Visual and Quantitative Analyses of Virus Genomic Sequences using a Metric-based Algorithm

ALEXANDRA BELINSKY
Mach-3dP Inc.
Burlington, ON, CANADA

GUENNADI A. KOUZAEV
Norwegian University of Science and Technology - NTNU
Trondheim, NORWAY

Abstract: - This work aims to study the virus RNAs using a novel accelerated algorithm to explore any-length repetitive genomic fragments in sequences using Hamming distance between the binary-expressed characters of an RNA and a query pattern. Primary attention is paid to the building and analyzing 1-D distributions (walks) of *atg*-patterns - codon-starting triplets in genomes. These triplets compose a distributed set called a word scheme of RNA. A complete genome map is built by plotting the mentioned *atg*-walks, trajectories of separate (*a*-, *c*-, *g*-, and *t*-symbols) nucleotides, and the lines designating the genomic words. The said map can be additionally equipped by gene's designations making this tool pertinent for multi-scale genomic analyses. The visual examination of *atg*-walks is followed by calculating statistical parameters of genomic sequences, including estimating walk-geometry deviation of RNAs and fractal properties of word-length distributions. This approach is applied to the SARS CoV-2, MERS CoV, Dengue, and Ebola viruses, whose complete genomic sequences are taken from GenBank and GISAID. The relative stability of these walks for SARS CoV-2 and MERS CoV viruses was found, unlike the Dengue and Ebola distributions that showed an increased deviation of their geometrical and fractal characteristics. The developed approach can be useful in further studying mutations of viruses and building their phylogenic trees.

Key-Words: - RNA, Hamming-distance metric measure, quantitative RNA walks, SARS Cov-2 virus, MERS CoV virus, Dengue virus, Ebola virus

Received: April 19, 2022. Revised: November 5, 2022. Accepted: November 24, 2022. Published: December 31, 2022.

1 Introduction

A virus is a tiny semi-live unit carrying genetic material (RNA or DNA – double-helix RNA structure) in a protein capsid covered by a lipid coat. The virus penetrates the cell wall and urges this bio-machine to 'manufacture' more viruses.

Some viruses are RNA-based and transfer genetic information by long chains of four organic acids, namely, Adenine (*a*), Cytosine (*c*), Guanine (*g*), and Uracil (*u*) [1]. DNA-based viruses and double-stranded genetic polymers carry the information by four nucleotides, but one of them is Thymine (*t*) instead of Uracil. In genomic databases, even the single-stranded viral RNAs are registered as complementary chains where Thymine substitutes Uracil due to some instrumental specifics [2] that do not hinder the mathematical aspects of the virus theories. These complimentary RNAs will be used further for numerical modeling in our paper.

A complete RNA is a chain of codons (exons) used to transfer genetic information and introns. Unfortunately, the latter's role is not well known [3]. Sequencing of RNA or DNA is

searching and identifying nucleotides by instrumental means. Codons in RNAs start with an '*aug*' combination of nucleotides and end with one of the following three combinations: '*uaa*', '*uag*', or '*uga*'. Additionally, '*aug*' may play a coding role in genomes. Some DNA strands consist of billions of nucleotides, so mathematical methods are widely used in genomics [4]. For instance, the RNA symbols are substituted by number values, and this process is called DNA/RNA mapping [5]-[8]. For example, in Ref. [5], eleven methods of numerical representation of genomic sequences are listed and analyzed to conclude that each is preferable in a particular application, and no universal mapping algorithm is equally advantageous for all genomic studies.

Different retrieval algorithms can be applied to genomic sequences, including signal-processing means [7],[9]-[14]. The numerical RNAs can be shown graphically for qualitative analyses. For instance, each nucleotide is represented by a unit vector in a 4-dimensional (4-D) space, and an imaginary walker moves along an RNA sequence, making a trajectory in this space that can be described by fractional order differential or integrodifferential equations [15],[16].

The nucleotides are combined in a certain way to avoid apparent difficulties with plotting walks in multidimensional spaces [17],[18]. For instance, each nucleotide is associated with one of four unit vectors in 2-D space, which projections on the plane axes can take positive (+1) or negative (–1) values [19],[20]. A trajectory is built moving along the consecutive number of a nucleotide in the studied genomic sequence.

In general, DNA walks allow the detection of codons and introns, discover hidden RNA periodicity [12]-[14], and calculate phylogenetic distances between genomic sequences [21], among others. Some additional results and reviews on DNA imaging can be found, for instance, in Refs. [17],[22]-[24], where the necessity to use specified walks for each class of genomic problems is shown.

Genomic walk analysis can be followed by calculating fractal properties of distributions of nucleotides [25]-[34]. Fractals are self-similar or scale-invariant objects. It means small 'sub-chain' geometry is repeated on larger geometry scales, although it is randomly distorted.

A biopolymer chain in a solution is bent fractally [28]. This fractality influences the chemical reaction rate, diffusion, and surface absorption of long-chain and globular molecules, among others [27],[28],[35]-[38]. Because polar solvents have frequency-dependent properties, they are adjusted by microwaves that influence the polymer fractal dimension. Thus, some bioreactions can be controlled by a weak high-gradient microwave field [39],[40].

Although many achievements are known in the numerical mapping of RNAs, some questions have not been resolved. For instance, the known genomic walks are designed to track single nucleotides or their pairs, leading to crowdy trajectories and overloaded plots that are challenging to analyze visually in one plot [5]-[8].

Meanwhile, the complete RNAs of viruses are composed chiefly of codons, and one repetitive pattern therein is their starting *atg*-triplet. We assign to these triplets a mathematical construction - a viral RNA scheme.

Our proposed pattern search algorithm calculates the triplet distributions along an RNA sequence. Additionally, the same algorithm can make the walks of each of the four nucleotides. These trajectories are found not twisted firmly in comparison to curves from Refs. [5],[8], for instance, and they are easier to be analyzed visually. These graphs can be equipped with marks pointing to genes and hyperlinks with the gene names, making these figures interactive means to analyze the genomes.

Some results of creating such a tool and applications to genomes of several viruses are given here. The source codes and visualizations

can be used for research and practical applications. One of them is the studies of the stability of the mentioned *atg*-schemes towards mutations, variation of codon fillings, and the fractality of *atg*-distributions, among others.

Section 2 considers our proposed calculation algorithms and plotting techniques in detail with application examples. The results of applying these techniques to the SARS CoV-2, MERS CoV, Dengue, and Ebola viruses are in Section 3. They are further discussed in detail in Section 4, and the conclusions are in Section 5. In Appendix 1, all necessary data for the analyzed RNAs taken from GenBank and GISAID are tabular, including the parameters calculated in this contribution.

2 Materials and Methods

2.1 Materials and Data Availability

In this paper, only *complete arbitrary-chosen* genomic sequences without missed symbols are studied taken from GenBank [41] and GISAID [42]. Among them are 36 SARS CoV-2 genomic sequences from GISAID and one from GenBank, 20 genomic sequences for the MERS coronavirus (GenBank), 25 for the Dengue virus from GenBank, and 15 Ebola virus genomic sequences (GenBank). Data from the GISAID are available after registration. All names of genomic sequences are given in figure legends and Tables 1–4.

2.2 Methods and Application Examples

2.2.1 Metric-based *atg*-triplet Algorithm

As has been stated above, for both DNA and RNA descriptions by characters, their alphabet consists of four nucleotides marked by *a*-, *c*-, *g*-, and *t*-symbols. These designations are used to study RNA or DNA if their physicochemical properties are outside the research scope.

In many cases, RNA/DNA have repeated patterns of nucleotide sequences, and these regions are better conserved in mutations [9], [43],[44]. As a rule, pattern discovery relates to nondeterministic polynomial time problems (NP-problems), i.e., solution time increases exponentially with the sequence length.

A typical algorithm compares a query character pattern with a length of nucleotides with the following one-symbol shift of query along a chain. In our code, we use these techniques, as well. Usually, the search algorithms working with characters with no assigned numerical values are slower by 1.43-2.37 times than those processing the binary variables, as shown in Refs. [45],[46]. Moreover, even the computer arithmetic logic units can be re-designed to fulfill the frequently repeated patterns of binary operations to accelerate them [47]. Then, in our code, all RNA's characters are transformed into binaries before all operations using a Matlab operator dec2bin(*character*) [48].

In computers, for instance, the UTF-8 format allows encoding all 1,112,064 valid character code points, and it is widely used for the World Wide Web [49]. The first 128 characters (US-ASCII) require only one byte (eight binary numbers) in this format. If binary units initially represent the sequences, then calculating the DNA sequence's numerical properties in this form reduces the computation time.

Because the binary sequences now write the DNA/RNA chains, they can be characterized quantitatively using a suitable technique. One calculates a metric distance between the binary-represented symbols and a query (base) 'moving' along a chain.

Many metric types are used in codes and big data [50]-[55]. The advantage of using metric estimates is that they can be applied in cluster analysis for similar grouping nucleotide or protein distribution patterns. For instance, it can help classify the virus RNAs [56],[57]. Notably, this distance can be the Hamming one [50] used further in this paper.

Consider a flow chart of our code on exploring patterns of arbitrary length n (Figure 1). It starts with importing sequence data A of the length N from any genomic database in FASTA format [41],[42] and defining a query pattern B of any size n. From both files, the empty spots should be removed [58].

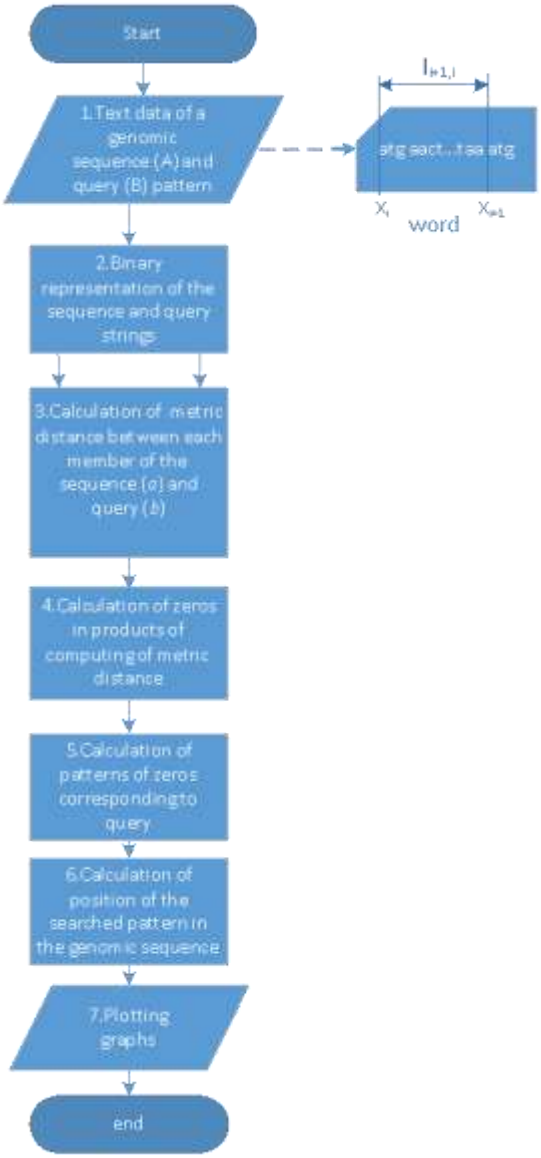

Fig. 1: Algorithm flowchart.

In the second step, the data files are transformed into binary strings, which are used to calculate Hamming distance between each binary symbol of A and B. This distance is a metric for comparing two binary values, and it is the number of bit positions in which the two bits are different.

To calculate Hamming distance d_H between two strings A and B, the XOR operation $(A \oplus B)$ is used, and the total number of '1's in the resultant string C is counted. The distance value is zero if the compared binary-represented symbols are the same. Only n characters are compared on each count; then, the query is moved to one symbol, and calculations are repeated.

Then, a string C of numerical estimates of the length $M = n \times N$ is a product of step 3 of this code. Not all registered genomic sequences are divisible by n. In this case, a needed number of characters a is added to the end of the sequence A, or the Hamming metric operation can be fulfilled using Levenstein's distance formula from Ref. [53], workable for compared sub-strings of arbitrary length.

The following two parts of our algorithm are calculating numbered (y_i) query positions in a sequence x_i according to the Hamming-distance data. All zeros in the string C are initially obtained (step 4, Figure 1). Then, only n neighboring zeros corresponding to a query are selected (step 5, Figure 1), and this query is numbered in a sequential manner starting with the first one found in RNA. The positions x_i of these numbered queries y_i in a complete RNA sequence are calculated analytically (step 6, Figure 1).

Let us take the coordinate of the first symbol in a numbered found query, then a set of points can be built along a studied sequence. These points, being connected, make a curve called a query walk.

In this paper, the codon start-up *atg*-triplet is used below as a query (pattern B). We define the positions x_i of the first symbols of the sequentially numbered *atg*-triplets in an RNA A, and the *atg*-walks are plotted.

Additionally, we calculate the word length $l_{i,i+1}^{(atg)} = x_{i+1} - x_i$. In our algorithm, a 'word' is a nucleotide sequence starting with an *atg*-triplet and all symbols up to the next *atg*-one (Figure 1, right side).

The proposed algorithm was realized in the Matlab environment [48], and it is a few-ten-line code. The following Matlab library functions were used:

1. *dec*2*bin*(character variable) – to transform a character variable into a binary one
2. *ptisd*2$(a, b,$ '*hamming*'$)$ – to calculate

the Hamming distance value between two binary values a and b

3. $zeros(\text{string})$ – calculation of numbers of zero-values in a string
4. $plot(y,x)$ – plot function $y(x)$

The developed algorithm was applied to many available virus complete genomes. Because the *atg*-triplets start the codons in most cases, the main attention was paid to building the distributions of these repeated patterns called the RNA word schemes. The *atg*-positions were compared with the found ones from the given genomes to verify these calculations.

2.2.2 Visualization Techniques

2.2.2.1 One-dimensional *atg*-walks

The viral RNAs, consisting of thousands of nucleotides, are challenging to analyze, and many visualizing methods are used. Among these techniques is the plotting the DNA walks projected on the spaces of appropriated dimensions considered above (see Introduction). Some contributions are full of symbolic designations of nucleotides and diagrams showing the positions of genes in the complete DNA sequences. The preference for a visualization method is dictated by the specificity of applications, although there is a need for a universal graphical tool.

This paper found that *atg*-walks could be plotted by coordinates of the numbered *atg*-triplets in a complete RNA sequence. Like the known DNA walks, these dots can be considered the points of a trajectory named here as *atg*-walk (Figure 2A). A diagram showing the positions of the symbol a in defined *atg*-triplets is given in Figure 2B.

Similar to single-symbol DNA walks, the *atg*-trajectories have fractal properties. Their type was defined by analyzing distributions of the coordinates of triplets shown by vertical lines along an RNA sequence (Figure 2B). These *atg*-distributions have repeating motifs on different geometry scale levels, i.e., they have fractal properties.

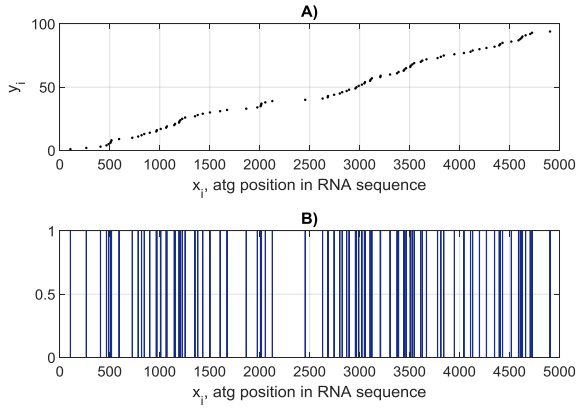

Fig. 2: Positions of *atg*-triplets along the genome sequence of SARS-CoV-2 virus MN988668.1 (row 1, Table 1, Appendix 1) given for the first 5 000 nucleotides provided by points (A) and vertical lines in diagram (B).

Below, our initial assumption about the fractality of *atg*-distributions will be 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 set rule [59].

2.2.2.2. Multi-scale Mapping of RNA Sequences

It is necessary to see full-scale virus RNA maps and analyze all types of mutations. Previously, the most attention was paid to mapping *atg*-triplets, thinking they constructed a scheme of RNA, a relatively stable structure. Besides the structural mutations changing the *atg*-distributions, the nucleotides vary their positions inside codons. Our algorithm considers even a single symbol as a pattern, allowing the calculation of distribution curves for each nucleotide similar to *atg*-ones. These curves can be considered the first level of spatial detailing of RNAs. The words in our definitions (see Figure 1) compose the second level. Words form a gene responsible for synthesizing several proteins, and the genes belong to the third level of spatial detailing of RNAs.

A combined plotting of elements of the hierarchical RNAs organization will be helpful in the visual analyses of genomes. One of the ways is shown in Figure 3.

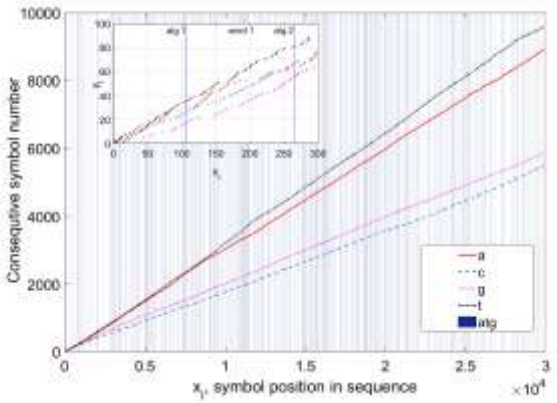

Fig. 3: Two-scale study results of a SARS-CoV-2 virus MN988668.1 (row 1, Table 1, Appendix 1). In the inlet, these symbols are pointed inside the words given for the first 300 nucleotides.

Here, positions of *a*-symbols of *atg*-triplets in a sequence are given by vertical lines (second level of detailing). Words take spaces between these vertical lines (see Figure 1). They are filled by numbered nucleotides (points of a different color), which are the first level of RNA detailing. It allows distinguishing nucleotides even at the beginning of coordinates where visual clutter is seen (inlet in Figure 3).

The next level of hierarchical RNAs organization is with genes. For instance, in GenBank [41], the list of symbols of a genomic sequence in FASTA format is followed by a diagram where the genes are given by horizontal bars with the gene's literal designations.

In our case, this diagram can be attached to a two-scale plot considered above. Another solution is to equip our figures with gene hyperlinks, an interactive means highlighting whose genes a nucleotide or codon belongs to, shown by a pointer.

Thus, the developed pattern search algorithm based on the Hamming distance applied to binary representations of nucleotide symbols allows building combined plotting of hierarchical organization of the RNAs of viruses. It can also be applied to the analyses of more complex protein structures. Different from many genomic walks, it produces spatially simple curves that can be analyzed visually and quantitatively.

2.2.3 Calculation Techniques for Fractal Properties of *atg*-Distributions

In many previous studies, the fractality of DNA/RNA sequences has been studied [6],[20]-[26],[29]-[38]. The motifs of small-size genetic 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, as is mentioned in many papers.

The large-size genomic data are often patterned, and each pattern can have its fractal dimension, i.e., the sequences can be multifractals [33]. This effect is typical in genomics, but it is also common in the theory of nonlinear dynamical systems, signal processing, and brain tissue morphology, among others [60]-[68].

Discovering the fractality of genomic sequences is preceded by their numerical representation, for instance, by walks of different types [6]-[8],[20],[22],[29],[33]. Then, each step value of a chosen walk is considered a sample of a continuous function, and the methods of signal processing theory are applied [12],[13].

The measure of self-similarity is its fractal dimension d_F which can be calculated using different methods. In our case, the fractal dimension calculations can be applied directly to the distribution of *atg*-consecutive numbers y_i (Figure 2a), but it gives this value close to 1 for the analyzed RNAs, i.e., the dependence $y_i(x_i)$ is close to the linear one. Then, these calculations are not effective in analyses due to their weak sensitivity. Instead, the word-length distributions along RNAs sequences $l_{i,i+1}^{(atg)}$ (see sub-section 2.2.1 and Figure 1, right side) are used.

A particular distribution of the word lengths is shown in Figure 4 by bars whose heights equal the word lengths. Then, the algorithms, usually applied to the sampled signals, can be used to compute the statistical properties of these word-length distributions.

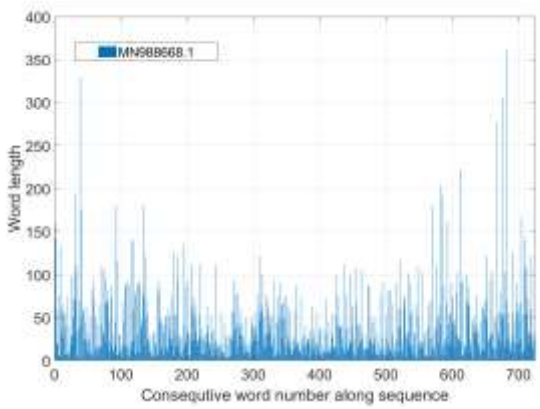

Fig. 4: Word-length distribution in a SARS Cov-2 virus MN988668.1 sequence (row 1, Table 1, Appendix 1).

In this paper, the fractal dimension of word length distributions was calculated using the software package FracLab 2.2 [67]. This code provides results with reasonable accuracy if the default parameters of FracLab are used. Although many researchers tested this code, it is again verified to calculate the Weierstrass function, which is synthesized according to a given fractal dimension value [68].

In a strong sense, the fractal dimension was defined for the infinite sequences. In our case, the studied RNAs have only 268-730 *atg*-triplets depending on the virus. Then, the fractal dimension values were estimated approximately.

3 Results

3.1. Study of *atg*-walks of SARS CoV-2 Genome Sequences

In this research, essential attention was paid to studying SARS CoV-2 complete RNA genome sequences. A recent comprehensive review of the genomics of this virus can be found in Refs. [69],[70], for instance.

The data used here and throughout this whole paper are from two genetic databases: GenBank [41] and GISAID [42]. A part of the studied genome sequences for this and other viruses is provided in Appendix 1.

Here, the main unit, called a 'word', is a nucleotide sequence starting with '*atg*' and the symbols up to the next triplet (Figure 1). The number of *atg*-s was calculated by our code and verified by a Matlab function $count(A, 'atg')$. These results are shown in the third column of Tables 1–4 (See Appendix 1). The Matlab functions $median(L_{\text{word}})$ and $rms(L_{\text{word}})$ calculated 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 in columns 4 and 5 of the mentioned tables.

Consider applying the developed approach to the complete genome of a Wuhan RNA sample MN988668.1 (GenBank) as an example. It consists of 29881 nucleotides and 725 *atg*-triplets (See row 1, Table 1, Appendix 1). Figure 2 shows the distribution by points of *atg*-triplets for the first 5000 nucleotides of this complete genome. In Figure 3, the entire study of this virus is shown. Word-length distribution of this sample is given in Figure 4.

Figure 5 illustrates the distribution (in lines) of *atg*-triplets along with complete genome sequences for thirty-seven SARS CoV-2 arbitrary-chosen virus samples, including Delta, Omicron, and a bat-corona sample (see Table 1, Appendix 1).

There were relatively compact localizations of the triplet curves despite the viruses being of different clades and lines. For instance, the relative difference $\delta y(x_i = 29455) = 100\% \cdot 2\Delta y_{1,21} / (y_1 + y_{21})$ of these curves 1-37 (Figure 5) is estimated at $x_i = 29455$ around only 1.6%. This confirms the conclusions of many specialists [71] that no new recombined strains have appeared up to this moment, despite many mutations found to date (2023), including the Omicron lineage.

Two inlets show the beginning and tails of these curves to illustrate details. Although, in general, these trajectories are woven firmly, the tails are between the bat's SARS-CoV-2 light-blue curve (*hCoV-19/bat/Cambodia/RShSTT182/2010*, row 6, Table 1, Appendix 1) and the black trajectory obtained for a sequence from Brazil (*hCoV-19/Brazil/RS-00674HM_LMM52649/2020*, row 14, Table 1, Appendix 1).

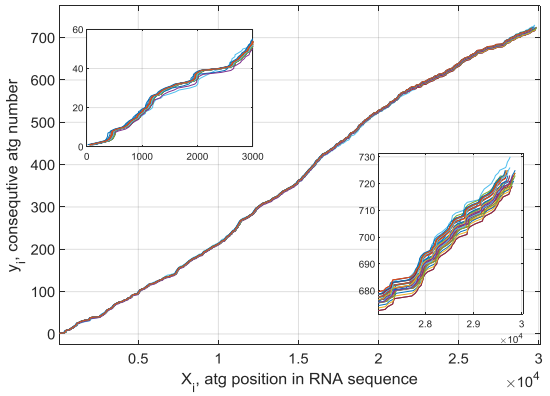

Fig. 5: Distributions of *atg*-triplets of 31 SARS Cov-2 complete RNA sequences (Table 1, Appendix 1). The inlets show these *atg*-distributions at the beginning and end of genome sequences.

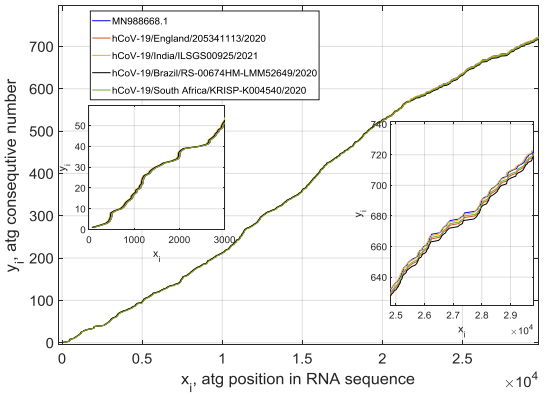

Fig. 6: Detailed distributions of *atg*-triplets for five trouble-making SARS Cov-2 complete RNA sequences (rows 1,8,3,14,19, Table 1, Appendix 1). The inlets show these *atg*-distributions at the beginning and end of genome sequences.

A detailed study of viruses from Table 1, Appendix 1 shows that each considered here sequence has individual *atg*-distribution. It means that most mutations are combined with the joint variations of word content, word length, and the number of these words. Other mutations with only word content variation may exist. However, the *atg*-walks cannot see them, and the single-symbol distributions considered below will help us to detect these modifications of viruses (See Section 2.2.2.2).

Figure 6 shows a detailed comparison of samples of five viruses causing increased trouble for specialists with the one from Wuhan, China. The inlets offer the details of these curves at their beginning and end. The tails of the three curves are closed between the Wuhan and Brazil trajectories. Although the difference between these curves is not significant, the mutations may have complicated consequences in the rate of contagiousness of viruses.

Different techniques for numerical comparison of sequences are known from data analytics, including, for instance, calculation of correlation coefficients of unstructured data sequences, data distance values, and clustering of data, among others [10],[56],[57]. Researching RNA sequences, we suppose that the error of nucleotide detection is essentially less than one percent; otherwise, the results of comparisons would be instrumentally noisy

We use a simplified algorithm for quantitative comparing *atg*-distributions of different virus samples. Each numbered *atg*-triplet (y_i) has its coordinate along a sequence (x_i). Thanks to mutations, the length of some coding words varied together with the coordinate (x_i) of a triplet.

In our case, we calculated the difference (deviation) between the coordinates (x_i) of *atg*-triplets of the same numbers (y_i) in the compared sequences. This operation was fulfilled only for the sequences of equal *atg*-triplets; otherwise, excessive coding words are neglected in comparisons. Of course, such a technique for the comparison of geometrical data has its disadvantages. Therefore, if a compared sequence has several *atg*-triplets fewer than the number of *atg*-ones in a reference sequence, the *atg*s of reference RNA are excluded from comparisons. Still, it allows for obtaining some information on mutations of viruses in a straightforward and resultative way that will be seen below.

Our approach supposes choosing a reference nucleotide sequence to compare the genomic virus data of other samples, and it is a complete genomic sequence MN988668.1 from GenBank (row 1, Table 1). Several virus samples from GenBank and GISAID have been studied in this way [44], and some results of comparisons are given in Figure 7.

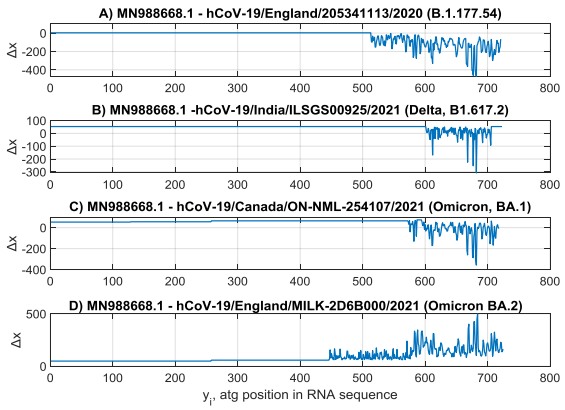

Fig. 7: Deviation of *atg*-coordinates in RNAs of four SARS CoV-2 viruses relative to the reference RNA MN988668.1 (see Table 1, Appendix 1).

The ordinary axis Δx in these plots shows the deviation of coordinates x_i of *atg*-triplets from the *atg*-coordinates of the reference sequence. As a rule, due to the different number of noncoding nucleotides at the beginning of complete RNA sequences, the curves in Figure 7 have constant biasing along the Δx axis.

The straight parts of these curves mean that the *atg* positions of a compared sequence are not perturbed regarding the corresponding coordinates in the reference RNA. It means that there are no mutations, or they are only with the variation of coding words without affecting their lengths if these mutations have a place.

In some studied samples (here and in Refs. [44],[72]), perturbations are near the end of the *orf1ab* gene, as is seen using a graphical tool of GenBank [41]. Perturbations were detected by calculating the x_i coordinates according to known y_i. The *atg*-perturbations could generally occur in any RNA's part, considering the random nature of mutations (Figure 7C, D). Relative deviation $|\Delta x_i|/N$ did not exceed 1–2% for compared viruses. Although this deviation is mathematically tiny, it may have severe consequences in the biological sense.

Our study shows that these difference curves (Figure 7) are individual for the studied samples. Although mutations without affecting the *atg*-distributions are possible, this individuality, theoretically, may be lost.

There are repeating motifs of comparison curves (Figure 7 and Refs. [44], [72]). The origin of this is unknown, but it was not coupled with the lineages of viruses and their clades. Other viruses can be studied in a similar manner.

3.2. Study of *atg*-walks of Complete Genome Sequences of the Middle East Respiratory Syndrome-related Coronavirus

Middle East Respiratory Syndrome-related (MERS) is a viral illness. The virus' origin is unknown, but it initially spread through camels and was first registered in Saudi Arabia [73],[74]. Most people infected with the MERS CoV virus developed a severe respiratory disease, which resulted in multiple human deaths.

Our simulation of *atg*-distributions of this virus shows the compactness of the calculated curves (Figure 8, and Table 2, Appendix 1), like the SARS CoV-2 characteristics. It follows that both viruses demonstrate relatively stable features towards the strong mutations connected with the recombination of the virus's parts. For instance, the relative difference $\delta_{1,10}y(x_i = 29982)$ of these curves is estimated at around 1% only. On average, the MERS RNAs have fewer number of *atg*-triplets, and longer nucleotide words than the SARS CoV-2 studied sequences.

In general, according to our more than three-year observation, the two studied coronaviruses (MERS CoV and SARS CoV-2) demonstrate relatively strong stability of their *atg*-distributions towards severe mutations, leading to the variation of codon positions, word lengths, and the number of words. It follows the conclusions of many scientists working in virology and virus genomics [71].

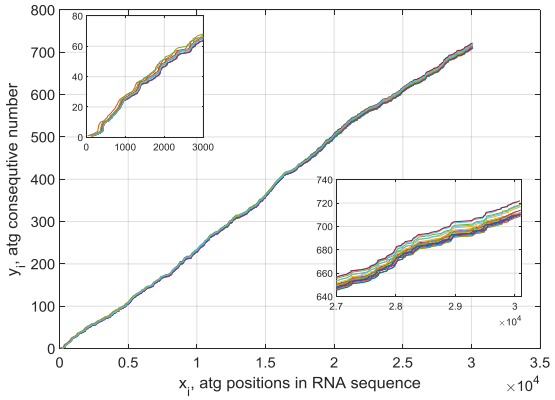

Fig. 8: Distributions of *atg*-triplets of 20 samples of the MERS CoV complete RNA sequences (Table 2, Appendix 1). Inlet shows the *atg*-distributions at the end of genome sequences. The inlets offer these distributions at the beginning and end of genome sequences.

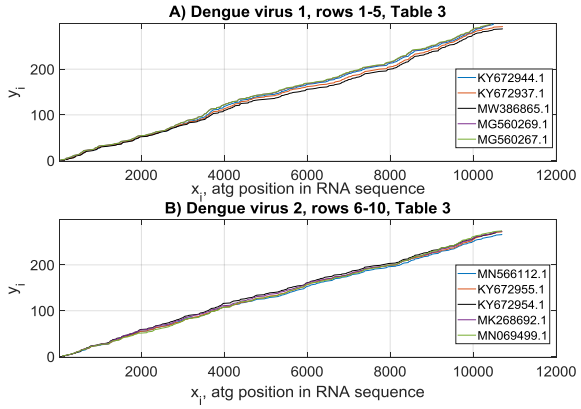

Fig. 9: Distributions of *atg*-triplets of complete RNA sequences of the Dengue virus-1 - (A) and Dengue virus-2 - (B). See rows 1-5 and 6-10, correspondingly from Table 3, Appendix 1.

3.3. Dengue Virus Study

The Dengue virus is spread through mosquito bites. For instance, a comprehensive review of the genomics of this virus can be found in Refs. [75],[76].

Unlike the coronaviruses, the Dengue virus (Table 3, Appendix 1) tends to form separate families, i.e., it is less stable than SARS-CoV-2 and MERS viruses. It has five genotypes (DENV 1–5) and around 47 strains. Unfortunately, the genomic data for this virus have been published less compared to coronaviruses. Some of them for which the complete genome data are available from GenBank have been studied below.

Figure 9A (rows 1-5, Table 3, Appendix 1) shows the *atg*-distributions of five sequences of Dengue virus-1 found in China. A rather large dispersion of these sequences is seen from these graphs.

Figure 9B gives the *atg*-distributions of five complete sequences of the Dengue virus-2. 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 an increased number of shorter words compared to the sequences of Dengue virus-1.

Figure 10A shows five data sets for different strains of Dengue virus-3 registered in many countries. They have about the same number of nucleotides, and comparable averaged lengths of words.

Figure 10B gives three *atg*-distributions of a Gabon-strain [76] of Dengue virus-3. It is supposed that this strain mutated from the earlier registered Dengue virus lines (Figure 10A). However, they are different in the length of complete genome sequences and their statistical characteristics, which will be considered in Section 3.5 below.

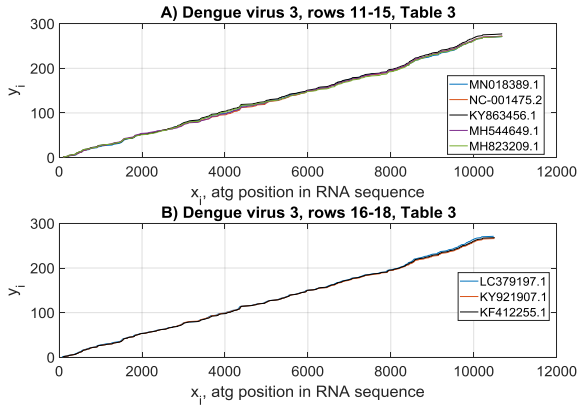

Fig. 10: Distributions of *atg*-triplets of complete RNA sequences of Dengue virus-3 (rows 11-15 –(A) and 16-18 –(B) Table 3, Appendix 1).

Figure 11A shows the *atg*-distributions of two Gabon-originated Dengue viruses that can relate to predecessors of other Dengue viruses of this family. Figure 11B 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.

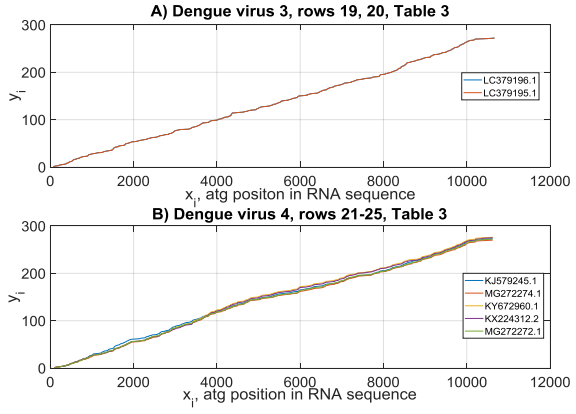

Fig. 11: Distributions of *atg*-triplets of complete RNA sequences of Dengue virus-3 (rows 19, 20 - (A), Table 3, Appendix 1) and Dengue-4 (rows 21-25 - (B), Table 4, Appendix 1).

A consolidated plot of all *atg*-curves of the Dengue RNAs studied here is shown in Figure 12. There is a substantial divergence of these trajectories in agreement with the mutation rate of this virus being relatively strong. For instance, this relative difference $\delta y_{1,25}$ estimated at $x_i = 10000$ is 14.1%.

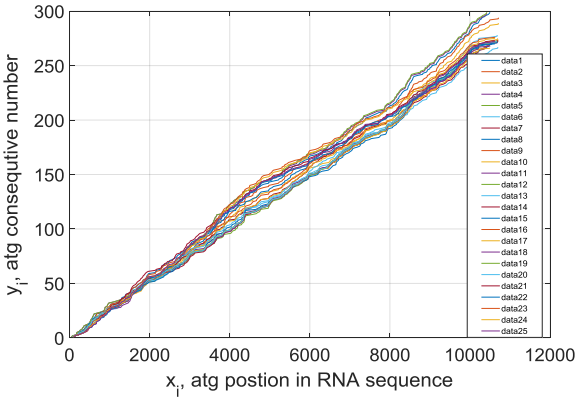

Fig. 12: Consolidated picture for all 25 Dengue virus arbitrary-chosen samples studied in this paper. The numbers of virus *atg*-curves correspond to Table 3, Appendix 1.

3.4. Analysis of *atg*-walks of Complete Genome Sequences of the Ebola Virus

There are four strains of the Ebola virus known worldwide, although many other mutations of this virus can be found. Like the Dengue virus, Ebola shows instability and an increased rate of mutations. 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 and humans. This fever is distinguished by a high death rate (from 25% to 90% of the infected individuals). A recent comprehensive review of the genomics of this virus can be found in Ref. [77].

The Ebola virus RNA consists of 19 000 nucleotides and more than three hundred *atg*-triplets. Figure 13A shows four sequences of this virus belonging to the EBOV strain registered in Zaire and Gabon. Three of them are very close to each other, but the mutant Zaire virus (in red) has some differences from the three others. The samples collected in Sudan (SUDV) are closer to each other (Figure 13B), but they have an increased number of *atg*-triplets and shorter words.

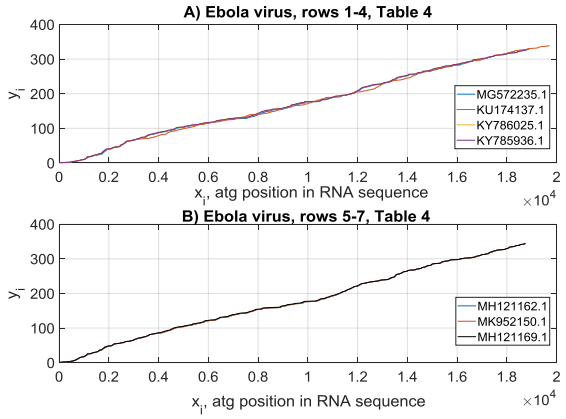

Fig. 13: Distributions of *atg*-triplets of complete RNA sequences of the Ebola - EBOV) virus from Zaire and Gabon (rows 1-4 - (A), Table 4, Appendix 1) and Ebola virus - SUDV from Sudan (rows 5-7 - (B), Table 4, Appendix 1).

The Bombali virus, registered in Sierra Leone, West Africa, is considered a new strain of Ebola. The *atg*-distributions of the five found in GenBank RNA sequences are different even visually from the two reviewed above, as seen in Figure 14A. Another Ebola strain that can be compared with the one studied above is the Bundibugyo (BDBV) virus, whose three *atg*-distributions are shown in Figure 14B.

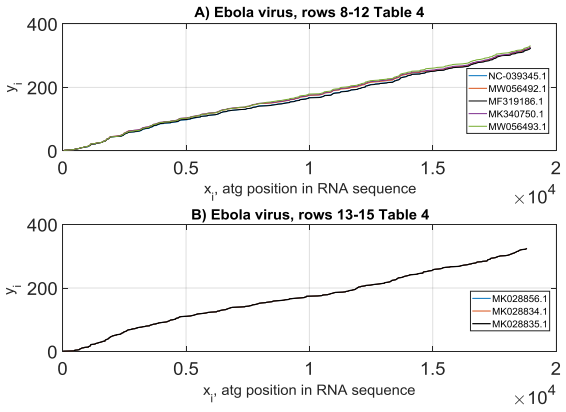

Fig. 14: Distributions of *atg*-triplets of five complete RNA sequences of the Ebola (Bombali) virus (rows 8-12 – (A), Table 4, Appendix 1) and three complete RNA sequences of the Ebola Bundibugyo (BDBV) virus. See rows 13-15 – (B), Table 4, Appendix 1.

The calculated 15 distributions are consolidated in Figure 15 to compare all four strains, where, instead of points, the results are represented by thin curves to make these distributions more visible. Here, the tendency of *atg*-curves to divergence and form clusters is seen. For instance, the relative difference of strains $\delta y_{1,15}$ is estimated at around 9%.

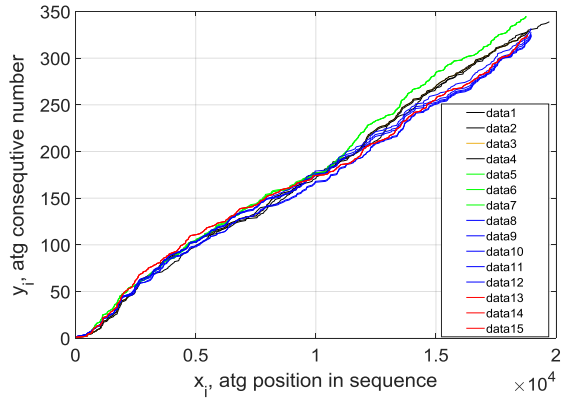

Fig. 15: Consolidated representation of 15 *atg*-distributions of four strains of the Ebola virus. Black color – EBOV; Green color – SUDV; Violet color – Bombali; Red color – BDBV. The numbers of virus *atg*-curves correspond to Table 4, Appendix 1.

It follows that the *atg*-walk is an effective visualization tool sensitive to the viral RNA mutations coupled to the number of codons and introns and word size variations. It allows the detection of viruses with essentially unstable genomes distinguished by their increased deviation of *atg*-walks and their fractal properties.

3.5 Statistical Characterisation of *atg*-walks: Calculating, Mapping, and Processing of the Inter-*atg* Distance Values

In this research, after applying the tool FracLab (See Section 2.2.3), it was discovered that all studied genomic sequences of the SARS CoV-2, MERS CoV, Dengue, and Ebola viruses have fractality in their word-length distributions. The results on this matter are placed in columns 6 of Tables 1-4, Appendix 1.

Figure 16A shows the fractal dimension values d_F of 37 genome sequences of SARS CoV-2 viruses. Figure 16B gives this parameter for 20 MERS coronavirus samples. Calculations show that the maximal normalized deviation (relative difference) of d_F is only 1.8% for SARS and 2% for MERS viruses, i.e., these results are in accordance with the conclusion on the stability of these viruses towards severe mutations with the codon-length and codon-content perturbations.

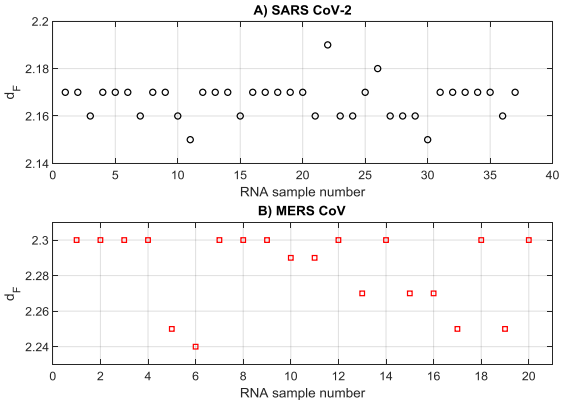

Fig. 16: Fractal dimensions d_F of word-length $l_{i,i+1}^{(atg)}$ distributions of 37 complete genome sequences of the SARS CoV-2 (A) and 25 complete genome sequences of the MERS CoV (B) viruses. The number of samples corresponds to Tables 1 and 2 of Appendix 1.

The Dengue virus has five families and 47 strains; they have different *atg*-distributions and fractal dimensions. According to the fractal calculations, some strains are close to each other (Figure 17).

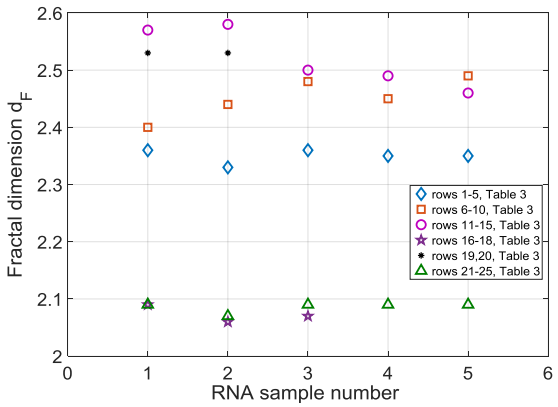

Fig. 17: Fractal dimensions d_F of word-length $l_{i,i+1}^{(atg)}$ distributions of 25 complete genome sequences of the Dengue 1-4 viruses and their strains (Table 3, Appendix 1).

At the same time, the maximal relative deviation is around 22% estimated for all studied samples, and it agrees that this virus is unstable from strain to strain.

The same conclusion is evident in Figure 18, where the fractal dimensions of several studied strains of the Ebola virus are given with a maximal relative deviation of around 18%.

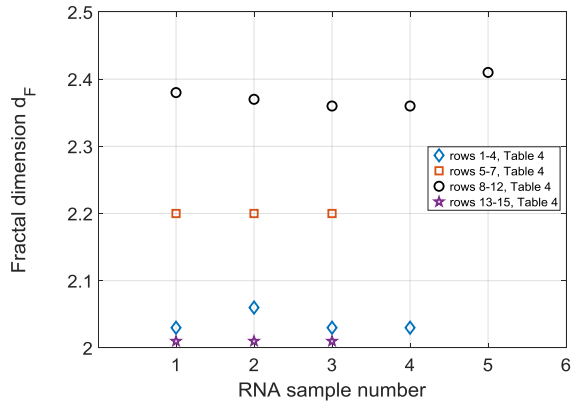

Fig. 18: Fractal dimensions d_F of word-length $l_{i,i+1}^{(atg)}$ distributions of 15 complete genome sequences of the Ebola virus strains (Table 4, Appendix 1).

4. Discussion

The research on the RNAs and DNAs of viruses and cellular organisms is a highly complex problem because of the 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 studies are exciting and can be fruitful.

In this paper, the viral RNAs were studied using a novel algorithm based on exploring the RNA patterns of arbitrary length. One of the operations of this algorithm is the numerical mapping of RNA characters, which is performed by calculating the Hamming distance between the preliminary binary-expressed queries and RNA symbols. This allows fulfilling these steps approximately twice as fast regarding the operations with real numbers [45]. The results of the application of this algorithm are verified by comparing them with complete RNA sequences. Considering this algorithm can search arbitrary-length patterns, the trajectories of separate symbols can be combined with the *atg*-walks for multi-scale plotting and RNA analysis, as shown in this contribution.

The mentioned *atg*-triplets compose relatively stable 1-D distributions called the RNA schemes in this paper. These distributions have been studied using our algorithm applied to arbitrary-chosen complete RNA sequences of the SARS CoV-2, MERS CoV, Dengue, and Ebola viruses.

The following properties of virus RNAs have been found in our research and not seen earlier:

1. Stability of *atg*-schemes towards intra-family mutations when the geometry of *atg*-curves is only slightly distorted according to the estimates of the relative difference of curve points (Section 3)
2. The highly compact *atg*-curve sets of the SARS CoV-2 and MERS CoV viruses, despite their continuous mutation (to the date of submission), estimated visually and quantitively calculating the relative difference of *atg*-coordinates (δy=1-1.6%), see Sections 3.1 and 3.2
3. More substantial divergence (δy=9-14%) of *atg*-curves of the Dengue and Ebola viruses in comparison to SARS CoV-2 and MERS CoV species (Sections 3.3 and 3.4 and 3.1, 3.2)
4. A visually found tendency towards clustering the *atg*-curves in the limits of one virus family (Ebola case, Section 3.4)
5. Distribution of single RNA's symbols and *atg*-triplets according to the random

fractal Cantor rule (Section 2.2.2.1)
6. Possible global correlation of the inter-triplet distances due to this fractality (Section 2.2.3)
7. Correlation of dispersion of fractal dimension values of *atg*-distributions with the instability of viruses (Section 3.5)

5. Conclusion

In this paper, the visual and quantitative analyses of viral RNAs have been performed using a novel algorithm to calculate the RNA pattern positions in the studied sequences. A part of this code uses binary symbols of RNA nucleotides for accelerated search. The algorithm allows more effective genomic studies by building 1-D distributions of different patterns and combining these sets on a single multi-scale plot.

The proposed techniques were applied to analyze the SARS CoV-2, MERS CoV, Dengue, and Ebola viruses. The 1-D distributions of *atg*-triplets (*atg*-schemes of RNAs) were calculated and plotted for these species. The levels of stability of these distributions have been estimated visually and numerically by calculating the divergence of triplet curves and deviation of word fractal dimensions of RNAs.

The main finding of this work is that the level of clustering of *atg*-curves is with the degree of stability of RNAs shown comparing the SARS-CoV-2 and MERS species and the Dengue and Ebola viruses. In addition, the rank of clustering is coupled with the deviation of fractal dimensions of codon-length distributions, which is more significant for the unstable Dengue and Ebola viruses mentioned. It may be found engaging in further study of the mutation of viruses and building their phylogenetic trees.

The developed approach is with the study of the RNA words, their lengths, and fractal properties of the word-length distributions. It can be applied in the research of mammalian DNAs where the gene length is the evolutionary dynamic and partly defines the gene expression level [78].

In addition, recent studies show that with aging, an imbalance of short and long genes in the transcriptome occurs, and the research of these phenomena by our algorithm may play an important role in the development of anti-aging treatments [79].

Thus, even quantitative and qualitative studies and modeling of short virus genes can be helpful in the genetics of more complicated mammalian DNAs consisting of billions of nucleotides.

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; NP: nondeterministic polynomial; UTF-8: Unicode Transformation Format-8 bit; US-ASCII: American Standard Code for Information Interchange; MERS CoV: Middle-East Respiratory Syndrome-related Corona Virus.

Acknowledgments

The authors thank the GenBank® [39] and GISAID [40] genetic data banks, and all researchers placed their genomic sequences in them. The online text processing service of https://onlinetexttools.com/ is appreciated.

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*., Vol. 11, 2015, pp. 951-967. https://dx.doi.org/10.1101%2Fpdb.top084970

[3]. G. Storz, An expanding universe of noncoding RNAs, *Science*, Vol. 296, 2002, pp. 1260-1263. https://doi.org/10.1126/science.1072249

[4]. C. Nello and M.W. Hahn, *Introduction to Computational Genomics: A Case Studies*

Approach. Cambridge, University Press, 2012. https://doi.org/10.1017/CBO9780511808982

[5]. H.K. Kwan and S.B. Arniker, Numerical representation of DNA sequences. *Proc. 2009 IEEE Int. Conf., Electro/Information Technology*, 2009, pp. 307-310. http://dx.doi.org/10.1109/EIT.2009.5189632

[6]. C. Cattani, Complex representation of DNA sequences, *Commun. in Computer and Inform. Sci.,* Vol. 13, 2008, pp. 528-537. http://dx.doi.org/10.1007/978-3-540-70600-7_42

[7]. P.D. Cristea, Conversation of nucleotide sequences into genomic signals, *J. Cell. Mol. Med*., Vol. 6, 2002, pp. 279-303. https://doi.org/10.1111/j.1582-4934.2002.tb00196.x

[8]. 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*., Vol. 62, 2015, pp. 150-156. https://doi.org/10.1016/j.jmgm.2015.09.011

[9]. B. Brejová, T. Vinar, and M. Li, Pattern discovery. In: Krawetz S.A., Womble D.D. (eds) *Introduction to Bioinformatics*, Humana Press, Totowa, NJ, 2003.

[10]. J. Zhang, *Visualization for Information Retrieval*, Springer, 2007. https://doi.org/10.1007/978-0-387-39940-9_954

[11]. M. Randic, M. Novic, and D. Plavsic. Milestones in graphical bioinformatics, *Int. J. Quantum Chem*., Vol. 113, 2013, pp. 2413-2446. https://doi.org/10.1002/qua.24479

[12]. P.P. Vaidyanathan, Genomics and proteomics: A signal processing tour, *IEEE Circ. Syst. Mag*., 4th Quarter, 2004, pp. 6-29. https://doi.org/10.1109/MCAS.2004.1371584

[13]. 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*, Vol. 4, 2009, pp. 28-40. https://doi.org/10.2174/157489309787158134

[14]. L. Das, S. Nanda, and J.K. Das, An integrated approach for identification of exon locations using recursive Gauss-Newton tuned adaptive Kaiser window, *Genomics*, Vol. 111, 2019, pp. 284-296. https://doi.org/10.1016/j.ygeno.2018.10.008

[15]. A. E. Lamairia, Nonexistence results of global solutions for fractional order integral equations on the Heisenberg group, *WSEAS Trans. Systems*, Vol. 21, 2022, pp. 382-386. http://dx.doi.org/10.37394/23202.2022.21.42

[16]. N. Viriyapong, Modification of Sumudu Decomposition method for nonlinear fractional Volterra integro-differential equations, *WSEAS Trans. Math*., Vol. 21, 2022, pp. 187-195. DOI: 10.37394/23206.2022.21.25

[17]. A. Czerniecka, D. Bielinska-Waz, P. Waz, and T. Clark, 20D-dynamic representation of protein sequences, *Genomics*, Vol. 107, 2016, pp. 16-23. https://doi.org/10.1016/j.ygeno.2015.12.003

[18]. E.R. Hamori and J. Raskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem*., Vol. 258, 1983, pp. 1318-1327. https://doi.org/10.1016/S0021-9258(18)33196-X

[19]. M.A. Gates, Simpler DNA representation, *Nature*, Vol. 316, 1985, pp. 219. https://doi.org/10.1038/316219a0

[20]. C.L. Berthelsen, J.A. Glazier, and M.H. Skolnick, Global fractal dimension of human DNA sequences treated as pseudorandom walks, *Phys. Rev. A*., Vol. 45, 1992, Paper No 89028913. https://doi.org/10.1103/PhysRevA.45.8902

[21]. P. Licinio and R.B. Caligiorne, Inference of phylogenetic distances from DNA-walk divergences, *Physica A*, Vol. 341, 2004, pp. 471-481. http://dx.doi.org/10.1016/j.physa.2004.03.098

[22]. J.A. Berger, S.K. Mitra, M. Carli, and A. Neri, Visualization and analysis of DNA sequences using DNA walks, *J. Franklin Inst*., Vol. 341, 2004, pp. 37-53. https://doi.org/10.1016/j.jfranklin.2003.12.002

[23]. A. Rosas, E. Nogueira Jr., and J.F. Fontanari, Multifractal analysis of DNA walks and trails, *Phys. Rev. E*, Vol. 66, 2002, Paper No 061906.

http://dx.doi.org/10.1103/PhysRevE.66.061906

[24]. A.D. Haimovich, B. Byrne, R. Ramaswamy, and W.J. Welsh, Wavelet analysis of DNA walks, *J. Comput. Biol.*, Vol. 13, 2006, pp. 1289-1298. https://doi.org/10.1089/cmb.2006.13.1289

[25]. H. Namazi, V.V. Kulish, F. Delaviz, and A. Delaviz, Diagnosis of skin cancer by correlation and complexity analyses of damaged DNA, *Onkotarget*, Vol. 6, 2015, pp. 42623-42631. https://dx.doi.org/10.18632%2Foncotarget.6003

[26]. 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.*, Vol. 35, 2019, pp. 2738-2748. https://doi.org/10.1093/bioinformatics/bty1021

[27]. K.S. Birdi, *Fractals in Chemistry, Geochemistry, and Biophysics*, N.-Y., Plenum Press, 1993.

[28]. T.G. Dewey, *Fractals in Molecular Biophysics*, Cambridge, Oxford University Press, 1997.

[29]. G. Abramson, H.A. Cerdeira, and C. Bruschi, Fractal properties of DNA walks, *Biosystems*, Vol. 49, 1999, pp. 63-70, https://doi.org/10.1016/s0303-2647(98)00032-x

[30]. C. Cattani, Fractals and hidden symmetries in DNA, *Math. Problems Eng.*, Vol. 2010, 2010, Paper No 507056(1-31). https://doi.org/10.1155/2010/507056

[31]. 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

[32]. B. Hao, H.C. Lee, and S. Zhang, Fractals related to long DNA sequences and complete genomes, *Chaos, Solitons and Fractals*, Vol. 11, 2000, pp. 825-836. https://doi.org/10.1016/S0960-0779(98)00182-9

[33]. Z.-Y. Su, T. Wu, and S.-Y. Wang, Local scaling and multifractality spectrum analysis of DNA sequences- GenBank data analysis, *Chaos, Solitons&Fractals*, Vol. 40, 2009, pp. 1750-1765. https://doi.org/10.1016/j.chaos.2007.09.078

[34]. 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.*, Vol. 4, 2019, pp. 267-278. https://doi.org/10.2478/AMNS.2019.1.00023

[35]. M.S. Swapna and S. Sankararaman, Fractal applications in bio-nanosystems, *Bioequiv. Availab.*, Vol. 2, 2019, Paper No OABB.000541.

[36]. X. Bin, E.H. Sargent, and S.O. Kelley, Nanostructuring of sensors determines the efficiency of biomolecular capture, *Anal. Chem.*, Vol. 82, 2010, pp. 5928–5931. https://doi.org/10.1021/ac101164n

[37]. J. Chen, Z. Luo, C. Sun, Z. Huang, C. Zhou, S. Yin, Y. Duan, and Y. Li, Research progress of DNA walker and its recent applications in biosensor, *TrAC Trends in Anal. Chem.*, Vol. 120, 2019, Paper No 115626. https://doi.org/10.1016/j.trac.2019.115626

[38]. A. Sadana, *Engineering Biosensors. Kinetics and Design Application*, San Diego, California, Acad. Press, 2001. https://doi.org/10.1016/B978-0-12-613763-7.X5015-0

[39]. G.A. Kouzaev, Frequency dependence of microwave-assisted electron-transfer chemical reactions, *Mol. Phys.*, Vol. 118, 2020, paper No e1685691. https://doi.org/10.1080/00268976.2019.1685691

[40]. S.V. Kapranov and G.A. Kouzaev, Nonlinear dynamics of dipoles in microwave electric field of a nanocoaxial tubular reactor, *Mol. Phys.*, Vol. 117, 2018, pp. 489-506. https://doi.org/10.1080/00268976.2018.1524526

[41]. GenBank® [https://www.ncbi.nlm.nih.gov/genbank/].

[42]. Global Initiative on Sharing All Influenza Data (GISAID) [https://www.gisaid.org/].

[43]. A. Belinsky and G.A. Kouzaev, Visual and quantitative analyses of virus genomic sequences using a metric-based algorithm,

bioArxiv preprint: bioArxiv 2021.06.17.448868; Europe PMC: PPR: PPR358597. https://doi.org/10.1101/2021.06.17.448868

[44]. A. Belinsky and G.A. Kouzaev, Geometrical study of virus RNA sequences, *bioArxiv preprint*: bioRxiv 2021.09.06.459135; https://doi.org/10.1101/2021.09.06.459135; *Europe PMC*: https://europepmc.org/article/PPR/PPR391263

[45]. R. Mian, M. Shintani, and M. Inoue, Hardware-software co-design for decimal multiplication, *Computers*, Vol. 10, 2021, pp. 17(1-19). https://doi.org/10.3390/computers10020017

[46]. N. Brisebarre, C. Lauter, M, Mezzarobba, and J.-M. Muller, Comparison between binary and decimal floating-point numbers, *IEEE Trans. Comput*., Vol. 65, 2016, pp. 2032-2044. https://doi.org/10.1109/TC.2015.2479602

[47]. A. Kostadinov and G.A. Kouzaev, A novel processor for artificial intelligence acceleration, *WSEAS Trans. Circ., Systems*, Vol. 21, 2022, pp. 125-141. https://doi.org/10.37394/23201.2022.21.14

[48]. Matlab® R2020b, version 9.9.0.1477703, [https://se.mathworks.com/products/matlab.html]

[49]. Chapter 2. *General Structure. The Unicode Standard* (6.0 ed.). Mountain View, California, US: The Unicode Consortium. ISBN 978-1-936213-01-6.

[50]. R.W. Hamming, Error detecting and error-correcting codes, *Bell Syst. Techn. J*., Vol. 29, 1950, pp. 147-160.

[51]. W.N. Waggener, *Pulse Code Modulation Techniques*, Berlin-Heidelberg: Springer Verlag, 1995.

[52]. G. Navarro and M. Raffinot, *Flexible Pattern Matching in Strings: Practical Online Search Algorithms for Texts and Biological Sequences*, Cambridge: Cambridge University Press, 2002. https://doi.org/10.1017/CBO9781316135228

[53]. V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, Vol. 10, 1966, pp. 707–710.

[54]. E. Gabidullin, Theory of codes with maximum rank distance, *Probl. Inform. Trans*., Vol. 21, 1985, pp. 1-76.

[55]. 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.

[56]. X. Yang, N. Dong, E. Chan, and S. Chen, Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries, *Emerging Microbes&Infect.*, Vol. 9, 2020, pp. 1287-1299. https://doi.org/10.1080/22221751.2020.1773745

[57]. J. Tzeng, H.H.-S. Lu, and W.-H. Li, Multidimensional scaling for large genomic data sets, *BMC Bioinformatics*, Vol. 9, 2008, Article No 179, pp. 1-17. https://doi.org/10.1186/1471-2105-9-179

[58]. *Online Text Tools* [https://onlinetexttools.com/].

[59]. J. Feder, *Fractals*, N.-Y., Plenum Press, 1988.

[60]. P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, *Physica D*, Vol. 9, 1983, pp. 189-208. https://doi.org/10.1016/0167-2789(83)90298-1

[61]. S.N. Rasband, *Chaotic Dynamics of Nonlinear Systems*. Weinheim, J. Wiley & Sons, 1989.

[62]. 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, pp. 1-39.

[63]. F. Roueff and J.L. Véhel, A regularization approach to fractional dimension estimation. In: *Proc. Int. Conf. Fractals 98*, Oct. 1998, Valletta, Malta. World Sci., 1998, pp. 1-14.

[64]. J.L. Véhel and P. Legrand, Signal and image processing with Fraclab, In: *Thinking in Patterns*. World Sci., 2003, pp. 321-322.

[65]. G.A. Kouzaev, *Application of Advanced Electromagnetics. Components and Systems*. Berlin-Heidelberg: Springer, 2013. https://doi.org/10.1007/978-3-642-30310-4

[66]. C. Guidolin, R. 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, pp. 138-156. http://dx.doi.org/10.1007/978-1-4939-3995-4

[67]. *FracLab 2.2. A fractal analysis toolbox for signal and image processing*. [https://project.inria.fr/fraclab/].

[68]. 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.

[69]. A. Rahimi, A. Mirzazadeh, and S. Tavakopolour, Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection, *Genomics*, Vol. 113, 2021, pp. 1221-1232. https://doi.org/10.1016/j.ygeno.2020.09.059

[70]. P. Forster. L. Forster, C. Renfrew, and M. Forster, Phylogenic network analysis of SARS-CoV-2 genomes. *PNAS,* Vol. 117, 2020, pp. 9241-9243. https://doi.org/10.1073/pnas.2004999117

[71]. V. Cooper, The coronavirus variants don't seem to be highly variable so far, *Sci. American*, 2021, March 24.

[72]. G.A. Kouzaev, The geometry of ATG-walks of the Omicron SARS CoV-2 Virus RNAs, *bioArxiv preprint*: bioRxiv doi: https://doi.org/10.1101/2021.12.20.473613; *Europe PMC*: PPR: PPR435860.

[73]. S.A. El-Kafrawy, V.M. Corman, A.M. Tolah, S.B. Al Masaudi, A.M. Hassan, M.A. Müller, T. Bleicker, S. M. Harakeh, A.A. Alzahrani, G.A.A. Abdulaziz, N. Alagili, A.M. Hashem, A. Zumla, C. Drosten, and E.I. Azhar, Enzootic patterns of Middle East respiratory syndrome coronavirus in imported African and local Arabian dromedary camels: a prospective genomic study, *The Lancet Planetary Health*, Vol. 3, 2019, pp. e521-e528. https://doi.org/10.1016/S2542-5196(19)30243-8

[74]. 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*., Vol. 9, 2020, pp. 2714-2720. https://doi.org/10.1080/22221751.2020.1861914

[75]. V.D. Dwivedi, I.P. Tripathi, R.C. Tripathi, S. Bharadwaj, and S.K Mishra, Genomics, proteomics and evolution of dengue virus, *Briefings in Functional Genomics*, Vol. 16, 2017, pp. 217-227. https://doi.org/10.1093/bfgp/elw040

[76]. 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. Infect. Diseases*, Vol. 91, 2020, pp. 129-136. https://doi.org/10.1016/j.ijid.2019.12.002

[77]. N. Di Paola, M. Sanchez-Lockhart, X. Zeng, J.H. Kuhn, and G. Palacios, Viral genomics in Ebola virus research, *Nature Rev. Microbiol*., Vol. 8, 2020, pp. 365–378. https://doi.org/10.1038/s41579-020-0354-7

[78]. V. Grishkevich and I. Yanai, Gene length and expression level shape genomic novelties, *Genome Research*, Vol. 24, 2014, pp. 1497-1503. https://doi.org/10.1101%2Fgr.169722.113

[79]. T. Stoeger, R.A. Grant, A.C. McQuattie-Pimentel, K.R. Anekalla, S.S. Liu, H. Tejedor-Navarro, B.D. Singer, H. Abdala-Valencia, M. Schwake, M.P. Tetreault, H. Perlman, W E. Balch, N.S. Chandel, K.M. Ridge, J.I. Sznajder, R.I. Morimoto, A.V. Misharin G R. Scott Budinger, and L.A.N. Amaral, Aging is associated with a systemic length-associated transcriptosome imbalance, *Nature Aging*, vol. 2, 2022, pp. 1191-1206. https://doi.org/10.1038/s43587-022-00317-6

Contribution of individual authors to the creation of a scientific article

All authors are contributed equally

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

Appendix 1. Results of statistical characterization of complete genetic sequences of the SARS CoV-2, MERS CoV, Dengue, and Ebola viruses

Table 1. Severe acute respiratory syndrome coronavirus 2, (GenBank, GISAID), *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 Word-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, (Delta) 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	57.96	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.1.177.54**, 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**, 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**,	29819	720	29	58.10	2.17

	Illumina MiSeq					
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
20	*hCoV-19/Canada/ON-NML-254107/2021*, **GR, BA.1 (Omicron),** Oxford Nanopore GridION	29685	718	29	57.73	2.17
21	*hCoV-19/England/MILK-2D6B000/2021*, GRA, **BA.2 (Omicron),** Illumina NovaSeq	29724	725	28.5	57.48	2.16
22	*hCoV-19/USA/CA-CDC-LC0934366/2022,* **GRA, BQ.1**, PacBio Sequel II	29642	718	28	57.31	2.19
23	*hCoV-19/USA/CA-CDC-LC0933659/2022, **GRA, BQ.1***, PacBio Sequel II	29668	725	28	57.5	2.16
24	*hCoV-19/Canada/QC-L00549819001/2022,* **GRA, BQ.1**, Illumina NextSeq	29719	725	28	57.5	2.16
25	*hCoV-19/Malaysia/IMR/CV05212/2022,* **GRA, BQ.1.23**, Nanopore GridION	29636	725	28	57.5	2.17
26	hCoV-19/Ireland/CO-CUH-S22C0165/2022, **GRA, BQ.1**, Nanopore MinION	29646	723	28	57.4	2.18
27	hCoV-19/Canada/QC-L00548493001/2022, **GRA, BA.4**, Oxford Nanopore PromethION	29709	724	29	57.5	2.16
28	*hCoV-19/Guatemala/7817-LNS/2022* **GRA, BA.4**, Illumina MiSeq	29714	724	29	57.48	2.16
29	*hcov-19/USA/UT-UPHL-221115393771/2022,* **GRA, BA.4.6**, Illumina NovaSeq 6000	29734	723	28.5	57.68	2.16
30	*hCoV-19/Germany/RP-USAFSAM-S20372/2022,* **GRA, BA.5,** Illumina_NextSeq_Mid	29693	722	29	57.66	2.15
31	*hCoV-19/Russia/MOW-CRIE-89961/2022,* **GRA, BA.5,** Oxford Nanopore	29646	719	29	57.65	2.17
32	*hCoV-19/Norway/OUS-26253/2022,* ***GRA,* BA.5.1**, Illumina Swift Amplicon SARS-CoV-2 protocol at Norwegian Sequencing Centre	29605	719	29	57.61	2.17
33	*hCoV-19/USA/CA-CDPH-FS48102807/2022,* **GRA, BA.5.1.1,** Element Biosciences	29607	723	29	57.36	2.17
34	*hCoV-19/USA/NY-NYULH9985/2022,* **GRA, XBB.1.5**, Amplicon (Illumina), Illumina NovaSeq	29773	726	29	57.44	2.17
35	*hCoV-19/USA/NJ-PHEL-V22054945/2022,* **GRA, XBB.1.5**, Oxford_Nanopore	29649	722	29	57.48	2.17
36	*hCoV-19/Iceland/L-3254/2022,* **GRA, XBB.1.5**, Illumina MiSeq	29669	725	28.5	57.47	2.16

37	*hCoV-19/Ireland/CO-CUH-S23C0065/2023.txt,* **GRA, CH.1.1**, Nanopore MinION	29652	722	29	57.49	2.17

Table 2. The Middle East respiratory syndrome-related coronavirus, (GenBank), *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 Word-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 [76]	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
11	*OL622035.1,* Middle East respiratory syndrome-related coronavirus isolate MERS-CoV_Riyadh_2016, 2021, Oxford Nanopore	29994	709	30	59.03	2.29
12	*OP712625.1*, Middle East respiratory syndrome-related coronavirus isolate MERS-CoV/dromedary camel/Egypt/NC4714/2016, 2022, Illumina	30108	719	30	57.88	2.3
13	*MH734114.1,* Middle East respiratory syndrome-related coronavirus isolate MERS-CoV camel/Kenya/C1215/2018, 2018	30033	721	29.5	58.05	2.27
14	*MG923468.1,* Middle East respiratory syndrome-related coronavirus isolate MERS-CoV camel/Ethiopia/AAU-EPHI-HKU4458/2017, 2018	30091	722	30	57.65	2.3
15	*KJ361503.1,* Middle East respiratory syndrome coronavirus isolate Hu-France - FRA2_130569-2013_Isolate_Sanger, 2014, Sanger dideoxy sequencing	30040	710	29	59.18	2.27
16	*KM210277.1,* Middle East respiratory syndrome coronavirus isolate England/4/2013, complete genome, 2014, Sanger dideoxy sequencing	30031	712	30	58.81	2.27
17	*KF958702.1,* Middle East respiratory syndrome coronavirus isolate MERS-CoV-Jeddah-human-1, 2013, Sanger dideoxy sequencing	29851	709	29	58.67	2.25
18	*OP654179.1,* Middle East respiratory syndrome-related coronavirus isolate MERS-CoV/dromedary camel/Egypt/NC270-P9/2015, 2022, Illumina	30131	711	30	58.87	2.3
19	MW086535.1, Middle East respiratory syndrome-related coronavirus isolate MERS-CoV/JC32/Ramtha, 2020, Illumina	29825	707	29.5	58.72	2.25
20	*MZ268405.1,* Middle East	30106	718	30	58.19	2.3

	respiratory syndrome-related coronavirus isolate MERS-CoV/Camel/Kenya/HKU-CAC10200/2020, 2021, Illumina					

Table 3. The Dengue virus, (GenBank), *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 Word-length Distribution
	1	2	3	4	5	6
1	*KY672944.1*, Dengue virus 1 isolate DENV-1/China/YN/YNH22 (2013), 2019, Sanger dideoxy	10709	299	23	47.74	2.36
2	*KY672937.1*, Dengue virus 1 isolate DENV-1/China/YN/DGRL-6(2014), 2019, Sanger dideoxy	10738	294	23	50.02	2.33
3	*MW386865.1*, Dengue virus 1 isolate YNBN04, China, 2020, Sanger dideoxy	10742	289	24	50.81	2.36
4	*MG560269.1*, Dengue virus 1 isolate P1253/China/GD/CZ/2014, 2018, Sanger dideoxy	10583	298	23	47.55	2.35
5	*MG560267.1,* Dengue virus 1 isolate P1258/China/GD/CZ/2014, 2018, Sanger dideoxy	10583	299	23	47.22	2.35
6	*MN566112.1,* Dengue virus 2 isolate New Caledonia-2018-AVS127, 2020, Illumina	10722	267	32	52.24	2.4
7	*KY672955.1*, Dengue virus 2 isolate DENV-2/China/YN/15DGR65(2015), 2019, Sanger dideoxy	10723	273	28	52.77	2.44
8	*KY672954.1*, Dengue virus 2 isolate DENV-2/China/YN/JH1516(2015), 2019, Sanger dideoxy	10665	271	29	51.50	2.48
9	*MK268692.1,* Dengue virus 2 isolate DENV-2/TH/1974, Thailand, 2019, Sanger dideoxy	10721	274	28	52.67	2.45
10	*MH069499.1,* Dengue virus 2 strain DENV-2/VE/IDAMS/910105, Venezuela, 2018, Illumina	10712	275	28	52.84	2.49
11	*MN018389.1*, Dengue virus 3 isolate D17011, China, 2020, Sanger dideoxy sequencing	10708	272	28	55.46	2.57
12	*NC_001475.3*, Dengue virus 3, Sri Lanka, 2019, Illumina	10707	273	27	55.05	2.58
13	*KY863456 .1*, Dengue virus 3 isolate 201610225, Indonesia, 2017, IonTorrent, Sanger	10707	278	28	52.84	2.5

	dideoxy sequencing					
14	*MH544649.1*, Dengue virus 3 isolate 449686_Antioquia_CO_2015, Colombia, 2018, Illumina; Sanger dideoxy sequencing	10707	273	28	52.84	2.49
15	*MH823209.1*, Dengue virus 3 isolate SMD-031, Indonesia, 2019, Illumina	10707	272	28	52.84	2.46
16	*LC379197.1*, Dengue virus 3 strain SYMAV-17/Gabon/2017 genomic RNA, 2019, Illumina	10641	271	29	52.85	2.06
17	*KY921907.1*, Dengue virus 3 isolate SG(EHI)D3/15095Y15, 2017, Singapore, Sanger dideoxy sequencing	10667	266	29	53.58	2.09
18	*KF041255.1,* Dengue virus 3 isolate D3/Pakistan/55505/2007, 2013, Sanger dideoxy sequencing	10675	268	29	53.55	2.07
19	*LC379196.1,* Dengue virus 3 strain SYMAV-09/Gabon/2016 genomic RNA, 2019, Illumina	10663	273	29	53.64	2.53
20	*LC379195.1,* Dengue virus 3 strain SYMAV-07/Gabon/2016 genomic RNA, 2019, Illumina	10663	273	29	53.64	2.53
21	*KJ579245.1*, Dengue virus 4 strain DENV-4/MT/BR23_TVP17909/2012, Brazil, 2020, Illumina	10649	273	26	53.12	2.09
22	*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
23	*KY672960.1*, Dengue virus 4 isolate DENV-4/China/YN/15DGR394 (2015), 2019, Sanger dideoxy	10661	276	26	52.63	2.09
24	*KX224312.2,* Dengue virus 4 isolate SG(EHI)D4/02990Y14, Singapore, 2017, Sanger dideoxy sequencing	10652	275	27	52.21	2.09
25	*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. The 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 Word-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_SLAB 000156, 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