JP Journal of Biostatistics

https://pphmjopenaccess.com
http://dx.doi.org/10.17654/0973514324017
Volume 24, Number 2, 2024, Pages 251-286 **P-ISSN: 0973-5143**

DNA WALKS IN VIRUS GENOMICS

A. Belinsky and G. A. Kouzaev*

Mach-3dP Inc., Burlington, ON
Canada
e-mail: designdigitalhealth@gmail.com

Department of Electronic Systems
Norwegian University of Science and Technology
Trondheim, Norway
e-mail: guennadi.kouzaev@ntnu.no
kouzaev@hotmail.com

Abstract

This paper studies published results in imaging and digital processing of virus RNAs (ribonucleic acid) using DNA (deoxyribonucleic acid) walks. The complicated nature and physicochemical properties of these nucleotide chains hinder the development of a universal method of numerical mapping and plotting of RNAs, and many algorithms that exist are reviewed here, including 2-D and 3-D DNA walks, walks in complex space, multi-dimensional dynamic representations of DNAs, etc. A detailed analysis is performed for a recently proposed query-walk algorithm and multi-level graphical representation of the

Received: November 6, 2023; Revised: December 22, 2023; Accepted: February 16, 2024
Keywords and phrases: virus RNAs, DNA walks, metric-based binary walk algorithm, ATG walk, SARS-CoV-2 virus, MERS-CoV virus, Dengue virus, Ebola virus.

*Corresponding author

How to cite this article: A. Belinsky and G. A. Kouzaev, DNA walks in virus genomics, JP Journal of Biostatistics 24(2) (2024), 251-286. http://dx.doi.org/10.17654/0973514324017

Published Online: April 20, 2024

traces of repeated patterns in RNA chains. They are represented by binary strings and compared with a sought query, calculating the Hamming distance in every comparison step. The coordinates of the found patterns or queries are defined, and a walk is composed of a set of consecutive numbers of these queries along the studied RNA. The primary attention of this review is paid to ATG triplets, which are starting nucleotides of codons (words) in most cases. As follows from the analyzed papers, the severe mutations of viruses touch the compactness of ATG curve sets of viruses and de-cluster the fractal dimension values of word-length distributions. The material of this review is helpful in the digital and visual studies of viruses.

1. Introduction

Viruses are tiny semi-life units carrying genetic information materials (RNAs or even DNAs) in a protein capsid covered by a lipid coat [1, 2]. They penetrate the cell wall and urge these cells to make more virus items, leading to severe illness or even bio-organism death. Although much attention had been paid by scientists before 2019, the tragic history of the SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) coronavirus pandemic and continued infection cases this year urge us to study further the nature of viruses and develop new treatment methods for virus-caused diseases and post-virus affections.

The RNAs of viruses are the sequences of base amino acids Adenine (A), Cytosine (C), Guanine (G), and Uracil (U). Due to some detection peculiarities, the Thymine's symbol (T) is used instead of Uracil in the RNA sequences. It does not influence the mathematical results of studying RNA's lines. This genetic information is represented in symbolic and graphical forms. Mathematical tools applied to RNA chains and giving quantitative and visualized information are very helpful in this case [3, 4]. Several database banks contain genetic information. For instance, among them are GenBank [5] and GISAID (Global Initiative on Sharing All Influenza Data) [6].

From the first look, the distributions of these symbols along the RNA chains are entirely stochastic, and it is used in multinomial theories of DNA sequences. Markov's models support the idea that the probability of a symbol appearing depends on the preceding nucleotide. Meanwhile, some more regularities have been found. There are short chains of introns (non-coding nucleotides) at the beginning of the virus RNAs. They do not carry vital genetic information but can enhance gene expression [7]. The RNA open reading frames [3] are composed of codons (coding triplets), which are started by AUG (ATG) set of nucleotides and ended by one of the following three combinations: UAA (TAA), UAG (TAG), or UGA (TGA). The triplets AUG (ATG) may also play independent coding roles. Long-range correlations of amino acids in RNAs are also known [8-10]. Many palindromes have been found in RNA chains [11]. All of these make the RNAs quasi-random.

Open reading frames are grouped into genes that may show relative stability during multiple generations of virus species. The message RNAs cleaned from introns define the synthesized proteins composed of 20 amino acids in various ways. The alterations of RNAs caused by chemical, physical agents, or rare mistakes of cellular machinery are called *mutations*, leading to protein changes.

The viruses can be divided by mutations into different branches and families, causing problems in the successful vaccinations of people and their disease treatments. Recognizing these mutations and predicting their consequences is an essential task for virology. It is reached by applying different algorithms of alignment of sequences, calculating the genetic distances between the RNA samples, and building phylogenetic trees of mutations [3].

There are many types of virus RNA variation. For instance:

- Point mutations: change of a base amino acid; adding a new one; nucleotide missing.

• Gross mutations: deletions of a part of RNA's sequence; insertions of extra-portion amino acids; rearrangements of RNA segments.

Remarkably, the most severe changes are with the point change of starting and stopping codons, reading frame length variations, and other gross mutations. The algorithms that can detect these variations are essential. For instance, these codes initially transform the RNA's alphabetical symbols into digital sequences called the *DNA walks* [10], which are evaluated further digitally and visually [12, 13].

This paper aims to review the progress in the development of DNA walk algorithms, applications of the digital signal processing methods to these walks, and new results obtained in the analysis of viruses, including SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), MERS-CoV (Middle-East Respiratory Syndrome-related Corona Virus), Dengue, and Ebola viruses using the introduced in [14-16] ATG tracing.

2. DNA Walks

Many algorithms allow the digitalization of literal representations of RNAs [17]. Among them are some DNA walks, which would enable not only the presentation of sequences in the form of curves [10] or 2-D patterns for visual analyses [18], but they can also provide quantitative information after applying signal processing algorithms, for instance. Each curve point is considered as a signal sample, and the obtained curve sample set can be processed numerically [12, 13].

One of the first results on the graphical representation of DNAs was published in [19]. Each mth nucleotide of a chain is represented by a vector:

$$G : \mathbf{g}_m(z) = -\mathbf{i} + \mathbf{j} - \mathbf{k}; \quad C : \mathbf{g}_m(z) = -\mathbf{i} - \mathbf{j} - \mathbf{k};$$

$$T : \mathbf{g}_m(z) = \mathbf{i} - \mathbf{j} - \mathbf{k}; \quad A : \mathbf{g}_m(z) = \mathbf{i} + \mathbf{j} - \mathbf{k}. \tag{1}$$

Each sequence point is calculated as $h(z_m) = \left| \sum_3 \mathbf{g}_m(z_m) \right|$ and a curve $h(z_m)$ is built, increasing the serial nucleotide number m. This technique was chosen because, otherwise, a nucleotide sequence trajectory should be created in 5-D space.

Difficulties in imaging genomes in multi-dimensional space urged the researchers to invent other ways in plane graphical representations [20-24].

For instance, Gates proposed to assign the following unit vectors for all four DNA nucleotides [20]:

$$\mathrm{G} : \mathbf{x}_0;\ \mathrm{C} : -\mathbf{x}_0;\ \mathrm{T} : \mathbf{y}_0;\ \mathrm{A} : -\mathbf{y}_0. \tag{2}$$

Then, the walks can be imaged on a plane. Unfortunately, these trajectories suffer from degeneracy because some of the combinations of nucleotides will have the same graphical representations. For instance, a genome fragment CCG gives "motion" on the same line. These walks can be used in practical imaging, as it follows from [25], for instance, to study the genomic alterations in lung cancer tissues. For example, Nandy's walks [22] were applied to 22 samples of Zika virus in [26]. It was found essential differences between the African, Asian, and American samples. At the same time, the patterns of this virus have some similarities with those of Dengue.

Degeneracy of DNA trajectories was decreased in modified Gates' walks in [27] by coding nucleotides on a plane in the following way:

$$\mathrm{G} : \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right);\ \mathrm{C} : \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right);\ \mathrm{T} : \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right);\ \mathrm{A} : \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right). \tag{3}$$

According to this method, the short-plotted patterns can be distinguished visually. Later, a method to obtain phylogenic trees was developed using this representation of viruses and other species' genomes [28]. For this purpose, the probabilistic distributions of each symbol are initially calculated for the compared sequences. Then, the dissimilarity measure is computed using Kullback-Leibler divergence or relative entropy [29]. Its value allowed calculation of the phylogenetic trees of 53 samples of the RNAs of Tomato

Yellow Leaf Curl Viruses and other non-virus species. It is essential that the comparisons of RNA were performed using a non-alignment method to avoid the solution of nondeterministic polynomial (NP)-problems in point-to-point comparisons of sequences.

Genomic walks can be mapped on a complex plane [10, 18, 30-32]. For a *k*th point of a sequence, the following values are defined [18]:

$$\mathrm{G}: z_k = -1;\ \mathrm{C}: z_k = -\sqrt{1};\ \mathrm{T}: z_k = \sqrt{-1};\ \mathrm{A}: z_k = +1. \tag{4}$$

Moving along the sequence, at each point k, a cumulative sum is calculated:

$$s(k) = \sum_{1}^{k} z_i. \tag{5}$$

Then, this walk is a dependence $k = F(s_k)$, i.e., a curve in a 3-D space $(k, \mathrm{Re}(s_k), \mathrm{Im}(s_k))$ that can be processed further by statistical methods to find, for instance, correlation properties in a studied sequence.

Genomic walks can be built in a 3-D complex space if quaternions code nucleotides [33]. In this case,

$$\mathrm{G}: -i - j - k;\ \mathrm{C}: i - j - k;\ \mathrm{T}: -i + j - k;\ \mathrm{A}: i + j + k, \tag{6}$$

where

$$i^2 = j^2 = k^2 = 1,\ i \cdot j = -j \cdot i = -1,$$
$$j \cdot k = k \cdot j = i,\ k \cdot i = -i \cdot k = j. \tag{7}$$

It was shown that the quaternionic representation of nucleotides in RNAs allows better detection of hidden periodicity of acid distributions in sequences.

Another way to avoid degeneracy is shown in [34], where two vectors code each nucleotide. The dual-vector (DV) curve technique is without degeneracy and loss of information, and it shows effective visualization,

according to the authors of the mentioned paper. This method was used to inspect and compare DNA sequences and their mutations.

Reviews on many 2- and 3-D simple walks developed before 2013 were published in [35-37]. The main conclusion of these results is that RNA/DNA sequences have very complicated properties, which are complex to be described by a single descriptor or mathematical method. Different models are needed for a more effective description of DNAs. The introduced walks should not distort the information in the sequences intended for visual analyses.

Further results on genomic walks are scattered in many journals, majoring in biology, physics, computer science, signal processing, etc. Many recent of them are reviewed below. For instance, the nucleotide sequences can be mapped in the polar system of coordinates, and separate circles or parts of unit circles are for the base nucleotides [38-40]. These diagrams allow visual inspections and numerical calculations to build different species' similarity dendrograms [39, 40].

Degeneracy of DNA walks can be avoided if each base is represented by a trigonometric function of a unit magnitude, the argument of which depends on the nucleotide [41]. Besides, these sinusoidal functions are placed on different levels along the *y*-axis. Points on these curves show the nucleotides, while the sequential number of a base is listed along the analyzed sequence. In the mentioned paper, the Euclidian distance values between the points of different sequences were calculated, and phylogenetic trees of the studied species were given. Several years ago, it was messaged that a software tool, Squiggle, was being developed for mapping DNAs using several known 2-D walks [42].

In addition to the pure mathematical representation of RNAs and DNAs, they can be described by walks coupled with the physicochemical properties of bases [43-46]. For instance, in [44], the maps $(X : \mathrm{pK}_a(\mathrm{COOH})$ and $Y : \mathrm{pK}_a(\mathrm{NH}_3^+))$ are composed along the studied sequences as zig-zag

curves. Then, similarity maps of sequences can be composed by comparing the mentioned graphs.

Grouping nucleotides into purine (A, G) and pyrimidine (C, T) allows the creation of 1-D walks on a plane assigning to purines (-1) step along the y-axis, $(+1)$ if pyrimidine is registered in a sequence. Then, excepting imaging, some numerical methods can be applied to this set of binary values to find, for instance, the global correlation of these nucleotide pairs in genomic sequences, namely, in intron-containing genes [47, 48].

A 3-D space modification of this idea is in [49], where a 4×4 matrix represents 16 DNA pairs, and a trajectory in 3-D space is calculated moving along a sequence from one pair to another couple of amino acids numbered in the following way:

$$\begin{aligned}
&\text{AG} : (0, 0, 0);\ \text{GA} : (0, 1, 0);\ \text{CT} : (0, 2, 0);\ \text{TC} : (0, 3, 0);\\
&\text{AC} : (1, 0, 0);\ \text{CA} : (1, 1, 0);\ \text{GT} : (1, 2, 0);\ \text{TG} : (1, 3, 0);\\
&\text{AT} : (2, 0, 0);\ \text{TA} : (2, 1, 0);\ \text{CG} : (2, 2, 0);\ \text{GC} : (2, 3, 0);\\
&\text{AA} : (3, 0, 0);\ \text{CC} : (3, 1, 0);\ \text{GG} : (3, 2, 0);\ \text{TT} : (3, 3, 0). \qquad (8)
\end{aligned}$$

The numbers in (8) are associated with the coordinates in the Cartesian system, and trajectories in 3-D space are built. These traces were compared with each other, calculating the cumulative distance between the points of compared curves in 3-D space, and it allowed recognition of significant similarities of the studied DNA sequences of several species.

Ten paired nucleotides are considered in [50] in a specific way to create 3-D walks of different species, including several tens of hantavirus sequences. The phylogenetic diagrams were built by calculating the Euclidean distance values between the studied numerically expressed sequences.

The bases can be grouped into triplets, and in [51], a distribution graph of 64 triplets of proteins loaded by their molecular weight is given as a cube with triplet vertices. A trajectory in a 3-D space is built using these cubes

and computer calculations. These models allowed computing the Euclidean distances between the compared sequences of H1N1 viruses and building phylogenetic trees of them. It was noticed that due to the possible degeneracy of the representation used, the method is preferable for analyses of long RNA chains, as the authors of the cited paper noted.

An interesting method for decreasing degeneracy was proposed in [52] called the *dynamic representation* of DNAs. It is based on the use of analogies with classical mechanics. Analogies are very common and effective in describing complicated objects and effects in science. In the mentioned paper, building a low-degeneracy DNA representation starts with mapping nucleotide trajectories by the unit vectors in the notations of Nandy [22, 26]:

$$\mathrm{G} : \mathbf{x}_0;\ \mathrm{C} : -\mathbf{y}_0;\ \mathrm{T} : \mathbf{y}_0;\ \mathrm{A} : -\mathbf{x}_0. \tag{9}$$

Then, a mass $m = 1$ is assigned to each point of this map. If motion is n-times degenerated, then the point weight is increased n-times. The trajectory is considered a rigid body of N material points. The coordinate of the center of mass is calculated according to simple formulas:

$$\mu_x = \frac{\sum\limits_i m_i x_i}{\sum\limits_i m_i};\quad \mu_y = \frac{\sum\limits_i m_i y_i}{\sum\limits_i m_i}, \tag{10}$$

where x_i and y_i are the coordinates of a genomic trajectory in Nandy's system (9).

Because the mass points are distributed nonhomogeneously on the plane, the tensor of inertia $\tilde{I}$ describes them. After that, the moments of inertia are computed as the eigenvalues $I_{1,2}$ of this tensor $\tilde{I}$ calculating this matrix determinant:

$$\det\begin{vmatrix} I_{xx} - I & I_{xy} \\ I_{yx} & I_{yy} - I \end{vmatrix} = 0, \tag{11}$$

where $I_{xy} = I_{yx} = -\sum_i m_i x_i^{\mu} y_i^{\mu}$, $I_{xx} = \sum_i m_i (y_i^{\mu})^2$, $I_{yy} = \sum_i m_i (x_i^{\mu})^2$, x_i^{μ} and y_i^{μ} are the coordinates of the mass m_i in the Cartesian system that originated at the center of mass. The two eigenvalues $I_{1,2}$ of equation (11) and the coordinates $\mu_{x,y}$ were chosen as the numerical descriptors (identifications) of genomic sequences.

Many genes were studied using this method. It was found that these descriptors of genomic sequences remove most of the cases of degeneration of 2-D methods known to the date of writing in [21]. More applications of this method are from [53, 54], where hundreds of samples of influenza viruses are studied.

Later, this 2-D method was applied to the study of fourteen Zika genomes [55]. The authors stated that applications of this non-alignment method of comparison allowed clear separation of genome clusters of the African and Asian-American Zika virus samples.

The mentioned 2-D algorithm was combined with a machine learning algorithm, and the prediction of novel subtypes of Influenza A reached 90% using these techniques [56].

This method was modified further in [57]. In this cited work, the nucleotides were represented by the four vectors of the unit length:

$$\mathrm{G} : (1, 0, 1); \mathrm{C} : (0, 1, 1); \mathrm{T} : (0, -1, 1); \mathrm{A} : (-1, 0, 1). \tag{12}$$

To each point in this 3-D space, a unit mass is assigned, and their coordinates are calculated similarly to (10). The tensor of inertia is a 3×3 matrix having three eigenvalues and vectors. The eigenvectors are orthonormal and can be used as a new coordinate system. In the cited work, the angles between the planes of the coordinate system x, y, z and the mentioned planes are the elements of a descriptor of a genomic sequence. Then, the full 3-D descriptor consists of the eigenvalues of the inertia tensor $\tilde{I}$ and the said angles. It allows an effective procedure for

similarity/dissimilarity analyses of genomic sequences. The 3-D representation used in the cited work avoids the typical degeneracy of 2-D graphs, and it was confirmed by analyzing the genes of many species in [57, 58], including influenza A and SARS-CoV-2 viruses and their evolution. In some cases, especially in the analyses of genomic sequences that are close to each other, the further developments of this method are fruitful.

The walks representing the sequences in 4-D space are proposed in [59, 60]. In this algorithm, the bases are coded in the following manner:

$$A : (1, 0, 0, 0); C : (1, 0, 0, 1); G : (1, 0, 1, 1); T : (2, 0, 1, 1). \quad (13)$$

It was established that the descriptors of the 4-D method are very sensitive, and they can even recognize a difference in one nucleobase without comparing sequences by nucleotides, i.e., in a non-alignment style. Notably, it was found that SARS-CoV-2 may have originated from bats or pangolins due to the proximity of their descriptors. One of the authors' conclusions of the cited method is the possible development of predictive models for virus evolution [59], using, for instance, machine learning algorithms and other artificial intelligence (AI) methods. Similarly, the proteins can be analyzed using 20-D representations of nucleotides [61].

Digital mapping and DNA walks generally allow the detection of codons, introns, and genes, the discovery of hidden RNA periodicity, and the calculation of phylogenetic distances between the genomic sequences. A fruitful approach is processing genomic sequences using AI software tools and computers equipped with AI accelerators [62]. It is required to use relatively sophisticated mathematical methods, for instance, from the signal processing area [12, 13].

Our analysis shows that one of the most developed methods in studying genomes by DNA walks is the dynamic representation of DNA and proteins [52-61]. It allows their visualization and the development of their phylogenetic trees in an alignment-free manner, tracing virus evolution and

predicting virus mutations connected with artificial intelligence tools. Nevertheless, the complexity of the virus world, its diversity, and its high mutation rate require further development and study of different mathematical models of virus genomes [35, 37, 58].

In complete RNAs, the ones of the repetitive patterns [63] therein are their codon's starting and ending triplets. They compose a scheme or "skeleton" of an RNA. In [14-16], it was proposed to study the ATG distributions along these RNAs.

The search algorithm used in the mentioned papers calculates the positions of any repeated patterns in RNA sequences, including separate nucleotides. These trajectories composed of consecutive numbers of patterns (ATG, A, C, G, and T) in sequences are found not twisted firmly compared to many known RNA walk curves, and they are easier to analyze visually and numerically. They allow highlighting genes and embedding hyperlinks with the gene's names on a single plot. This approach, reviewed below, is being applied to many samples of several viruses, including SARS-CoV-2, MERS-CoV, Dengue, and Ebola ones, and it allowed us to estimate the proximity of genomes using alignment and alignment-free techniques.

3. Research Methodology in [14-16]

3.1. Metric-based repeated-pattern walk algorithm

Repeated patterns in RNAs are composed of several or a single nucleotide. The most known repeated triplets are the starting and stopping codons. Discovery of them is an NP problem. Its solution time increases exponentially with the sequence length. The search time decrease is significant, especially for long sequences.

Initially, a character query pattern $\{A\}$ is defined as a certain length n, that particularly can be an ATG triplet. This pattern is sought in a genomic sequence $\{B\}$ of the length N. Then, a search algorithm compares the pattern $\{A\}$ with the corresponding set of nucleotides taken from the sequence $\{B\}$. Afterward, the comparison is shifted to one amino acid symbol. Other parts

of this algorithm identify the positions of the pattern, requiring the complete identity of query and sequence symbols during n comparisons. The number of these operations grows with the sequence and query sizes. It is known that classical computers do not resolve the NP problems, but some decreases in simulation time are very fruitful [63].

For instance, the search algorithms working with characters are slower by 1.43-2.37 times than those processing the binary variables [64, 65]. Moreover, even the computer arithmetic logic units can be re-designed to accelerate the fulfillment of the frequently repeated patterns of binary operations [62].

In [14], all RNA's characters are expressed by binaries using a Matlab library function dec2bin (character) [66] before all operations, taking into account that the UTF-8 (Unicode Transformation Format-8 bit); encoding all 1,112,064 valid character code points [67]. Because the binary sequences now represent the RNA chains and queries, they can be compared using a suitable binary technique, such as a metric distance between the binary RNA chains and queries.

Several metric types are known in coding and big data [68-73]. The metric estimates are applied in cluster analysis for grouping nucleotide or protein distribution patterns. Metrix can help classify the virus RNAs [74, 75]. Particularly, the Hamming one [68] is used in [14-16] for calculation of virus ATG walks. A detailed description of the developed algorithm is given in [14].

Particularly, the codon's start-up ATG triplet is used as a query $\{B\}$ in [14], and it calculates the position x_i of these triplets. The trajectories composed of these points along a sequence are the ATG walks.

Additionally, this algorithm calculates the word length distributions $l_{i,i+1}^{(\mathrm{ATG})} = x_{i+1} - x_i$. In the reviewed works [14-16], a 'word' is a nucleotide sequence starting with an ATG triplet and all symbols up to the next ATG-one (Figure 1).

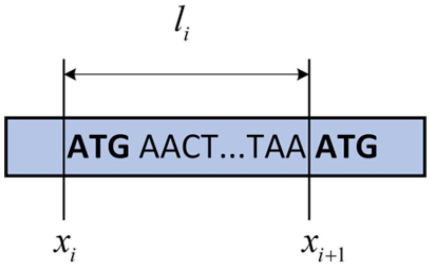
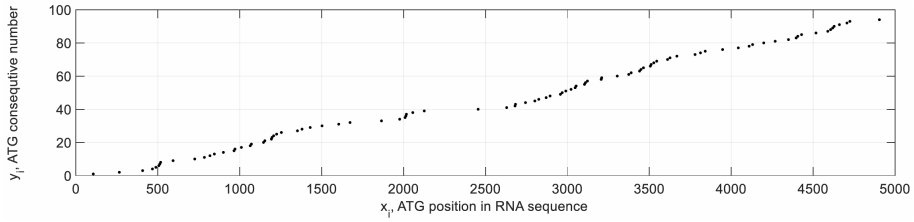


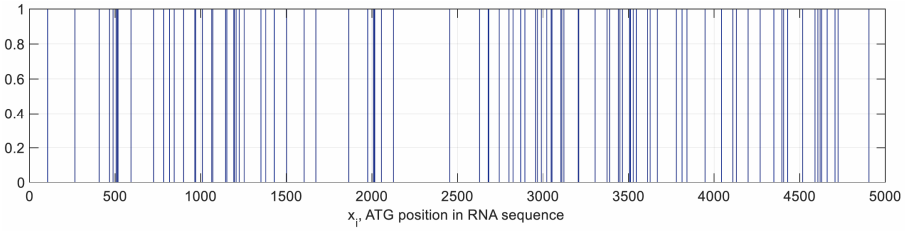
Figure 1. A word of the length l_i.

This code was realized in the Matlab environment [66] and applied to many available virus complete genome sequences. The obtained ATG positions were compared with those found in the studied genomes to verify these calculations.

An example of an ATG walk built using this algorithm is given in Figure 2(A). In addition, the coordinates of ATG points are shown by vertical bars distributed along sequences in a fractal manner (Figure 2(B)) that was confirmed by computations [14].



(A)



(B)

Figure 2. Positions of ATG triplets along the genome sequence of a SARS-CoV-2 virus MN988668.1 (GenBank) are given for the first 5,000 nucleotides provided by points (A) and vertical lines in diagram (B).

3.2. Multi-scale mapping of RNA sequences

The viral RNAs, consisting of thousands of nucleotides, are challenging to analyze, and many visualizing methods are used. One of them is the ATG walk considered above. Let us define an ATG walk as a series of coordinates of its first triplet symbol along an RNA (Figure 1). Taking repetitive distributions of ATG triplets, these coordinates are found by using the mentioned search-pattern algorithm. It is supposed that these walks allow us to detect the variations of codon's lengths of RNAs coupled to rather severe mutations of viruses.

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 repetitive pattern, allowing the calculation of distribution curves for each nucleotide similar to ATG ones. These curves can be viewed as the first level of spatial detailing of RNAs. The words in our definitions compose the second level. They 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 RNA organization will be helpful in the visual analyses of genomes. One of the ways is shown in Figure 3.

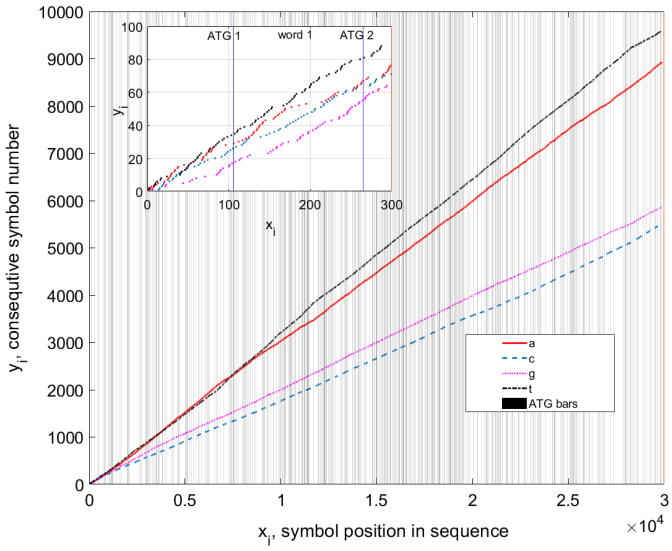


Figure 3. Two-scale study results of a SARS-CoV-2 virus MN988668.1 (GenBank). In the inlet, these symbols are pointed inside the words for the first 300 nucleotides.

Here, positions of symbol "A" of ATG triplets in a sequence are given by vertical lines (second level of detailing). Words take spaces between these vertical bars (see Figure 2(B)). They are filled by numbered nucleotides (points of a different color) corresponding to the first level of RNA detailing. It allows for distinguishing the nucleotides at the beginning of coordinates where visual clutter is seen (inlet in Figure 3). It is seen that the shapes of ATG (Figure 2(A)) and nucleotide curves (Figure 3) are close to lines [15], and this can be explained by physical modeling of the staking of these molecules [76, 77]. For instance, in [15], linear functions approximated such curves.

These ATG distributions have repeating motifs on different geometry scale levels (see bars in Figures 2(B) and 3), i.e., fractality, confirmed in [14] by studying genomic sequences. Presumably, the ATG triplets are distributed along with the RNA sequences of studied viruses according to the random Cantor multifractal set rule [78].

The next level of hierarchical RNA organization is with genes. For instance, in GenBank [5], 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 the case of Figure 3, this diagram can be attached to a two-scale plot considered above. Another solution is to equip these figures with gene interactive hyperlinks.

Thus, the pattern search algorithm developed in [13] allows building combined plotting of hierarchical organization of the RNAs of viruses. It can also be used for the analyses of more complex protein structures. Unlike many genomic walks, it produces spatially simple curves, as seen in Figures 2 and 3, that can be analyzed visually and quantitatively.

3.3. Calculation of fractal properties of ATG distributions

Fractality of DNA/RNA sequences has been studied in many works [10, 18, 79-93]. In this case, 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.

The sequences can be multifractals because the large-size genomic data are often patterned, and each pattern can have its fractal dimension [88]. 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 [94-101].

Discovering the fractality of genomic sequences is preceded by their numerical representation, for instance, by walks of different types [18, 31, 51, 79, 84, 88]. Then, each 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. For instance, the calculations can be applied directly to the distribution of ATG consecutive numbers (Figure 2). Still, it gives $d_F \approx 1$ for the analyzed RNAs, i.e., the dependence $y_i(x_i)$ is close to the linear one. Then, these

calculations are not practical in analyses due to their weak sensitivity. Instead, the word-length distributions $l_{i,i+1}^{(\mathrm{ATG})}$ (inter-ATG distance distributions) along the RNA sequences (see Subsection 3.1, Figures 2 and 3) are used. For instance, the inter-nucleotide distances were considered earlier in [102]. A particular distribution of the word lengths is shown in [14] by bars whose height is proportional to the word length. Then, the algorithms usually applied to the sampled signals can be used to compute the statistical properties of these inter-ATG distributions. The fractal dimension of these word-length distributions was calculated with the software package FracLab 2.2 [101]. This code provides the results with reasonable accuracy if its default parameters are used.

4. Review of the Main Results Obtained in the Study of SARS-CoV-2, MERS-CoV, Dengue, and Ebola Viruses Using ATG Walks

4.1. ATG walks

Numerous papers have been published dedicated to the genomics of these dangerous viruses. Here, only a few references are given that can be followed to be introduced to this topic [103-112].

The mentioned viruses were studied by computing the ATG walks and comparing their statistical characteristics [14-16]. The complete genomic sequences were taken from genetic databases GenBank [5] and GISAID [6], and the list of these sequences was given in the appendices of [14]. Some results of this study are shown below.

Figure 4(A) illustrates the distributions (in lines) of ATG triplets of complete genome sequences for thirty-six SARS-CoV-2 arbitrary-chosen virus samples registered during 2020-2023 and a bat-corona sample (hCoV-19/bat/Cambodia/RShSTT182/2010, GISAID). Figure 4(B) shows the ATG distributions for 20 MERS viruses. ATG trajectories of 25 Dengue RNA sequences are given in Figure 4(C). The ATG distributions of 15 Ebola sequences are shown in Figure 4(D).

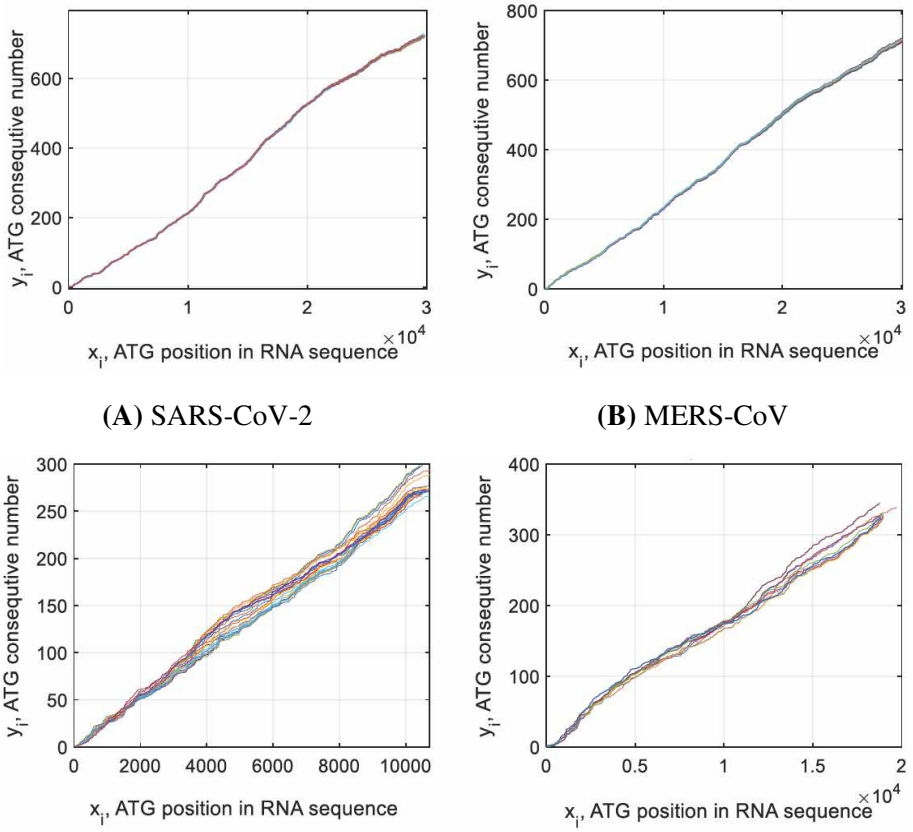


Figure 4. (A) Distributions of ATG triplets of 36 SARS-CoV-2 and one bat-corona (Cambodia, 2010) complete RNA sequences. (B) Distributions of 20 ATG triplets of MERS complete RNA sequences. (C) Distributions of ATG triplets of 25 Dengue complete RNA sequences. (D) Distributions of 15 ATG triplets of Ebola complete RNA sequences.

Considering Figure 4(A) on SARS-CoV-2 viruses, it can be seen that all triplet curves have relatively compact localizations despite the viruses being of different clades and lines. For instance, the relative difference $\delta y_{1,37}(x_i = 29506) = 100\% \cdot 2\Delta y_{1,37}/(y_1 + y_{37})$ of these SARS-CoV-2 curves 1-37 (Figure 4(A)) is estimated at $x_i = 29506$ around only 1.5%. It confirms the conclusions of many specialists [107] that no new recombined

strains have appeared up to that time (the beginning of 2023) despite many mutations found to date, including the Omicron lineage.

A detailed study of these SARS-CoV-2 samples [14] shows that each considered sequence has individual ATG distribution. It means some strong mutations are combined with the joint variations of word content, word length, and the number of these words in sequences. Due to the variations of the non-coding nucleotides, the ATG curves of complete sequences can also be shifted. Other mutations with only word content may exist. However, the ATG walks cannot see them, and the single-symbol distributions considered below will help us detect these virus modifications (see Subsection 3.2).

Visual comparison of a bat virus (hCoV-19/bat/Cambodia/RShSTT182/2010) ATG trajectory with a human's SARS-CoV-2 set of ATG curves [14] shows the proximity of all these walks to each other, which is in accordance with a conclusion of [60] that SARS-CoV-2 virus may originate from bats.

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 data clustering, among others [3, 37, 75, 86, 113, 114]. Researching the 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.

In addition to calculating cumulative divergence of the curves at certain points of a set of ATG curves $\delta y(x_i)$, in [15], a simplified algorithm for quantitative comparing ATG distributions of different virus samples is used (Figure 5). Each numbered ATG triplet (y_i) has its position along a sequence (x_i). Due to mutations, the length of some coding words varied together with the coordinate (x_i) of a triplet.

The difference (deviation) between the coordinates (x_i) of ATG triplets of the same numbers (y_i) in the compared sequences was calculated. 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 comparing geometrical data has some limitations. Therefore, if a sequence has several ATG triplets fewer than the number of ATG-ones in the referenced genome, then the ATGs of the last RNA are excluded from comparisons. Still, it allows for obtaining some information on mutations of viruses straightforwardly and effectively, which will be seen below.

This 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. Several virus samples from GenBank and GISAID have been studied this way in [15]. Some results of the comparisons are given in Figure 5.

The ordinate axis Δx in these plots shows the deviation of the coordinates x_i of ATG triplets from the ATG coordinates of the reference sequence. As a rule, due to the different number of non-coding nucleotides at the beginning of complete RNA sequences, the curves in Figure 5 have constant biasing along the Δx axis.

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

In some studied samples (here and in [15, 16]), perturbations are near the end of the *orf*1*ab* gene, as is seen using a graphical tool of GenBank [5]. The ATG perturbations could generally occur in any RNA's part, considering the random nature of mutations [14, 15]. Relative deviation $|\Delta x_i|/N$ did not

exceed 1-2% for the compared viruses. Although this deviation is mathematically tiny, it may have severe consequences in the biological sense.

It is seen that these difference curves (Figure 5) can be 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 5 and [15, 16]). The origin of this is unknown, but it was not coupled with the lineages of viruses and their clades in this research, handling only a few virus samples. Other virus genomes can be studied in the same way.

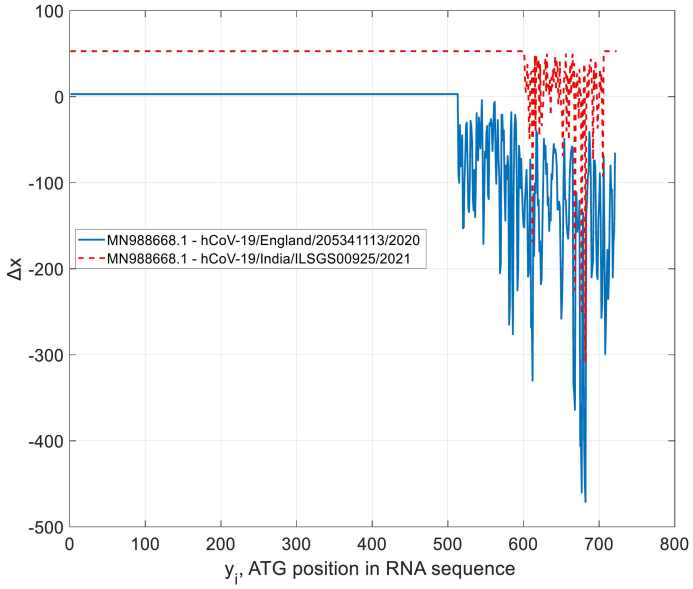


Figure 5. Deviation of ATG coordinates in RNAs of two SARS-CoV-2 viruses (GISAID) relative to the reference virus RNA MN988668.1 (GenBank).

Similarly, the ATG distributions of twenty Middle East Respiratory Syndrome-related (MERS) viruses [108, 109] are built (Figure 4(B)). On average, the studied MERS RNAs have fewer number ATG triplets (701-722 against 719-726) and longer words compared to the SARS-CoV-2

sequences [14]. They show (Figure 4(B)) increased compactness of MERS ATG distributions, similar to Figure 4(A). The relative difference $\delta_{1,20}y(x_i = 29478)$ of these traces is estimated at around 2%.

These two studied coronaviruses (MERS-CoV and SARS-CoV-2) demonstrate relatively strong stability of their ATG distributions towards severe mutations that lead to the variation of codon positions, word lengths, and the number of words. It corresponds to many scientists' conclusions during three years of virus investigations, including ATG walk observations [14-16, 107].

Unlike the studied above two coronaviruses, the Dengue virus [110, 111] has five genotypes (DENV 1-5) and around 47 strains. Its RNAs are comprised of more than 10,000 nucleotides and about 270-300 ATGs. A consolidated plot of 25 ATG curves of the arbitrary-chosen Dengue virus samples (GenBank) is shown in Figure 4(C).

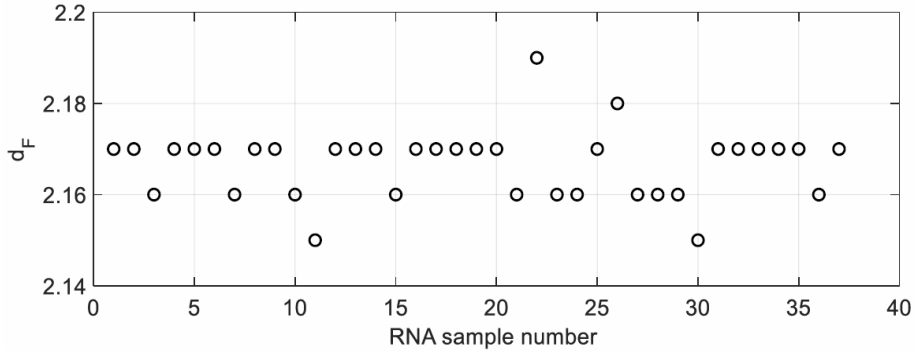
A substantial divergence of these trajectories is seen. This new finding corresponds to this virus's increased mutation rate found by other researchers. For instance, the relative difference $\delta y_{1,25}$ estimated at $x_i = 10{,}000$ which is 14.1%. It means some RNA mutations are coupled with the change in the length of coding words.

Similar to the Dengue virus, the Ebola one is highly variable [112] (see [14] for a list of studied viruses). Four strains of the Ebola virus are known worldwide, and more families can be formed in the future. This illness causes 25%-90% of the death rate for the infected people.

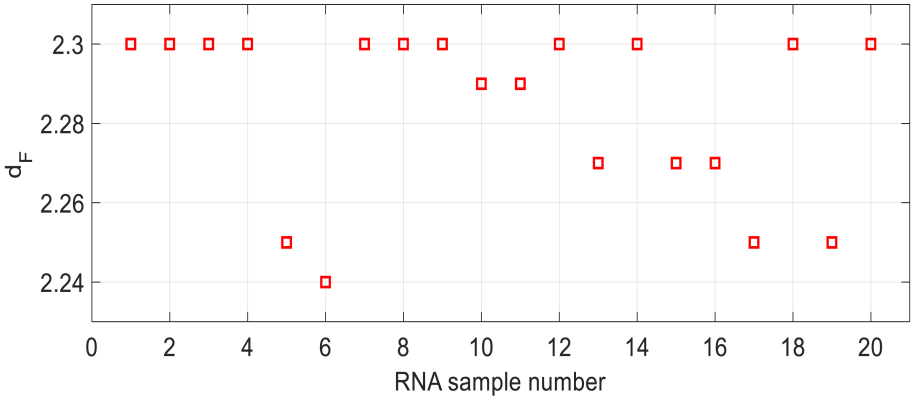
The Ebola virus RNA comprises about 19,000 nucleotides and three hundred ATG triplets. The ATG walks for 15 RNA samples of four strains found in GenBank are shown in Figure 4(D). It is found that the relative difference for 15 samples $\delta y_{1,15}(x_i = 18639) = 9.49\%$ is essentially more significant than the same parameter of coronaviruses.

4.2. Fractal properties of word-length distributions of SARS-CoV-2, MERS, Dengue, and Ebola RNAs

In [14], after applying the FracLab tool (see Subsection 3.3 and [101]), it was discovered that all studied genomic sequences of the SARS-CoV-2, and MERS-CoV, Dengue, and Ebola viruses have fractality measured by its dimension values d_F of word-length distributions of genomic sequences (Figures 6 and 7).

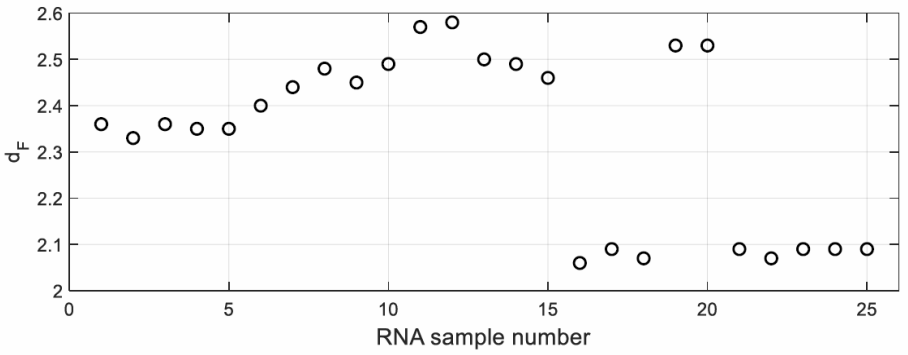


(A) SARS-CoV-2

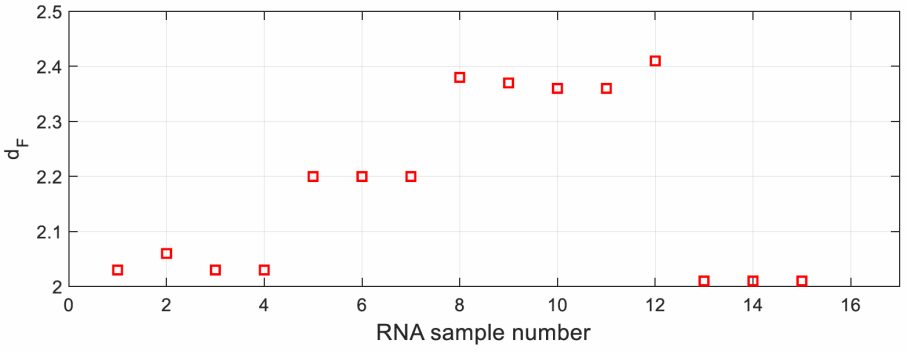


(B) MERS-CoV

Figure 6. Fractal dimensions d_F of word-length $l_{i,i+1}^{(\mathrm{ATG})}$ distributions of 36 complete genome sequences of the SARS-CoV-2, one bat-corona (A), and 20 complete genome sequences of the MERS-CoV (B) viruses.



(A) Dengue



(B) Ebola

Figure 7. Fractal dimensions d_F of word-length $l_{i,i+1}^{(\mathrm{ATG})}$ distributions of 25 sequences of the Dengue (A) and 15 samples of Ebola (B) viruses.

A visual analysis of these figures indicates relatively compact distributions of the parameters d_F for SARS-CoV-2 and MERS coronaviruses (Figure 6). The fractal parameters d_F of the Dengue and Ebola viruses are distributed less compactly (Figure 7).

The quantitative estimates of deviations of fractal parameters d_F of each studied set of viruses were calculated in [14], and it is $\delta d_F = 100\% \cdot 2(d_{F_{\max}} - d_{F_{\min}})/(d_{F_{\max}} + d_{F_{\min}})$. Their values vary in the limits of

1.8%-25%, being essentially more significant for the Dengue and Ebola viruses.

5. Discussion

For a summary of the discussion in [14], let us compose two tables (Tables 1 and 2) to compare the divergence of ATG curves and deviation of parameters of fractal properties of the word-length distributions of the genomic sequences of the four studied viruses.

Table 1. Maximal normalized difference δy of ATG trajectories

Virus name	SARS-CoV-2	MERS-CoV	Dengue	Ebola
Number of samples	37	20	25	15
δy, maximal normalized difference, %	1.5	2	14.1	9.6

Table 2. Maximal normalized deviation of fractal parameters δd_F

Virus name	SARS-CoV-2	MERS-CoV	Dengue	Ebola
Number of samples	37	20	25	15
δd_F, maximal normalized difference, %	1.8	2	22	25

It is seen that the values of the normalized difference (δy) of ATG curves and their normalized deviation of fractality parameter δd_F calculated for word distributions correlate with the mutability of viruses, which is significant for the Dengue and Ebola ones. The method applied in [14-16] for comparison of genomic sequences relates to the free-alignment techniques and is not coupled with point-to-point calculations of the genomic distances.

6. Conclusions

The research on the RNAs and DNAs of viruses and cellular organisms is a highly complicated process because of the many nucleotides of these

organic polymers and the unclear mechanisms of their synthesis. Although many mathematical tools have been developed, new studies are exciting and can be fruitful. Among them are the DNA walks, allowing graphical representations of RNA chains and their mathematical processing.

In this review, *more than* 100 *papers* have been analyzed on the methods of digital mapping and imaging of virus RNAs. Unfortunately, no universal tools allowing complete studies of all properties of DNAs are known. They often give complicated 2-D or 3-D maps or projections of curves from 4-D space. Some of them may distort RNA information by the chosen method of digitalization and mapping. Then, the research on better representation of DNAs is still in high demand, considering modern progress in computer graphics. The primary review attention was paid to such DNA walks, allowing the quantitative and qualitative studies of RNAs, as proposed in 2007, dynamic representations of DNAs [52].

A new tracing method was given in 2021 consisting of consecutive mapping of positions of the repetitive patterns along RNA chains [14-16]. These positions are sought using calculations of Hamming distance between the binary-expressed patterns (queries) and studied sequences. It allows representations of 1-D traces of all four nucleotides, other repeated patterns along a studied RNA, and information on the inter-pattern distance distributions and genes on one image. As the repetitive patterns, the ATG codons (starting triplets) and separate nucleotides were chosen. These trajectories (pattern walks) are mapped all together, and no essential intertwining is found in virus RNAs that is convenient for visual analysis of all mutations.

These techniques were applied to observe 37 SARS-CoV-2 sequences, 20 samples of MERS coronavirus, 25 RNAs of the Dengue one, and 15 sequences of the Ebola virus RNAs. It was found that the instability of viruses and their tendency to separate into different families are with the divergence of ATG walks and deviation of fractal dimension values calculated for the genomic word-length (inter-ATG) distributions, which is typical for the last two viruses. Unlike them, the SARS-CoV-2 and MERS

coronaviruses demonstrated the increased clustering of their ATG walks and stability of the mentioned fractal parameters according to the data analyzed up to the beginning of this (2023) year.

The results and methods of [14-16] are interesting not only in the study of virus RNAs but also in the research of mammalian DNAs, where the length of genes is related to the evolution of organisms [115]. An individual bio-object can show a gene-length imbalance with aging, as was established in [116]. The modeling of DNAs by using the reviewed methods can be effective in research and developing new drugs and gene treatments.

Acknowledgement

The authors thank the anonymous referees for their valuable suggestions and constructive criticisms which improved the presentation of the paper.

References

[1] H. Fletcher and I. Hickey, Genetics, 4th ed., Garland Science, 2013.

[2] G. Meister, RNA Biology: An Introduction, Wiley-VCH, 2011.

[3] C. Nello and M. Hahn, Introduction to Computational Genomics: A Case Studies Approach, University Press Cambridge, 2012.

[4] A. Pinho, S. Garcia, D. Pratas and P. J. S. G. Ferreira, DNA sequences at a glance, Plos One 8 (2013), e79922(1-11).

[5] GenBank® [www.ncbi.nlm.nih.gov/genbank].

[6] Global Initiative on Sharing All Influenza Data (GISAID) [www.gisaid.org].

[7] J. Blayney et al., Super-enhancers include classical enhancers and facilitators to fully activate gene expression, Cell 186 (2023), 5826-5839.

[8] W. Li, T. Marr T and K. Kaneko, Understanding long-range correlations in DNA sequences, Phys. D 75 (1994), 392-416.

[9] G. Villani, Affinity and correlation in DNA, Multidisciplinary Sci. J. 5 (2022), 214-231.

[10] J. Berger, S. Mitra, M. Carli and A. Neri, Visualization and analysis of DNA sequences using DNA walks, J. Franklin Inst. 341 (2004), 37-53.

[11] M. Tibatan and M. Sarısaman, Unitary structure of palindromes in DNA, Biosystems 211 (2022), 104565(1-8).

[12] P. Vaidyanathan, Genomics and proteomics: a signal processor's tour, IEEE Circ. Syst. Mag. 4 (2004), 7-29.

[13] J. Lorenzo-Ginori, A. Rodríguez-Fuentes, R. Ábalo and R. S. Rodríguez, Digital signal processing in the analysis of genomic sequences, Current Bioinformatics 4 (2009), 28-40.

[14] A. Belinsky and G. Kouzaev, Visual and quantitative analyses of virus genomic sequences using a metric-based algorithm, WSEAS Trans. Circ. Syst. 21 (2022), 323-348.

[15] A. Belinsky and G. Kouzaev, Geometrical study of virus RNA sequences, BioRxiv preprint: 2021.09.06.459135. https://doi.org/10.1101/2021.09.06.459135; Europe PMC: PPR: PPR391263.

[16] G. Kouzaev, The geometry of ATG-walks of the Omicron SARS-CoV-2 virus RNAs, BioRxiv preprint: https://doi.org/10.1101/2021.12.20.473613; Europe PMC: PPR: PPR435860.

[17] H. Kwan and S. Arniker, Numerical representation of DNA sequences, Proc. 2009 IEEE Int. Conf., Electro/Information Technology, Windsor, ON, Canada, 2009, pp. 307-310.

[18] C. Cattani, Complex representation of DNA sequences, M. Elloumi et al., eds., Bioinformatics Research and Development, BIRD 2008, Communications in Computer and Information Science, Vol. 13, Springer, 2008, pp. 528-537.

[19] E. Hamori and J. Raskin, Curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem. 258 (1983), 1318-1327.

[20] M. Gates, Simpler DNA sequence representations, Nature 316 (1985), 219.

[21] R. Voss, Evolution of long-range fractal correlations and $1/f$ noise in DNA sequences, Phys. Rev. Lett. 68 (1992), 3805-3808.

[22] A. Nandy, A new graphical representation and analysis of DNA sequence structure, I. Methodology and applications to globin genes, Curr. Sci. 66 (1994), 309-314.

[23] A. Nandy, Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences, Cabios 12 (1996), 55-62.

[24] P. Leong and S. Morgenthaaler, Random walk and gap plots of DNA sequences, Comput. Appl. Biosci. 11 (1995), 503-507.

[25] B. Hewelt et al., The DNA walk and its demonstration of deterministic chaos-relevance to genomic alterations in lung cancer, Bioinformatics 35 (2019), 2738-2748.

[26] A. Nandy et al., Characterizing the Zika virus genome - a bioinformatics study, Curr. Comp. Aided Drug Design 12 (2016), 87-97.

[27] S. S. T. Yau et al., DNA sequence representation without degeneracy, Nucleic Acids Res. 31 (2003), 3078-3080.

[28] C. Yu, M. Deng and S. S. T. Yau, DNA sequence comparison by a novel probabilistic method, Inform. Sci. 181 (2011), 1484-1492.

[29] T. Cover and J. Thomas, Elements of Information Theory, J. Wiley and Sons, 1991.

[30] J. Berger et al., New approaches to genome sequence analysis based on digital signal processing, Proc. Workshop on Genomic Signal Processing and Statistics (GENSIPS), IEEE, Raleigh, North Carolina, USA, 11-13 Oct. 2002, CP2-08. 2002, pp. 1-4.

[31] P. Cristea, Conversion of nucleotide sequences into genomic signals, J. Cell. Mol. Med. 6 (2002), 279-303.

[32] L. Das, S. Nanda and J. Das, An integrated approach for identification of exon locations using recursive Gauss Newton tuned Kaiser window, Genomics 111 (2019), 284-296.

[33] A. Brodzik and O. Peters, Symbol-balanced quaternion periodicity transform for latent pattern detection in DNA sequences, Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'05), 2005, Philadelphia, PA, USA, 2005, Vol. 5, pp. v/373-v/376.

[34] Z. J. Zang, DV-curve: a novel intuitive tool for visualizing and analyzing DNA sequences, Bioinformatics 25 (2009), 1112-1117.

[35] A. Nandy, M. Harle and S. Basak, Mathematical descriptors of DNA sequences: development and applications, Arkivoc (2006), 211-238.

[36] H. Kwan and S. Arniker, Numerical representation of DNA sequences, Proc. 2009 IEEE Int. Conf. Electro/Inf. Technol., Windsor, ON, Canada, 2009, pp. 307-310.

[37] M. Randić, M. Novič and D. Plavšić, Milestones in graphical bioinformatics, Int. J. Quantum Chem. 113 (2013), 2413-2446.

[38] V. Aram, A. Iranmanesh and Z. Majid, Spider representations of DNA sequences, J. Comput. Theor. Nanoscience 11 (2014), 418-420.

[39] Y. Li, Q. Liu and X. Zheng, DUC-curve, a highly compact 2D graphical representation of DNA sequences and its application in sequence alignment, Phys. A 456 (2016), 256-270.

[40] Z. Mo et al., One novel representation of DNA sequence based on the global and local position information, Sci. Rep. 8 (2018), 7592(1-7).

[41] G. S. Xie et al., Graphical representations and similarity analysis of DNA sequences based on trigonometric functions, Acta Biotheor. 66 (2018), 113-133.

[42] B. Lee, Squiggle: a user-friendly two-dimensional DNA sequence visualization tool, Bioinformatics 35 (2018), 1425-1426.

[43] J. Moroz and P. Nelson, Torsional directed walks, entropic elasticity, and DNA twist stiffness, Proc. Natl. Acad. Sci. USA 94 (1997), 14418-14422.

[44] M. Randić, 2-D graphical representation of proteins based on physico-chemical properties of amino acids, Chem. Phys. Lett. 476 (2009), 281-286.

[45] M. Mahmoodi-Reihani, F. Abbasitabar and V. Zare-Shahabadi, A novel graphical representation and similarity analysis of protein sequences based on physicochemical properties, Phys. A 510 (2018), 477-485.

[46] N. Marascio et al., Molecular characterization and cluster analysis of SARS-CoV-2 viral isolates in Kahramanmaras city, Turkey: The Delta VOC wave within one month, Viruses 15 (2023), 802(1-12).

[47] C. Peng et al., Long-range correlations in nucleotide sequences, Nature 356 (1992), 168-170.

[48] A. Grigoriev, Analyzing genomes with cumulative skew diagrams, Nucleic Acids Res. 26 (1998), 2286-2290.

[49] X. Q. Qi, J. Wen and Z. H. Qi, New 3D graphical representation of DNA sequence based on dual nucleotides, J. Theor. Biol. 249 (2007), 681-690.

[50] C. Li et al., Novel graphical representation and numerical characterization of DNA sequences, Appl. Sci. 6 (2016), 63(1-15).

[51] F. Bai et al., Vector representation and its application of DNA sequences based on nucleotide triplet codons, J. Mol. Graph Model. 62 (2015), 150-156.

[52] D. Bielińska-Wąż et al., 2D-dynamic representation of DNA sequences, Chem. Phys. Lett. 442 (2007), 140-144.

[53] A. Nandy et al., Characteristics of influenza HA-NA interdependence determined through a graphical technique, Curr. Comuter. Aided Drug Design 10 (2014), 285-302.

[54] A. Nandy and S. Basak, Prognosis of possible reassortments in recent H5N2 epidemic influenza in USA: implication for computer-assisted surveillance as well as drug/vaccine design, Curr. Comput. Aided Drug Design 11 (2015), 110-116.

[55] D. Panas et al., 2D-dynamic representation of DNA/RNA sequences as a characterization tool of the Zika virus genome, MATCH Commun. Math. Comput. Chem. 77 (2017), 321-332.

[56] D. Panas et al., An application of the 2D-dynamic representation of DNA/RNA sequences to the prediction of influenza a virus subtypes, MATCH Commun. Math. Comput. Chem. 80 (2018), 295-310.

[57] P. Wąż and D. Bielińska-Wąż, 3D-dynamic representation of DNA sequences, J. Mol. Model. 20 (2014), 2141(1-7).

[58] D. Bielińska-Wąż, P. Wąż and D. Panas, Applications of 2D and 3D-dynamic representations of DNA/RNA sequences for a description of genome sequences of viruses, Comb. Chem. High Throughput Screening 25 (2022), 429-438.

[59] P. Wąż and D. Bielińska-Wąż, Non-standard bioinformatics characterization of SARS-CoV-2, Comp. Biol. Med. 131 (2021), 104247(1-14).

[60] D. Bielińska-Wąż et al., 4D-dynamic representation of DNA/RNA sequences: studies on genetic diversity of Echinococcus multilocularis in red foxes in Poland, Life 12 (2022), 877(1-23).

[61] A. Czernieka et al., 20D-dynamic representations of protein sequences, Genomics 107 (2016), 16-23.

[62] A. Kostadinov and G. Kouzaev, A novel processor for artificial intelligence acceleration, WSEAS Trans. Circ. Systems 21 (2022), 125-141.

[63] B. Brejová, T. Vinar and M. Li, Pattern discovery, Introduction to Bioinformatics, S. Krawetz and D. Womble, eds., Humana Press, 2003, pp. 491-522.

[64] R. Mian, M. Shintani and M. Inoue, Hardware-software co-design for decimal multiplication, Computers 10 (2021), 17(1-19).

[65] N. Brisebarre et al., Comparison between binary and decimal floating-point numbers, IEEE Trans. Comput. 65 (2016), 2032-2044.

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

[67] Chapter 2. General Structure, The Unicode Standard (6.0 ed.), The Unicode Consortium: Mountain View, California, US.

[68] R. Hamming, Error detecting and error-correcting codes, Bell. Syst. Techn. J. 29 (1950), 147-160.

[69] W. Waggener, Pulse Code Modulation Techniques, Springer-Verlag, 1995.

[70] G. Navarro and M. Raffinot, Flexible Pattern Matching in Strings: Practical Online Search Algorithms for Texts and Biological Sequences, Cambridge University Press, 2002.

[71] V. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Phys. Doklady 10 (1966), 707-710.

[72] E. Gabidullin, Theory of codes with maximum rank distance, Problemy Peredachi Informatsii (Probl. Inform. Trans.) 21 (1985), 3-16.

[73] 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.

[74] X. Yang et al., Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries, Emerging Microbes and Infect. 9 (2020), 1287-1299.

[75] J. Tzeng, H. H. S. Lu and W. H. Li, Multi-dimensional scaling for large genomic data sets, BMC Bioinformatics 9 (2008), 179(1-17).

[76] A. Taghavi et al., Evaluating geometric definitions of staking for RNA dinucleoside monophosphates using molecular mechanics calculations, J. Chem. Theory Comput. 18 (2022), 3637-3653.

[77] A. Melkich and A. Khrennikov, Nontrivial quantum and quantum-like effects in biosystems: Unsolved questions and paradoxes, Progress Biophys. Mol. Biol. 119 (2015), 137-161.

[78] J. Feder, Fractals, Plenum Press, 1988.

[79] C. Berthelsen, J. Glazier and M. Skolnick, Global fractal dimension of human DNA sequences treated as pseudorandom walks, Phys. Rev. A 45 (1992), Paper No 89028913.

[80] P. Licinio and R. Caligiorne, Inference of phylogenetic distances from DNA-walk divergences, Phys. A 341 (2004), 471-481.

[81] A. Rosas, E. Nogueira Jr. and J. Fontanari, Multifractal analysis of DNA walks and trails, Phys. Rev. E 66 (2002), 061906(1-6).

[82] A. Haimovich et al., Wavelet analysis of DNA walks, J. Comput. Biol. 13 (2006), 1289-1298.

[83] H. Namazi et al., Diagnosis of skin cancer by correlation and complexity analyses of damaged DNA, Oncotarget 6 (2015), 42623-42631.

[84] G. Abramson, H. Cerdeira and C. Bruschi, Fractal properties of DNA walks, Biosystems 49 (1999), 63-70.

[85] C. Cattani, Fractals and hidden symmetries in DNA, Math. Probl. Eng. 2010 (2010), 507056(1-31).

[86] S. Ouadfeul, Multifractal analysis of SARS-CoV-2 coronavirus genomes using the wavelet transforms, BioRxiv preprint: https://doi.org/10.1101/2020.08.15.252411.

[87] B. Hao, H. C. Lee and S. Zhang, Fractals related to long DNA sequences and complete genomes, Chaos Solitons Fractals 11 (2000), 825-836.

[88] Z. Y. Su, T. Wu and S. Y. Wang, Local scaling and multifractality spectrum analysis of DNA sequences - GenBank data analysis, Chaos Solitons Fractals 40 (2009), 1750-1765.

[89] G. Durán-Meza, J. López-García and J. del Río-Correa, The self-similarity properties and multifractal analysis of DNA sequences, Appl. Math. Nonlin. Sci. 4 (2019), 267-278.

[90] M. Swapna and S. Sankararaman, Fractal applications in bio-nanosystems, Bioequiv. Availab. 2 (2019), pp. OABB.000541(1-4).

[91] X. Bin, E. Sargent and S. Kelley, Nanostructuring of sensors determines the efficiency of biomolecular capture, Anal. Chem. 82 (2010), 5928-5931.

[92] J. Chen et al., Research progress of DNA walker and its recent applications in biosensor, TrAC Trends in Anal. Chem. 120 (2019), 115626(1-14).

[93] A. Sadana, Engineering Biosensors, Kinetics and Design Application, Acad. Press, 2001.

[94] P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, Phys. D 9 (1983), 189-208.

[95] S. Rasband, Chaotic Dynamics of Nonlinear Systems, Dover Publications, 2015.

[96] B. Henry, N. Lovell and F. Camacho, Nonlinear dynamics time series analyses, Nonlinear Biomedical Signal Processing: Dynamic Analysis and Modeling, M. Akay, ed., IEEE Press, 2000, pp. 1-39.

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

[98] J. Véhel and P. Legrand, Signal and image processing with Fraclab, Thinking in Patterns, World Sci. (2003), 321-322.

[99] G. Kouzaev, Application of Advanced Electromagnetics, Components and Systems, Springer-Verlag, 2013.

[100] C. Guidolin et al., Does a self-similarity logic shape the organization of the nervous system? The Fractal Geometry of the Brain, A. Di Leva, ed., Springer-Verlag, 2016, pp. 138-156.

[101] FracLab 2.2. A Fractal Analysis Toolbox for Signal and Image Processing. [www.project.inria.fr/fraclab]

[102] X. H. Xie et al., A novel genome signature based on inter-nucleotide distances profiles for visualization of metagenomic data, Phys. A 482 (2017), 87-94.

[103] X. Yang et al., Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries, Emerging Microbes and Infect. 9 (2020), 1287-1299.

[104] C. Cao et al., The architecture of the SARS-CoV-2 RNA genome inside virion, Nature Commun. 12 (2021), 3917(1-14).

[105] A. Brant et al. SARS-CoV-2: from its discovery to genome structure, transcription, and replication, Cell and Bioscience 11 (2021), 136(1-17).

[106] C. Wu et al., Structure genomics of SARS-CoV-2 and its Omicron variant: drug design templates for COVID-19, Acta Pharm. Sinica 43 (2022), 3021-3033.

[107] V. Cooper, The coronavirus variants do not seem to be highly variable so far, Sci. American, 2021.

[108] S. El-Kafrawy et al., Enzootic patterns of Middle East respiratory syndrome coronavirus in imported African and local Arabian dromedary camels: a prospective genomic study, The Lancet Planetary Health 3 (2019), e521-e528.

[109] M. Kim et al., An infectious cDNA clone of a growth attenuated Korean isolate of MERS coronavirus KNIH002 in clade B, Emerg. Microbes Infect. 9 (2020), 2714-2720.

[110] V. Dwivedi et al., Genomics, proteomics and evolution of dengue virus, Briefings in Functional Genomics 16 (2017), 217-227.

[111] H. Abea et al., 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 91 (2020), 129-136.

[112] N. Di Paola et al., Viral genomics in Ebola virus research, Nature Rev. Microbiol. 8 (2020), 365-378.

[113] J. Zhang, Visualization for Information Retrieval, Springer-Verlag, 2007.

[114] M. Vračko et al., Cluster analysis of coronavirus sequences using computational sequence descriptors: with applications for SARS, MERS and SARS-CoV-2 (CoVID-19), Curr. Comput. Aided Drug Design 17 (2021), 936-945.

[115] V. Grishkevich and I. Yanai, Gene length and expression level shape genomic novelties, Genome Research 24 (2014), 1497-1503.

[116] T. Stoeger et al., Aging is associated with a systemic length-associated transcriptome imbalance, Nature Aging 2 (2022), 1191-1206.