations

RNA: Ribonucleic acid; DNA: Deoxyribonucleic acid; SARS CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2; cDNA: complementary DNA; GISAID: Global Initiative on Sharing All Influenza Data; UTF-8: Unicode Transformation Format-8 bit; *US-ASCII:* American Standard Code for Information Interchange*;* MERS: Middle-East Respiratory Syndrome-related.

Acknowledgments

The author thanks the GenBank® [9] and GISAID [10] genetic data banks, and all researchers placed their genomic sequences in them. The online text processing service of https://onlinetexttools.com/ is appreciated.

Authors contributions

All authors are contributed equally.

Funding

Not applicable.

Declarations

Ethical approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no conflicts of interest that are relevant to this research paper.

References

1. G. Meister, *RNA Biology: An Introduction*. Weinheim, Wiley-VCH, 2011.
2. K.R. Kukurba and S.B. Montgomery, RNA sequencing and analysis. *Cold Spring Harb Protoc.*, **11**, 951-967, 2015.
3. D.R. Garalde, E.A. Snell, D. Jachimowicz, et al., Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, **15**, 201-206, 2018.
4. A. Viehweger, S. Krautwurst, K. Lamkiewicz, R. Madhugiri, J. Ziebuhr, M. Hölzer, and M. Marz, Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis, *Genome Res*, **29**, 1545-1554, 2019.
5. G. Storz, An expanding universe of noncoding RNAs, *Science*, **296**, 1260-1263, 2002.

6. C. Nello and M.W. Hahn, *Introduction to Computational Genomics: A Case Studies Approach*. Cambridge, University Press, 2012.
7. Vaidyanathan PP. Genomics and proteomics: A signal processing tour. *IEEE Circ. Syst. Mag.*, 4th Quarter, 6-28, 2004.
8. J.V. Lorenzo-Ginori, A. Rodríguez-Fuentes, R.G. Ábalo, R. Grau, and R.S. Rodríguez, Digital signal processing in the analysis of genomic sequences, *Current Bioinformatics*, **4**, 28-40, 2009.
9. *GenBANK ®*, [https://www.ncbi.nlm.nih.gov/genbank/].
10. *Global Initiative on Sharing All Influenza Data (GISAID)*, [https://www.gisaid.org/].
11. J.A. Berger, S.K. Mitra, M. Carli, and A. Neri, Visualization and analysis of DNA sequences using DNA walks. *J. Franklin Inst.*, **341**, 37-53, 2004.
12. C.L. Berthelsen, J.A. Glazier, and M.H. Skolnik, Global fractal dimension of human DNA sequences treated as pseudorandom walks, *Phys. Rev. A.*, **45**, 89028913, 1992.
13. A.D. Haimovich, B. Byrne, R. Ramaswamy, and W.J. Welsh, Wavelet analysis of DNA walks, *J. Comput. Biol.*, **13**, 1289-1298, 2006.
14. C. Cattani, Complex representation of DNA sequences. In: *Bioinformatics Research and Development. BIRD 2008*. Edited by Elloumi M., Küng J., Linial M., Murphy R.F., Schneider K., and Toma C. Communications in Computer and Information Science, vol 13. Springer: Berlin-Heidelberg, 528-537, 2008.
15. A. Czerniecka, D. Bielinska-Waz, P. Waz, and T. Clark, 20D-dynamic representation of protein sequences. *Genomics*, **107**, 16-23, 2016.
16. F. Bai, J. Zhang, J. Zheng, C. Li, and L. Liu, Vector representation and its application of DNA sequences based on nucleotide triplet codons, *J. Mol. Graphics Modell.*, **62**, 150-156, 2015.
17. S.-A. Ouadfeul, Multifractal analysis of SARS-CoV-2 Coronavirus genomes using the wavelet transforms, bioRxiv Preprint: https://doi.org/10.1101/2020.08.15.252411.
18. M. Randic, M. Novic, and D. Plavsic. Milestones in graphical bioinformatics, *Int. J. Quantum Chem.*, **113**, 2413-2446, 2013.
19. S. Patodia, A. Bagaria, and D. Chopra, Molecular dynamics simulation of proteins: A brief overview. *J. Phys. Chem. Biochem.*, **4**, 1000166, 2014.
20. R.F. Voss, Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences, *Phys. Rev. Lett.*, **68**, 3805-3808, 1992.
21. G. Abramson, H.A. Cerdeira, and C. Bruschi, Fractal properties of DNA walks, *Biosystems*, **69**, 63-70, 491999, 1999.
22. C. Cattani, Fractals and hidden symmetries in DNA, *Math. Problems Eng.*, **12**, 507056, 2010.
23. B. Hao, H.T. Lee, and S. Zhang, Fractals related to long DNA sequences and complete genomes, *Chaos, Solitons and Fractals*, **11**, 825-836, 2000.
24. Z.-Y. Su, T. Wu, and S.-Y. Wang, Local scaling and multifractality spectrum analysis of DNA sequences-GenBank data analysis, *Chaos, Solitons and Fractals*, **40**, 1750-1765, 2009.

25. G. Durán-Meza, J. López-García, and J.L. del Río-Correa, The self-similarity properties and multifractal analysis of DNA sequences, Appl. Math. Nonlin. Sci., **4**, 267–278, 2019.
26. K.S. Birdi, *Fractals in Chemistry, Geochemistry, and Biophysics*. N.Y., Plenum Press, 1993.
27. T.G. Dewey, *Fractals in Molecular Biophysics*, Cambridge, Oxford University Press, 1997.
28. G.A. Kouzaev, Frequency dependence of microwave-assisted electron-transfer chemical reactions, *Mol. Phys*., **118**, e1685691, 2020.
29. S.V. Kapranov and G.A. Kouzaev, Nonlinear dynamics of dipoles in microwave electric field of a nanocoaxial tubular reactor, *Mol. Phys*., **117**, 489-506, 2019.
30. M.S. Swapna and S. Sankararaman, Fractal applications in bio-nanosystems, *Bioequiv. Availab*., **2**, OABB.000541, 2019.
31. X. Bin, E.H. Sargent, and S.O. Kelley, Nanostructuring of sensors determines the efficiency of biomolecular capture, *Anal. Chem.,* **82**, 5928–5931, 2010.
32. J. Chen, Z. Luob, C. Sunac, Z. Huang, C. Zhoua, S. Yin, Y. Duan, and Y. Li, Research progress of DNA walker and its recent applications in biosensor, *TrAC Trends in Anal. Chem.,* **120**, 115626, 2019.
33. A. Sadana, *Engineering Biosensors. Kinetics and Design Application*. San Diego, California, Acad. Press, 2001.
34. H. Namazi, V.V. Kulish, F. Delaviz, and A. Delaviz, Diagnosis of skin cancer by correlation and complexity analyses of damaged DNA, *Onkotarget*, **6**, 42623-42631, 2015.
35. B. Hewelt, H. Li, M.K. Jolly, P. Kulkarni, I. Mambetsariev, and R. Salgia, The DNA walk and its demonstration of deterministic chaos—relevance to genomic alterations in lung cancer, *Bioinformat.*, **35**, 2738–2748, 2019.
36. D.B. Searls, String variable grammar: a logic grammar formalism for the biological language of DNA, *J. Logic Programm.*, **24**, 73-102, 1995.
37. S. Ji, The linguistics of DNA: Words, sentences, grammar, phonetics, and semantics, *Ann. N. Y. Acad. Sci.*, **870**, 411-417,1999.
38. R. Carrascosa, F. Coste, M. Gallé, and G. Infante-Lopez, Searching for smallest grammars on large sequences and application to DNA, *J. Discr. Algorithms,* **11**, 62-72, 2011.
39. A.A. Green, J. Kim, D. Ma, P. Silver, J.J. Collins, and P. Yin, Complex cellular logic computation using ribocomputing devices, *Nature*, **548**, 117–121, 2017.
40. K.A. Dill, A. Lucas, J. Hockenmaier, L. Huang, D. Chiang, and A.K. Joshi, Computational linguistics: A new tool for exploring biopolymer structures and statistical mechanics, *Polymer*, **48**, 4289-4300, 2007.
41. S. Li, K.-L. Ng, and M.C. Chung, Quantitative linguistic study of DNA sequences, *Physica A*, **321**, 189-192, 2003.
42. H.S. Sichel. On a distribution law for word frequencies, *J. Am. Stat. Assoc*., **70**, 542-547, 1975.
43. *Chapter 2. General Structure. The Unicode Standard (6.0 ed.)*. Mountain View, California, US: The Unicode Consortium. ISBN 978-1-936213-01-6.
44. R.W. Hamming, Error detecting and error-correcting codes, *Bell Syst. Techn. J.*, **29**, 147–160, 1950.
45. B. Waggener, *Pulse Code Modulation Techniques*. Berlin-Heidelberg: Springer Verlag, 1995.

46. G. Navarro and M. Raffinot, *Flexible Pattern Matching in Strings: Practical Online Search Algorithms for Texts and Biological Sequences*. Cambridge: Cambridge University Press, 2002.
47. V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**, 707–710, 1966.
48. E. Gabidullin, Theory of codes with maximum rank distance, *Probl. Inf. Trans.*, **21**, 1-76, 1985.
49. E. Polityko, Calculation of distance between strings (https://www.mathworks.com/matlabcentral/fileexchange/17585-calculation-of-distance-between-strings, MATLAB Central File Exchange. Retrieved March 3, 2021.
50. V. Cooper, The coronavirus variants don't seem to be highly variable so far, *Sci. American*, 2021, March 24.
51. H. Abea, Y. Ushijimaa, M.M. Loembe, R. Bikangui, G. Nguema-Ondo, P.I. Mpingabo, V.R. Zadeh, C.M. Pemba, Y. Kurosaki, Y. Igasaki, S.G. deVries, M.P. Grobusch, S.T. Agnandji, B. Lell, and J. Yasuda, Re-emergence of Dengue virus serotype 3 infections in Gabon in 2016–2017, and evidence for the risk of repeated Dengue virus infections. *Int. J. Inf. Diseases*, **91**, 129-136, 2020.
52. J. Feder, *Fractals*. N.Y., Plenum Press, 1988.
53. P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, *Physica D*, **9**, 189-208, 1983.
54. S.N. Rasband, *Chaotic Dynamics of Nonlinear Systems*, Weinheim, J. Wiley & Sons, 1989.
55. B. Henry, N. Lovell, and F. Camacho, *Nonlinear Dynamics Time Series Analyses*, in Nonlinear Biomedical Signal Processing: Dynamic Analysis and Modeling. Edited by Akay M., IEEE, 2000, 1-39.
56. F. Roueff and J.L. Véhel, *A regularization approach to fractional dimension estimation.* In Proc. Fractals 98, Oct. 1998, Valletta, Malta. World Sci., 1998, 1-14.
57. J.L. Véhel and P. Legrand, *Signal and image processing with Fraclab,* In Thinking in Patterns. World Sci., 2003, 321-322.
58. G.A. Kouzaev: *Application of Advanced Electromagnetics. Components and Systems*. Berlin-Heidelberg: Springer, 2013.
59. D. Guidolin, C. Tortorella, R. De Caro, and L.F. Agnati, Does a self-similarity logic shape the organization of the nervous system? In The Fractal Geometry of the Brain. Edited by Di Leva A: Berlin-Heidelberg: Springer Verlag, 2016, 138-156.
60. FracLab 2.2. A fractal analysis toolbox for signal and image processing. https://project.inria.fr/fraclab/
61. J. Monge-Álvarez, Weierstrass Cosine Function (WCF) (https://www.mathworks.com/matlabcentral/fileexchange/50292-weierstrass-cosine-function-wcf), MATLAB Central File Exchange. Retrieved March 21, 2021.
62. M. Kim, H. Cho, S.-H.. Lee, W-J. Park, J.-M. Kim, J.-S. Moon, G.-W. Kim, W. Lee, H.-G. Jung, J.-S. Yang, J.-H. Choi, J.-Y. Lee, S.S. Kim, and J.-W. Oh, An infectious cDNA clone of a growth attenuated Korean isolate of MERS coronavirus KNIH002 in clade B, *Emerg. Microbes Infect*., **9**, 2714-2720, 2020.

Appendix 1. Results of statistical characterization of complete genetic sequencies of SARS CoV-2, MERS CoV, Dengue, and Ebola virus

Table 1. Severe acute respiratory syndrome coronavirus 2 (SARS Cov-2), *atg*-walk

	GenBank or GISAID Virus Name, Clade, Lineage, Registration Year, Sequencing Technology	Number of Nucleotides in the Sequence	Number of *atg*-Triplets in the sequence	Word Median Length	RMS Word Length	Fractal Dimension of the Length Distribution
	1	2	3	4	5	6
1	GenBank: *MN988668.1*, Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV WHU01, Wuhan, China, 2020, Illumina	29881	725	29	57.93	2.17
2	*hCoV-19/Japan/NGY-NNH-075/2021*, GR, B.1.1.64, Illumina MiSeq, Sanger	29848	722	29	58.03	2.17
3	hCoV-19/India/ILSGS00925/2021, G, B.1.617.2, Illumina NextSeq550	29782	723	28.05	57.77	2.16
4	hCoV-19/South Korea/KDCA3504/2021, GH, B.1.497, Illumina Miseq	29901	722	29	57.9602	2.17
5	*hCoV-19/Taiwan/TSGH-34/2020*, S, A.1, Illumina NovaSeq4000	29903	724	29	57.79	2.17
6	*hCoV-19/bat/Cambodia/RShSTT182/2010,* A.1, (bat virus), 2021, Illumina NextSeq	29787	730	29	55.81	2.17
7	hCoV-19/Austria/CeMM3224/2021, GR, B.1.1.244, Illumina NovaSeq	29782	721	30	59.03	2.16
8	*hCoV-19/England/205341113/2020,* GV, B.1.177, Illumina NextSeq	29862	721	29	57.97	2.17
9	*hCoV-19/Ireland/D-NVRL-e84IRL94434/2021*, GV, B.1.177, Illumina	29523	719	29	59.56	2.17
10	*hCoV-19/Netherlands/UT-RIVM-13868/2021*, GH, B. 1.160, Nanopore MinION	29782	720	28	58.17	2.16
11	*hCoV-19/Norway/0179/2021,* GH, B.1.36, 2021, Nanopore GridIon	29782	723	28	57.88	2.15
12	*hCoV-19/Russia/IVA-CRIE-L188N0202/2021,* GR, B.1.1.317, Illumina	29735	720	29	57.77	2.17
13	*hCoV-19/Spain/RI-IBV-99016064/2021*, GV, B.1.221, Illumina MiSeq	29865	719	29	59.56	2.17
14	*hCoV-19/Brazil/RS-00674HM_LMM52649/2020*, GR, B.1.1.33, Illumina Miseq	29867	719	29	58.31	2.17
15	*hCoV-19/Canada/ON-S2383/2021,* GH, B. 1.36.38, Illumina MiniSeq	29830	722	29	57.89	2.16
16	*hCoV-19/Mexico/CMX-INER-0222/2020,* G, B.1.551, Illumina NextSeq	29885	724	29	57.83	2.17
17	*hCoV-19/USA/TX-HHD-2102044112/2021*, GR, B.1.1.244, Illumina MiSeq	29819	720	29	58.10	2.17
18	*hCoV-19/USA/CA-LACPHL-AF00513/2021,* GH, B.1.429, Illumina MiSeq	29844	723	29	57.86	2.17
19	*hCoV-19/South Africa/KRISP-K004540/2020*, GR, B.1.1.56, Illumina MiSeq	29851	722	29	57.90	2.17

Table 2. Middle East respiratory syndrome-related coronavirus (MERS), *atg*-walk

	GenBank Virus Name and Accession Number, Registration Year, Sequencing Technology	Nucleotides Number	Number of ATG-Triplets	Word Median Length	RMS Word Length	Fractal Regularization Dimension of the Length Distribution
	1	2	3	4	5	6
1	*MF598617.1*, Middle East respiratory syndrome-related coronavirus strain camel/UAE_B25_2015, United Arabian Emirates, AE, 2017, Illumina; Sanger dideoxy sequencing	30123	712	30	58.8	2.30
2	*MF598595.1,* Middle East respiratory syndrome-related coronavirus strain camel/UAE_B2_2015, United Arabian Emirates, 2017, Illumina; Sanger dideoxy	30123	709	30	59.04	2.30
3	*NC-019843.3*, Middle East respiratory syndrome-related coronavirus isolate HCoV-EMC/2012, Saudi Arabia, 2020, Sanger dideoxy	30119	717	30	58.48	2.30
4	*KY673148.1,* Middle East respiratory syndrome-related coronavirus strain Hu/Oman_50_2015, 2017, Sanger dideoxy	30123	714	29	58.74	2.30
5	*KT225476.2*, Middle East respiratory syndrome coronavirus isolate MERS-CoV/THA/CU/17_06_2015, Oman/Thailand, 2017, Sanger dideoxy	29809	703	30	59.03	2.25
6	*MG923479.1*, Middle East respiratory syndrome-related coronavirus isolate MERS-CoV camel/Nigeria/NV1712/2016, 2018, Sanger dideoxy	29455	701	30	58.08	2.24
7	*MK967708.1,* Middle East respiratory syndrome-related coronavirus isolate Merscov/Egypt/Camel/AHRI-FAO-*1/2018*, 2019, CLC genomic workbench	30106	711	30	58.05	2.30
8	*MT361640.1*, Mutant Middle East respiratory syndrome-related coronavirus clone MERS-CoV YKC, South Korea, 2021, sequencing technology is described in [62]	30136	710	30	58.90	2.30
9	*KT326819.1*, Middle East respiratory syndrome coronavirus strain MERS-CoV/KOR/KNIH/001_05_2015, South Korea, 2017, Illumina and Sanger dideoxy	29995	711	30	58.86	2.30
10	*MK129253.1*, Middle East respiratory syndrome-related coronavirus isolate MERS-CoV/KOR/KCDC/001_2018-TSVi, South Korea, 2019, Sanger dideoxy	30150	712	30	58.81	2.29

Table 3. Dengue Virus, *atg*-walk

#	GenBank Virus Name, Registration Year, Sequencing Technology	Nucleotides Number	Number of ATG-Triplets	Word Median Length	RMS Word Length	Fractal Regularization Dimension of the Length Distribution
	1	2	3	4	5	6
1.1	*KY672944.1*, Dengue virus 1 isolate DENV-1/China/YN/YNH22 (2013), 2019, Sanger dideoxy	10709	299	23	47.74	2.36
1.2	*KY672937.1*, Dengue virus 1 isolate DENV-1/China/YN/DGRL-6(2014), 2019, Sanger dideoxy	10738	294	23	50.02	2.33
1.3	*MW386865.1*, Dengue virus 1 isolate YNBN04, China, 2020, Sanger dideoxy	10742	289	24	50.81	2.36
1.4	*MG560269.1*, Dengue virus 1 isolate P1253/China/GD/CZ/2014, 2018, Sanger dideoxy	10583	298	23	47.55	2.35
1.5	*MG560267.1,* Dengue virus 1 isolate P1258/China/GD/CZ/2014, 2018, Sanger dideoxy	10583	299	23	47.22	2.35
2.1	*MN566112.1,* Dengue virus 2 isolate New Caledonia-2018-AVS127, 2020, Illumina	10722	267	32	52.24	2.4
2.2	*KY672955.1*, Dengue virus 2 isolate DENV-2/China/YN/15DGR65(2015), 2019, Sanger dideoxy	10723	273	28	52.77	2.44
2.3	*KY672954.1*, Dengue virus 2 isolate DENV-2/China/YN/JH1516(2015), 2019, Sanger dideoxy	10665	271	29	51.50	2.48
2.4	*MK268692.1,* Dengue virus 2 isolate DENV-2/TH/1974, Thailand, 2019, Sanger dideoxy	10721	274	28	52.67	2.45
2.5	*MH069499.1,* Dengue virus 2 strain DENV-2/VE/IDAMS/910105, Venezuela, 2018, Illumina	10712	275	28	52.84	2.49
3.1	*MN018389.1*, Dengue virus 3 isolate D17011, China, 2020, Sanger dideoxy sequencing	10708	272	28	55.46	2.57
3.2	*NC_001475.2*, Dengue virus 3, Sri Lanka, 2019, Illumina	10707	273	27	55.05	2.58
3.3	*KY863456 .1*, Dengue virus 3 isolate 201610225, Indonesia, 2017, IonTorrent, Sanger dideoxy sequencing	10707	278	28	52.84	2.5
3.4	*MH544649.1*, Dengue virus 3 isolate 449686_Antioquia_CO_2015, Colombia, 2018, Illumina; Sanger dideoxy sequencing	10707	273	28	52.84	2.49
3.5	*MH823209.1*, Dengue virus 3 isolate SMD-031, Indonesia, 2019, Illumina	10707	272	28	52.84	2.46
3.6	*LC379197.1*, Dengue virus 3 strain SYMAV-17/Gabon/2017 genomic RNA, 2019, Illumina	10641	271	29	52.85	2.06
3.7	*KY921907.1*, Dengue virus 3 isolate SG(EHI)D3/15095Y15, 2017, Singapore, Sanger dideoxy sequencing	10667	266	29	53.58	2.09
3.8	*KF041255.1,* Dengue virus 3 isolate D3/Pakistan/55505/2007, 2013,	10675	268	29	53.55	2.07

	Sanger dideoxy sequencing					
3.9	*LC379196.1,* Dengue virus 3 strain SYMAV-09/Gabon/2016 genomic RNA, 2019, Illumina	10663	273	29	53.64	2.53
3.10	*LC379195.1,* Dengue virus 3 strain SYMAV-07/Gabon/2016 genomic RNA, 2019, Illumina	10663	273	29	53.64	2.53
4.1	*KJ579245.1*, Dengue virus 4 strain DENV-4/MT/BR23_TVP17909/2012, Brazil, 2020, Illumina	10649	273	26	53.12	2.09
4.2	*MG272274.1*, Dengue virus 4 isolate D4/IND/PUNE/IRSHA-FG-03 (S-49), complete genome, India, 2018, Ion Proton System	10652	270	27	53.07	2.07
4.3	*KY672960.1*, Dengue virus 4 isolate DENV-4/China/YN/15DGR394 (2015), 2019, Sanger dideoxy	10661	276	26	52.63	2.09
4.4	*KX224312.2,* Dengue virus 4 isolate SG(EHI)D4/02990Y14, Singapore, 2017, Sanger dideoxy sequencing	10652	275	27	52.21	2.09
4.5	*MG272272.1,* Dengue virus 4 isolate D4/IND/PUNE/IRSHA-FG-01 (1028), India, 2018, Ion Proton System	10652	272	27	54.57	2.09

Table 4. Ebola Virus (GenBank), *atg*-walk

	Genbank Virus Name, Registration Year, Sequencing Technology	Nucleotide Number	Number of ATG-Triplets	Word Median Length	RMS Word Length	Fractal Dimension of the Length Distribution
	1	2	3	4	5	6
1	*MG572235.1*, Zaire ebolavirus isolate Ebola virus/H.sapiens-tc/COD/1995/Kikwit-9510621, Zaire, 2019, PacBio; Illumina	18957	329	40.5	85.29	2.03
2	*KU174137.1*, Mutant Zaire ebolavirus isolate Ebola virus/H.sapiens-rec/COD/1976/Yambuku-Mayinga-eGFP-BDBV_GP, Zaire, 2019, Illumina	19774	339	41.5	84.02	2.06
3	*KY786025.1,* Ebola virus strain Ebola virus/M.fascicularis-wt/GAB/2001/untreated-CCL053D5, Gabon, 2018, IonTorrent	18871	327	40.5	85.06	2.03
4	*KY785936.1*, Ebola virus strain Ebola virus/M.fascicularis-wt/GAB/2001/100mg-CA470D5, Gabon, 2018, IonTorrent	18871	327	40.5	85.06	2.03
5	*MH121162.1,* Sudan ebolavirus isolate Ebola virus/H.sapiens-tc/Sudan/1976/Boniface-R4142L, 2019, Illumina	18831	345	37.5	77.79	2.2
6	*MK952150.1,* Sudan ebolavirus isolate Ebola virus/H.sapiens-wt/SSD/1976/Maridi-BNI/DT, South Sudan, 2020, Illumina	18847	345	37	77.84	2.2
7	*MH121169.1,* Sudan ebolavirus isolate Ebola virus/H.sapiens-tc/Sudan/2004/Yambio-HCM/SAV/017, 2019, PacBio; Illumina	18849	345	37	77.84	2.2
8	*NC_039345.1*, Bombali ebolavirus isolate Bombali ebolavirus/Mops condylurus/SLE/2016/PREDICT_SLAB000156, Sierra Leone, 2018, Sanger dideoxy, Illumina	19043	325	36	84.24	2.38
9	*MW056492.1*, Bombali ebolavirus isolate X030, Kenya, 2020, Illumina	19025	326	36	84.11	2.37
10	*MF319186.1*, Bombali ebolavirus isolate Bombali virus/C.pumilus-wt/SLE/2016/Northern Province-PREDICT_SLAB000047, Sierra Leone, 2019, Sanger dideoxy	19043	324	36	85.27	2.36
11	*MK340750.1*, Bombali ebolavirus isolate B241, Kenya, 2019, Illumina	19025	328	36	83.59	2.36
12	*MW056493.1*, Bombali ebolavirus isolate Z153, Kenya, 2020, Illumina	19025	332	36	81.46	2.41
13	*MK028856.1,* Bundibugyo ebolavirus isolate Ebola virus/H.sapiens-tc/Uganda/2007/Bundibugyo-200706291, 2019, PacBio; Illumina	18940	325	39	84.62	2.01
14	*MK028834.1,* Bundibugyo ebolavirus isolate Ebola virus/H.sapiens-tc/Uganda/2007/Bundibugyo-R4386L, Uganda, 2019, PacBio; Illumina	18917	325	39	84.62	2.01
15	*MK028835.1*, Bundibugyo ebolavirus isolate Ebola virus/H.sapiens-tc/Uganda/2007/Bundibugyo-200706291, 2019, PacBio; Illumina	18936	325	39	84.63	2.01

Table 5. Severe acute respiratory syndrome coronavirus 2 (SARS Cov-2), *c*-walk

	GenBank or GISAID Virus Name, Lineage, Registration Year, Sequencing Technology	Number of Nucleotides in the Sequence	Number of 'c's in the sequence	Inter-'c' distance Median Length	RMS of Inter-'c' Distance Length	Fractal Dimension of the Inter-'c' Distance Distribution
	1	2	3	4	5	6
1	GenBank: *MN988668.1*, Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV WHU01, Wuhan, China, 2020, Illumina	29881	5492	4	7.30	2.25
2	GISAID: *hCoV-19/Japan/NGY-NNH-075/2021*, B.1.1.64, Illumina MiSeq, Sanger	29848	5486	4	7.30	2.31
3	*hCoV-19/India/ILSGS00925/2021*, G, B.1.617.2, Illumina NextSeq550	29782	5459	4	7.32	2.34
4	*hCoV-19/South Korea/KDCA3504/2021*, GH, B.1.497, Illumina Miseq	29901	5486	4	7.31	2.22
5	*hCoV-19/Taiwan/TSGH-34/2020*, S, A.1, Illumina NovaSeq4000	29903	5491	4	7.30	2.25
6	*hCoV-19/bat/Cambodia/RShSTT182/2010*, A.1, 2021, Illumina NextSeq	29787	5471	4	7.32	2.31
7	GISAID: *hCoV-19/Austria/CeMM3224/2021*, B.1.1.244, Illumina NovaSeq	29782	5465	4	7.32	2.34
8	GISAID: *hCoV-19/England/205341113/2020,* B.1.177*,* Illumina NextSeq	29862	5477	4	7.32	2.31
9	*hCoV-19/Ireland/D-NVRL-e84IRL94434/2021*, B.1.177, Illumina	29523	5394	4	7.69	2.35
10	*hCoV-19/Netherlands/UT-RIVM-13868/2021*, B. 1.160, Nanopore MinION	29782	5464	4	7.32	2.34
11	GISAID: *hCoV-19/Norway/0179/2021,* B.1.36, 2021, Nanopore GridIon	29782	5453	4	7.33	2.34
12	hCo*V-19/Russia/IVA-CRIE-L188N0202/2021*, GR, B.1.1.317, Illumina	29735	5465	4	7.30	2.30
13	*hCoV-19/Spain/RI-IBV-99016064/2021*, B.1.221, Ilumina MiSeq	29865	5434	4	7.75	2.32
14	*hCoV-19/Brazil/RS-00674HM_LMM52649/2020*, GR, B.1.1.33, Illumina Miseq	29867	5467	4	7.34	2.34
15	*hCoV-19/Canada/ON-S2383/2021,* GH, B. 1.36.38, Illumina MiniSeq	29830	5478	4	7.31	2.29
16	GISAID: *hCoV-19/Mexico/CMX-INER-0222/2020,* B.1, Illumina NextSeq	29885	5483	4	7.31	2.25
17	*GISAID: hCoV-19/USA/TX-HHD-2102044112/2021*, B.1.1.244, Illumina MiSeq	29819	5472	4	7.32	2.31
18	*hCoV-19/USA/CA-LACPHL-AF00513/2021*, GH, B.1.429, Illumina MiSeq	29844	5475	4	7.32	2.31
19	*hCoV-19/South Africa/KRISP-K004540/2020*, GR, B.1.1.56, Illumina MiSeq	29851	5486	4	7.31	2.22