Visualization and Characterization of Nucleotide Sequences in Bioinformatics: State-of-Art in the Past and Present

Dmitry A. Zimnyakov[1,2,3], Marina V. Alonova[1*], Anatoly V. Skripal[3], Onega V. Ulianova[3], and Valentina A. Feodorova[3]

[1] Yury Gagarin State Technical University of Saratov, 77 Polytechnicheskaya str., Saratov 410054, Russia
[2] Institute for Precision Mechanics and Control Problems of the Russian Academy of Sciences (IPTMU RAS), 24 Rabochaya str., Saratov 410028, Russia
[3] Saratov State University, 83 Astrakhanskaya str., Saratov 410012, Russia

***e-mail:** alonova_marina@mail.ru

Abstract. The review considers the main trends in the development of the methods for representation and characterization of DNA-associated nucleotide sequences, starting from the key works performed in the nineties of the last century and ending with the current state of the art. The basic principles of algorithms for processing nucleotide sequences and generating representing 2D and 3D objects are discussed in relation to such popular methods as polyline representation (generation of H-, Z-, C-, … curves), chaos game representation, and sequence logo representation. In addition, novel approaches to representation of genetic information using the principles of coherent optical information processing, such as GB speckle technique and polarization mapping, are considered. The discussion is illustrated by the examples of graphic images obtained using the basic algorithms for model nucleotide sequences or sequences borrowed from open sources.

Keywords: nucleotide sequences; graphical imaging; polyline representation; chaos game representation; sequence logo; coherent optical imaging.

Paper #8980 received 29 May 2023; revised manuscript received 29 Aug 2023; accepted for publication 29 Aug 2023; published online 21 Nov 2023. doi: 10.18287/JBPE23.09.040201.

1 Introduction

The growing challenges facing humanity due to pandemic nature of emergence and spread of various pathogenic microorganisms with high variability in the past few decades have led, in particular, to a rapid development of such interdisciplinary field as bioinformatics. This area of modern science combines the latest fundamental and applied achievements of such sciences as biochemistry, biophysics, information theory, mathematical statistics, the theory of random processes, etc. Currently, the procedure for obtaining and processing genetic information about various biological objects includes two essential stages, the first of which is sequencing of the raw biological material using the modern next-generation sequencing (NGS) technology: short-read and long-read sequencing techniques (see, e.g., [1]). In these technologies, the identification of certain nucleotides or their groups is usually carried out using fluorescent techniques (for instance, applying specific fluorescent labels). In addition, the identification can be based on measurements of ion currents (for example, due to changes in the concentration of hydrogen ions during DNA polymerization when interacting with certain types of deoxyribonucleotide triphosphates); this technique is defined as the ion semiconductor sequencing [1]. As a result, quasi-random sequences of four characters (A, C, T, G) associated with the nitrogenous bases (A – adenine, C – cytosine, T – thymine, G – guanine) that make up DNA are formed. These are heterocyclic organic compounds, which are derivatives of pyrimidine and purine. Adenine ($C_5H_5N_5$) and guanine ($C_5H_5N_5O$) are purine derivatives, while cytosine ($C_4H_5N_3O$) and thymine ($C_5H_6N_2O_2$) are

pyrimidine derivatives. Between the nitrogenous bases of deoxyribonucleotides of two different chains, three or two hydrogen bonds are formed (guanine binds to cytosine with three bonds, and adenine with thymine with two) [2].

The length of sequencing-recovered fragments, beginning from the so-called start codons, can vary from several hundred to tens of thousands letters. Protein synthesis occurs in ribosomes, and information from a DNA molecule is transferred by an RNA molecule, which is an intermediary. The RNA polymerase enzyme moves along the DNA chain and, falling on the ATG triplet (the start codon), initiates the beginning of the synthesis of the RNA molecule. The synthesis continues until the RNA polymerase occurs at the stop codon [2].

Identification, analysis, and visualization of these sequence fragments are the main tasks of modern bioinformatics, related to the second essential stage of genetic information processing. Over the past thirty years, abundance of various approaches has been developed to solve the problems of analyzing and visualizing symbol sets associated with sequenced nucleotide sequences. The key principles of these approaches differ significantly and are based on a variety of mathematical methods and algorithms, ranging from information entropy estimates according to Kolmogorov and Shannon to application of the concepts of the chaos game theory. In the last few years, some works related to the possible use of coherent light fields as carriers and converters of specially encoded genetic information have been published.

The purpose of this review is to consider the current state of the art of the methods for analyzing and visualizing genetic information presented in the form of symbolic sequences, to systematize these methods, and to discuss their advantages and disadvantages.

2 Inherent Statistical Properties of Nucleotide-Associated Symbol Sequences

Despite a seemingly random nature of the sequencing-recovered symbol series, they are characterized by a certain statistical heterogeneity. At the lower level, corresponding to individual symbols, this heterogeneity manifests itself in different relative frequencies of manifestations of each of the 4 symbols during sequential reading of the letter series. As an example, Fig. 1 shows relative frequencies of manifestation of adenine, cytosine, thymine, and guanine in nucleotide sequences corresponding to three pathogenic objects: the model Spike gene of the SARS-CoV-2 virus [3] Wuhan strain (the GenBank Acc. No. EPI_ISL_402124) [4], Influenza A virus (A/Weiss/43 (H1N1)) neuraminidase (NA) gene, complete cds GenBank: AF250365.2 [5], and Variola virus, complete genome NCBI Reference Sequence: NC_001611.1 [6].

At a higher level, corresponding to combination of individual nucleotide-associated symbols into groups of a given length, more flexible and versatile possibilities open up for identifying and analyzing nucleotide sequences. Traditionally, the sequences are divided into groups of three characters (triplets), beginning from the start codon (e.g., ATA, CTG, AAT, CAC, etc.). Thus, the number of different triplets is determined by a combinatorial evaluation of a number of possible combinations of four objects in groups of three items, which is 64.

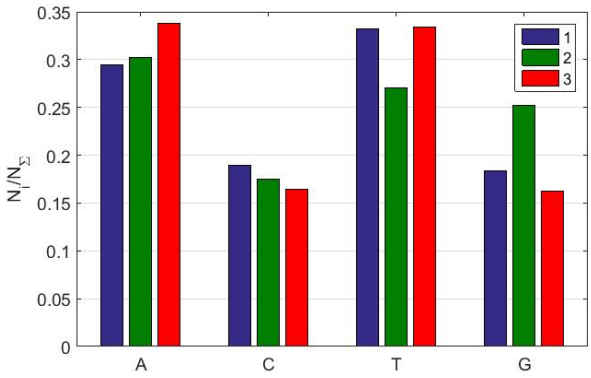


Fig. 1 The relative frequencies of appearance of four base nucleotides in DNA sequences of three pathogenic objects: 1 – Wuhan strain of the SARS-CoV-2 virus; 2 – Influenza A/H1N1 virus; 3 – Variola virus.

The hypothesis that DNA-associated nucleotide series have fractal properties was put forward in the nineties of the past century by a group of researchers [7] from the Institute of Molecular Genetics and Institute of Chemical Physics of the Russian Academy of Sciences. These properties are associated with existence of long-range power-law correlations along the DNA sequences established by a number of researchers almost at the same time [8–13]. In particular, a characteristic manifestation of such long-range correlations is the deviation of the exponent α of the so-called Peng function [8] $P(l) \infty l^{\alpha}$ from the trivial value of 0.5 towards larger values (of the order of 0.6–0.7). The Peng function is introduced as the r.m.s. value of the purine (A,G) – pyrimidine (C,T) difference in sequence fragments with the length of l depending on l.

Consideration of the fractal properties of nucleotide sequences through the estimates of the so-called Hurst exponent or a set of critical exponents can be classified as one of the approaches to a generalized description of inherent features of these sequences. As another generalized approach, we can note the definitions of complexity measures carried out using various techniques. Generally speaking, the idea of evaluating complexity of the texts goes back to the fundamentals of the information theory and was particularly discussed by Kolmogorov [14]. In 1976, J. Ziv and A. Lempel proposed an efficient technique for estimating the complexity measure of symbolic sequences of the finite length [15], which in various modifications has been applied to bioinformatics problems (in particular, for the compression of genetic information) [16–18]. In the Ziv-Lemplel method, the complexity measure of a given symbol sequence is evaluated as a minimal number of

steps necessary for the sequence synthesis in the framework of a certain procedure. This procedure allows for two operations: generation of a new symbol or copying a fragment from an already synthesized sequence.

In addition to the Ziv-Lempel complexity measure, the measure of linguistic complexity is also applied in bioinformatics [19, 20]. This parameter of one-dimensional symbol sequences is introduced in terms of vocabulary usage. This is the ratio of actual vocabulary of words (symbol combinations) of a given length L_W to the maximal possible vocabulary for the given sequence. The linguistic complexity a product of all vocabulary usages for $1 \le L_W \le N_t - 1$ (N_t is the number of symbols in the sequence).

In addition to the integral estimates of complexity and entropy [21, 22] of quasi-random symbol strings obtained by sequencing, other statistical approaches have been successfully applied in the generalized analysis of the properties of nucleotide series. These approaches include, for example, the method of the Bayes prediction [23], features of the Markov [24, 25] or Jukes-Cantor [26] processes, the Monte-Carlo method [27], specially designed probabilistic models [28], and multidimensional scaling and clustering techniques [29, 30].

3 Polyline Representation of Nucleotide Sequences in 2D and 3D Spaces

The eighties and nineties of the past century as well as the 2000s were marked by emergence of a large number of approaches to a polyline-based 2D- or 3D-visualization of DNA-associated symbol sequences. A variety of approaches that display distributions of the basic nucleotides in the analyzed fragments of the sequenced DNA in various ways in the plane and in space, gave rise to the concepts of imaging curves of the H-type, Z-type, etc. In general, the noticeable differences between these imaging techniques come down to two main points:

– the choice of the initial basis for 2D or 3D display, based on existence of only four basic nucleotides (A, C, T, G) and with account to the differences between them in terms of membership in the pyrimidine or purine groups, amino or keto groups, as well as in strength of the hydrogen bonds (weak or strong);

– the algorithm of an iterative procedure used to generate the subsequent start-to-end coordinates of the line segments in 2D or 3D space at a given iteration step.

Certainly, one of the pioneering works in this area was carried out by E. Hamori and J. Riskin in 1983 [31]. In fact, they established the general principles for specifying the bases for dot or polyline imaging of nucleotide sequences, which were then applied and modified in a number of further works. The Hamori-Riskin displaying technique was defined as a synthesis of the H-curve, which gradually falls down along the z-direction in the 3D Cartesian space, depending on the current position of a displayed nucleotide in the sequence. This current position is associated with a descending z-coordinate, and the (A, C, T, G) basis is localized on the plane $(x, y, z = 0)$. A visualized nucleotide at the position z is associated with a specific vector $\bar{g}(z)$ introduced following the rule: $A \to \bar{g}(z) = \bar{i} + \bar{j} - \bar{k}$; $C \to \bar{g}(z) = -\bar{i} - \bar{j} - \bar{k}$; $T \to \bar{g}(z) = \bar{i} - \bar{j} - \bar{k}$; $G \to \bar{g}(z) = -\bar{i} + \bar{j} - \bar{k}$. Accordingly, the resulting vector $\bar{h}(n)$, which defines the end of a line segment belonging to the H-curve at the *n*-position, is determined by summation of the $\bar{g}(z)$ vectors: $\bar{h}(n) = \sum_{z=1}^{n} \bar{g}(z)$. Following Hamori and Riskin, the H-curve "will be a space curve consisting of a series of n joined base vectors assembled in a head-to-tail fashion in the $-z$ direction" [31].

Based on the algorithm for generating the H-curve, it can be concluded that the *y*-projection of the vector $\bar{h}(n)$ characterizes excess in the number of purine nucleotides in the analyzed sequence over the pyrimidine nucleotides. In a similar way, the *x*-component indicates the dominance of amino nucleotides over the keto ones (note that these dominances can be both positive and negative). Among the advantages of the considered H-curve technique, the authors highlighted a possibility for identifying specific features of the visualized nucleotide sequences (in particular, preponderance of purine nucleotides, richness of A and T nucleotides, presence of only pyrimidine nucleotides, etc.) by visual inspection of the generated H-curve.

In the decade since the publication of Hamori and Riskin's work, several approaches to the polyline 2D-visualization of nucleotide sequences have been considered [32–40]. All of them are based on the general principle of a step-by-step generation of a displaying polyline along the rendered sequence in (A, C, T, G)-assigned coordinates. However, the rules for finding the start and end coordinates for a line segment at a given step can slightly differ from one approach to another. In particular, Nandy [33] used a segment generation rule in 2D space that is in general similar to the Hamori & Riskin's algorithm for the 3D imaging. Leong and Morgenthaker [34] considered a random walk presentation for DNA sequences on the (A, C, T, G) plane. In this approach, a step-by-step generation of the imaging curve is controlled by the following simple rule: if a displayed nucleotide is A, then the corresponding segment of the unit length is directed to the right; in the case of C it is directed to the left. Similarly, occurrence of T causes the up direction, and G corresponds to the down shift. Accordingly, the line that mimics a random walk trajectory is obtained (as an example, see Fig. 2).

Further development of this methodology in relation to the polyline 3D-imaging of DNA-associated symbol sequences was carried out in Refs. [35–37].

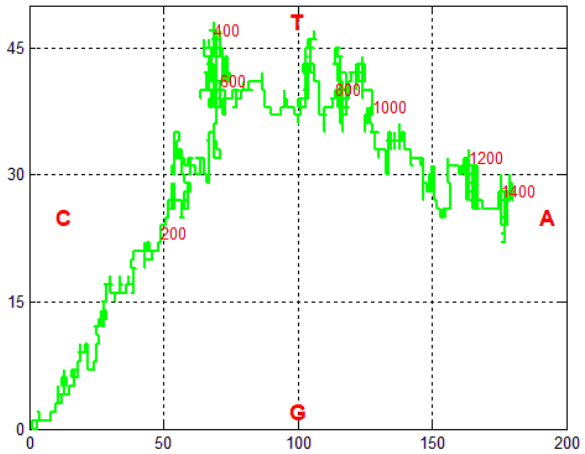


Fig. 2 Example of the random walk representation of the symbol sequence associated with the Influenza A/H1N1 virus.

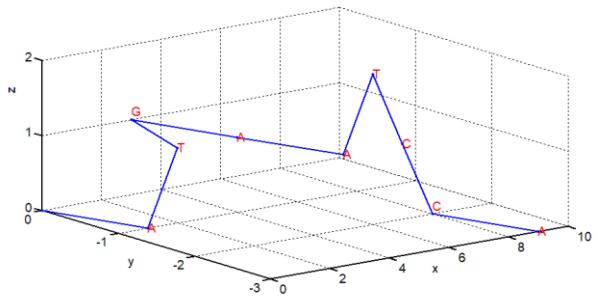


Fig. 3 Example of the recovered initial fragment of the RY-curve for the symbol sequence associated with the Influenza A/H1N1 virus.

However, despite certain advantages of these imaging techniques (in particular, obviousness), they all have a common drawback associated with information losses due to a possible overlapping and degeneracy of the displaying curve [38]. Generation of RY-, MK-, and WS-curves in 3D space, which is free of these disadvantages, was considered in Ref. [38]. In this case, the base points for a pair of nucleotides that are similar in one of three basic features (pyrimidine/purine, amine/keto, or weak hydrogen bond/strong hydrogen bond) are located on one of the coordinate planes (for example, *xy*), and for the other pair – on a perpendicular plane (e.g., *xz*). Similar to other polyline techniques, the end coordinates of the displaying segments associated with (A, C, T, G) symbols are generated sequentially using a recursive procedure when moving from a previous position in the symbol series to the given element. A recursive procedure considered in Ref. [38] is based on summing the end coordinates of symbol-associated basic points, following from the order of symbols in the displayed sequence. As a result, a polyline interpreted as the RY-, MK-, or WS-curve occurs in the 3D space. The starting point for all three curves is the origin.

As an example, let us consider the procedure of 3D visualization of the sequence fragment *ATGAATCCA* (the initial 9-base fragment of the symbol sequence for the Influenza A/H1N1 virus, [5]) in the RY basis (separation of nucleotides according to the "pyrimidine/purine" feature). In this case, the set of base points in 3D Cartesian coordinates is defined as: A → (1, −1, 0), G → (1, 1, 0), T → (1, 0, 1), C → (1, 0, −1); thus, the base points for purines (A, G) are located in the *xy*-plane, whereas the points corresponding to the pyrimidines lie in the *xz*-plane. Sequential generation of the nodal points of the RY-curve for a fragment of the model sequence is as follows:

$$\begin{cases} origin \\ (0,0,0) \end{cases} \xrightarrow{A} (1,-1,0) \xrightarrow{T} (2,-1,1) \xrightarrow{G} (3,0,1) \xrightarrow{A} (4,-1,1) \xrightarrow{A} (5,-2,1) \xrightarrow{T} (6,-2,2) \xrightarrow{C} (7,-2,1) \xrightarrow{C} (8,-2,0) \xrightarrow{A} (9,-3,0) \Big\}, \quad (1)$$

and the corresponding fragment of the RY-curve is displayed in Fig. 3.

The MK-curve and the WS-curve can be recovered in a similar way. In the first case, the basis is created as: A → (1, −1, 0), C → (1, 1, 0), G → (1, 0, 1), T → (1, 0, −1). Accounting for the strength of the hydrogen bonds (the WS-curve) corresponds to the following basis: A → (1, −1, 0), T → (1, 1, 0), G → (1, 0, 1), C → (1, 0, −1). Xie and Mo [38] substantiated such inherent properties of the RY-, MK-, and WS-curves as the absence of loop formation and degeneration, as well as a one-to-one correspondence between the visualized nucleotide sequence and the curves. In addition, the *x*-coordinate of the end node for all the curves defines the length of a visualized nucleotide sequence. One of important features of all the three curves is that analysis of the coordinates of their projections on the *xy*- and *xz*-planes makes it possible to determine the relative frequencies of occurrence of base nucleotides in the visualized sequences (i.e., has clear biological implication).

Among the discussed variety of polyline representations of nucleotide sequences, it is also necessary to note the technique for generating C-curves, proposed in Refs. [39, 40]. Generation of polyline node coordinates in this case is performed using a much more sophisticated procedure compared to those discussed above. On the (x, y) plane, a set of 64 base points is formed with the coordinates x = (−4; −3; −2; −1; 1, 2, 3 4); y = (−4; −3; −2; −1; 1, 2, 3, 4). Each base point is associated with a certain triplet of nucleotides in accordance with the following scheme: 16 triplets located in the first quadrant of the (x, y) plane contain adenine (A) as the first nucleotide. Accordingly, guanine (G) is the first nucleotide for the triplets of the second quadrant, cytosine (C) for the triplets of the third quadrant, and thymine (T) for the triplets of the fourth quadrant. In each

triplet, the remaining nucleotides are distributed according to the following scheme:

$$\begin{matrix} IGC & IGG & IAG & IAC \\ IGT & IGA & IAA & IAT \\ ICT & ICA & ITA & ITT \\ ICC & ICG & ITG & ITC \end{matrix}, \qquad (2)$$

where I denotes the first nucleotide. For example, the triplet TAT corresponds to the point (4, −2) in the (x, y) plane, and the triplet CGA is assigned to the point (−3, −2). Using this base, sequential generation of the coordinates of the C-polyline nodes is carried out as:

$$\begin{cases} x_n = \sum_{i=1}^{n} x_i; \\ y_n = \sum_{i=1}^{n} y_i; \\ z_n = \sum_{i=1}^{n} x_i y_i, \end{cases} \qquad (3)$$

where n is the current position of a displayed nucleotide in the sequence. According to Jafarzadeh and Iranmanesh [40], there are no manifestations of looping and degeneration in the C-polylines synthesized in this way. However, the declared advantages of this method in comparison with other 3D polyline representation techniques are not obvious.

4 The Concept of a Chaos Game Representation (CGR) and its Application in Bioinformatics

Application of the chaos game representation (CGR) in bioinformatics started with the pioneering work of Jeffrey [41], who extended the principles of generating two-dimensional fractal objects to the DNA analysis.

The basic idea of the chaos game representation is to sequentially fill a square area with the dots associated with sequentially located symbols in the mapped DNA-related symbol series. The vertices of the square are related to the (A, C, T, G) symbols in a certain way, and the choice of this relationship is determined by setting the representation configuration. As a rule, three basic configurations are applied, similarly to those used for polyline visualization in 2D and 3D spaces. Thus, if we take the upper left vertex of the square as the starting point and arrange the characters clockwise: CGTA (the RY configuration), GCTA (the MK configuration), and CTGA (the WS configuration). These designations are associated with the symbols on the main and minor diagonals of the square. For example, the RY configuration corresponds to the purines ($R = \{A, G\}$) on the minor diagonal, while the pyrimidines ($Y = \{C, T\}$) are on the main diagonal of the square. The MK configuration takes into account the correspondence of the basic nucleotides to the amino ($M = \{A, C\}$, the minor diagonal) or keto ($K = \{G, T\}$, the main diagonal) groups. Finally, the WS configuration corresponds to the nucleotides of weak hydrogen bonds $W = \{A, T\}$ at the minor diagonal, and of strong bonds $S = \{C, G\}$ at the main diagonal.

Mapping of the symbol sequence is carried out in accordance with the following recursive procedure (Fig. 4): the first symbol in the sequence is marked by the dot at the midpoint between the square center and the vertex denoted by the same symbol. The dot position corresponding to the second symbol is obtained by placing the dot at the midpoint between the position of the first dot and the vertex square corresponding to the dotted symbol. This procedure is repeated until the end of the mapped sequence. The choice of the base configuration (RY, MK, or WS) drastically affects the recovered CGR maps. A remarkable feature is the expressed large-scale inhomogeneities of the dot density across the map areas; this feature is directly related to the above mentioned existence of long-range correlations along the nucleotide sequences. Note that pure random distributions of A, C, T, G letters in the sequences are displayed by uniform distributions of the dots across the synthesized CGR maps.

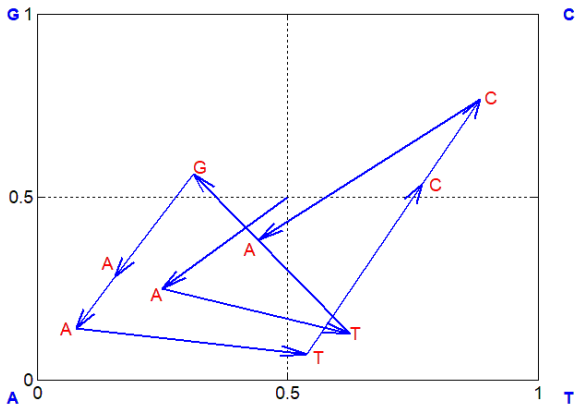
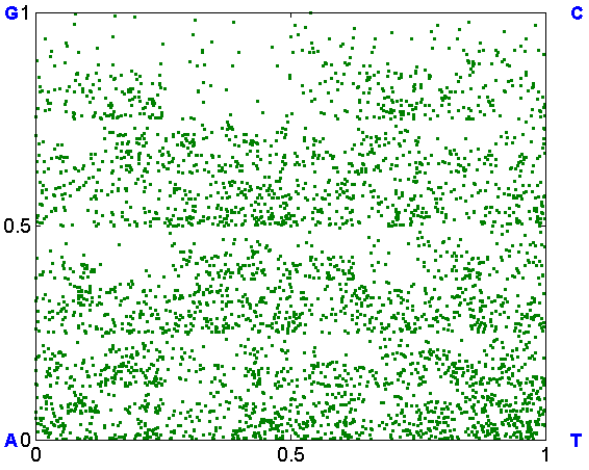
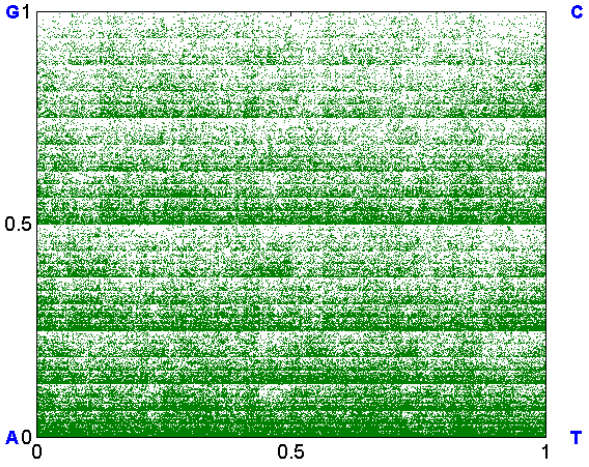


Fig. 4 Example of the initial stage of the CGR map generation in the MK basis for the symbol sequence associated with the Influenza A/H1N1 virus.

A common problem in the synthesis of the CGR maps stems from the limited lengths of the symbol sequences available from DNA sequencing. In other words, the low filling density of the synthesized maps with nucleotide-representing dots reduces reliability of detecting global and local features in distributions of nucleotides along sequenced sequences. As an example, Fig. 5 shows the synthesized maps in the MK base for two symbol sequences corresponding to the available DNA data for the Wuhan strain of the SARS-CoV-2 virus [3] (a) and the Variola virus [6] (b). In the first case, the length of the available sequence was 3822 symbols, while in the second case, 185578 symbols were processed.



(a)



(b)

Fig. 5 The synthesized CGR maps for symbol sequences associated with (a) the Wuhan strain of the SARS-CoV-2 virus (3822 bases are available) and (b) the Variola virus (185578 bases are available).

Obviously, the level of identification of significant details on the synthesized chaos game maps dramatically falls down with a decreasing length of the available sequence.

Further development of the CGR technique led to appearance of a frequency chaos game representation (FCGR), in which the problem of small lengths of the analyzed sequences is partially solved. In the FCGR technique, which can be interpreted as a "coarse-grained CGR", the display square is divided into $N \times N$ sub-squares and the number of imaging dots falling within each square is counted. Thus, the set of 8×8 sub-squares corresponds to the selection of 3-mers (i.e., triplets). Visualization of the obtained FCGR data can be done, for example, using the grayscale coding. As an example, Fig. 6 displays the recovered FCGR grayscale map in the MK base for the Wuhan strain of the SARS-CoV-2 virus (see the left panel in Fig. 5 for comparison).

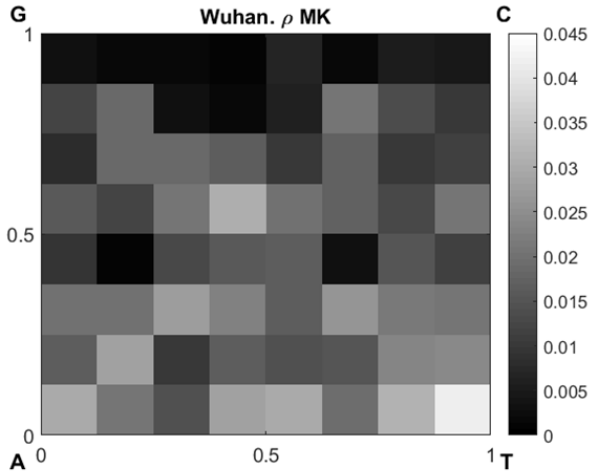


Fig. 6 The (8×8) FCGR grayscale map in the MK base for the Wuhan strain of the SARS-CoV-2 virus.

The basic principles of the FCGR technique were developed by Burma et al. [42]; further progress in this direction was achieved in the works of Huynen et al. ([43], analysis of the GC content in histones), Hill et al. ([44], frequency analysis of dinucleotides for human globin and alcohol dehydrogenase genes). Almeida et al. [45] proposed a more sophisticated way of dividing the imaging square into sub-squares (quadrants) based on the logarithmic relartionship between the number q of the quadrants and the number k of nucleotides in the groups: $k = \left(\log_2(q)\right)/2$. This approach obviously leads to a non-integer resolution of the FCGR mapping. However, the authors justified the use of this logarithmic scaling for addressing redundancy in the genomic structures, which is a major feature, e.g., in the case of amino acid encoding.

Typically, the FCGR maps are plotted in 2D using the gray-scale or color coding; however, it is necessary to note attempts to extend this approach to the 3D space (see, e.g., [46–49]). In particular, in the synthesized 3D CGR maps, the numbers of k-mers are represented along the z-axis instead of a gray scale.

In recent years, various modifications of the CGR method have been actively used to analyze variability of the SARS-CoV-2 virus. In particular, several groups applied the CGR and FCGR techniques to recover phylogenetic trees of coronaviruses [50–52]. Coupling of CGR with an artificial neural network provided additional opportunities to represent the genome and to classify intra-coronavirus sequences [53]. An original approach to the analysis of DNA sequences based on a combination of CGR and modeling of small-angle light scattering is considered in Refs. [54, 55].

5 Sequence Logo Representation

Among a variety of techniques for 2D representation of nucleotide sequences actively used now in bioinformatics, it is necessary to mention the sequence

logo technique introduced by T. D. Schneider and R. M. Stephens in 1990 [56]. This technique aims to identify the most conserved bases across the collections of aligned nucleotide sequences and is based on the estimates of the information entropy following C. Shannon. A sequence logo can be viewed as a specific histogram for a set of nucleotide sequences of the same biological object; this histogram is created according to the following procedure:

– at the first step, the analyzed sequences are aligned with respect to each other;

– next, the frequency sets of occurrences of each nucleotide at each position in the associated sequences are generated;

– with the frequency sets generated, the amount of information (in bits per position) for each position i along the aligned sequences can be defined as:

$$R(i) = 2 + \sum_{m=A}^{G} f(m,i)\log_2\left[f(m,i)\right] - e(n), \quad (4)$$

– where $f(m,i)$ is the frequency of occurrence of the nucleotide m ($m = A,C,T,G$) at the given position, the second term on the right side is the measure of information uncertainty (entropy) according to Shannon, and the third term is a correction factor that takes into account the number of nucleotide sequences in the analyzed collection;

– contribution of a given symbol to the content of information at a given position is estimated as $f(m,i)\cdot R(i)$.

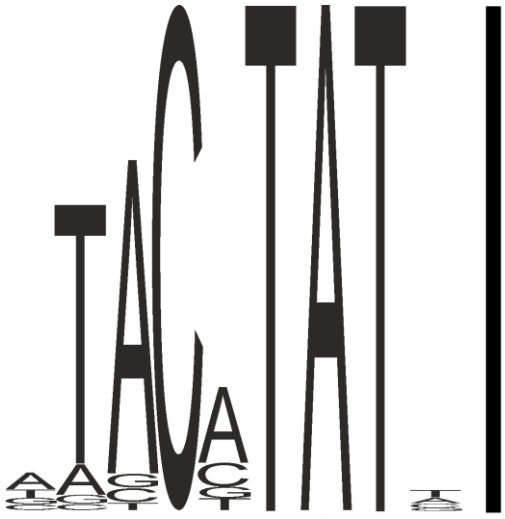


Fig. 7 The fragment of synthesized sequence logo for a model set of the aligned (A, C, T, G) series; the relative frequencies for (A, C, T, G) symbols at each position are presented below the logo; the maximal information content in the stacks (2 bits) is marked by the vertical black bar.

In the sequence logo technique, representation is carried out by setting the height $h(m,i)$ of the visualized symbols at the given position i in accordance with the expression $h(m,i) = f(m,i)\cdot R(i)$. Then the symbols are stacked on top of each other in an increasing order of their frequencies and plotted; this procedure is illustrated in Fig. 7.

6 Visualization and Analysis of Gene-Based Structures Using the Principles of Coherent Optics

The idea of coherent-optical methods for the analysis and visualization of genetic information goes back to the principles of optical information processing. These principles proceed from an unique correspondence between the amplitude-phase characteristics of the coherent light field diffracted by some object and structural properties of an object. Accordingly, by forming some 2D object that in a certain way imitates the structure of the analyzed 1D nucleotide sequence, we can analyze the diffracted light field as a fingerprint of the sequence structure. The GB speckle technology [57] uses this principle and is based on the idea of converting linear A, C, T, G sequences into two-dimensional quasi-random phase matrices read by a coherent light beam. As a result, a speckle pattern is formed in the far diffraction zone with spatial distributions of the amplitude, phase, and intensity that uniquely correspond to distribution of the symbols in the original sequence (abbreviation “GB” corresponds to the term “gene-based”). In particular, the following algorithm was proposed for converting the triplet-divided symbol sequences into two-dimensional phase matrices:

1) each triplet is assigned a certain number in the range from 0 to 63 in accordance with the following rule:

$$D_T = 16a_3 + 4a_2 + a_1 - 21, \quad (5)$$

where the lower indices in the right side define the symbol positions in the triplet and the following associations are accepted: $A \to a = 1; C \to a = 2, T \to a = 3, G \to a = 4$. Accordingly, the zero number corresponds to the combination AAA in the triplet, and the value of 63 is assigned to the GGG triplet;

2) from the obtained numeral sequence, a fragment of length N is selected that satisfies the condition $N = n_{\max}^2$, where $n_{\max}^2$ is the maximum square of an integer not exceeding the number of triplets in the analyzed sequence; then this fragment is divided into n sub-fragments of n-th length which are considered as consecutive rows of the matrix $\{a_{ij}\}$ being the basis for the GB phase-modulating matrix $\{\Delta\varphi_{ij}\} = K_\varphi\{a_{ij}\}$, where K_φ is the phase-modulating factor and $a_{ij} = D_T$ for a given triplet;

3) when reading the phase-modulating matrix $\{\Delta\varphi_{ij}\}$ by a collimated laser beam, the GB speckle structure is formed in the far diffraction zone (for example, in the focal plane of the Fourier transform lens).

The choice of the K_φ value is determined by the condition for the formation of a developed speckle pattern [58] in the far diffraction zone. This condition corresponds to the fact that the variance of phase fluctuations in the boundary light field (directly behind the synthesized phase screen) must significantly exceed π^2. Accordingly, for the case under discussion (the values of D_T are in the range from 0 to 63), the value of the phase modulation factor should significantly exceed $\pi/63$. The developed speckle pattern formed in this case is characterized by a zero mean value of the complex amplitude and negative exponential statistics of intensity fluctuations [58].

As discussed in Refs. [59, 60], GB speckle patterns created using laser radiation, or their digital representations, can be analyzed using coherent optical methods of information processing. However, it should be noted that numerical modeling of decorrelation of complex amplitude spatial distributions in the GB speckle fields at small nucleotide substitutions in the initial sequences shows that such decorrelation is not very significant. Accordingly, the use of GB speckle technology for identification and characterization of mutational changes in the DNA structure is currently rather debatable. In this regard, another approach to coherent-optical visualization and analysis of DNA-associated symbol sequences is of particular interest. This approach is based on polarization coding of the triplets in the analyzed sequences and formation of binary maps of extreme local polarization states in the diffraction zone, which are uniquely associated with the analyzed gene structures [59, 60].

The idea of polarization coding of the triplets is to represent each triplet in the analyzed sequence with the (2×2) submatrix $\{a_{ij}\}$, where each element is associated with one of the four nucleotides (A, C, T, G), and the value of the element corresponds to the content of this nucleotide in the triplet. For example, the triplet ATA is represented as:

$$ATA \rightarrow \begin{pmatrix} 2 & 0 \\ 1 & 0 \end{pmatrix}. \qquad (6)$$

Here the following associations are used: $a_{11} \leftrightarrow A$; $a_{12} \leftrightarrow C$; $a_{21} \leftrightarrow T$; $a_{22} \leftrightarrow G$. An inherent property of the submatrices is that the sum of all elements is always 3. After this transformation, a square phase-modulating matrix $\{\tilde{a}_{km}\}$ is assembled from the resulting set of submatrices by combining submatrices line by line. In contrast to the $(N \times N)$-sized phase-modulating matrices synthesized within the framework of the GB speckle technique, the $\{\tilde{a}_{km}\}$ matrices used in polarization mapping have the size $(2N \times 2N)$. Another difference lies in a significantly lower bit depth of polarization coding in comparison with the GB speckle technique.

In the modification of the polarization mapping technique considered in Ref. [60], information is read from the matrix $\{\tilde{a}_{km}\}$ by a linearly polarized collimated laser beam with the polarization plane oriented along the main diagonal of the matrix. The phase modulating factor for orthogonally polarized components of the readout beam is set to π; in addition, for one of the components of the readout beam, the phase shift of $\pi/2$ is introduced with respect to the other orthogonally polarized component. Thus, a quasi-random low-bit set of local polarization states is formed directly behind the phase-modulating matrix over the cross section of the transmitted readout beam.

Note that the mentioned algorithm of polarization modulation of the readout light beam (the phase-modulating factor of π and the phase shift of $\pi/2$ for one of the components) makes it possible to provide a set of close-to-circular local polarization states in the readout plane. As found, this set is very sensitive to local changes in the structure of the encoded nucleotide sequences. In Ref. [60], a slightly different polarization modulation algorithm with reading of local polarization states in the paraxial region of the readout plane is also considered, which allows performing frequency analysis of the nucleotide content in encoded sequences.

The polarization-dependent diffraction pattern corresponding to the analyzed symbol sequence is analyzed in the focal plane of the Fourier-transforming lens in terms of the Stokes vector components $\{s^r_{t,p}\}$. The $\{s^r_{p,t}\}$ values in the (p,t) point of the readout plane are determined by the amplitudes of *x*- and *y*- polarized components and the phase shift between them in the considered point. Thus, the local value of the first component of the Stokes vector ($r = 0$) defines the total intensity of the diffracted light in the (p,t) point; the $\{s^1_{p,t}\}$ value corresponds to the difference of intensities of the *x*- and *y*- components in the observation point. The third component $\{s^2_{p,t}\}$ also characterizes the difference in intensities, but in the 45°-rotated coordinate system. Finally, the fourth component $\{s^3_{p,t}\}$ characterizes a contribution of the circular polarization (right or left, depending on the sign) to the total polarization state at the detection point. For convenience of the analysis, it is advisable to consider the normalized values $\{s^{1,2,3}_{p,t}\} = \{S^{1,2,3}_{p,t} / S^0_{p,t}\}$. The possible values of $\{s^{1,2,3}_{p,t}\}$ are in the range from -1 to 1, and proximity of these values to ± 1 corresponds to the polarization close to the extreme state at the detection point. Accordingly, the binarized maps of distributions of extreme local polarization states

in the readout plane can be considered as unique two-dimensional identifiers of the analyzed symbol sequences. As an example, Fig. 8 (a, b) displays the binarized distributions of $s^3_{p,t}$ extreme states ($-0.999 > s^3_{p,t} \geq -1$), which correspond to the symbol sequences for the Wuhan and Omicron [61] strains of the SARS-CoV-2 virus. Fig. 8(c) displays the result of the pixel-by-pixel conjunction of Fig. 8(a) and Fig. 8(b) as a representation of differences between these binary maps.

To quantify the differences in the binary distributions of extreme polarization states, the correlation coefficient $R_{1,2}$ can be introduced as $R^{1,2,3}_{1,2} = \sum_{p,t} \tilde{s}^{1,2,3}_{(p,t),1} \times \tilde{s}^{1,2,3}_{(p,t),2} \Big/ \sum_{p,t} \tilde{s}^{1,2,3}_{(p,t),1}$. Here, the upper index defines the type of extreme polarization state used for analysis, and the lover indices (1, 2) correspond to the reference and analyzed binary patterns. Statistical modeling has shown that the correlation coefficient introduced in this way is highly sensitive to small measurements in the structure of the analyzed sequence relative to the reference sequence (with the number of nucleotide substitutions of the order of 1–3). As an example, Fig. 9 shows dependence of $R^3_{1,2}$ on the number of substitutions when the symbol sequence for the Wuhan strain of the SarsCov2 virus is applied as a reference. Also, the correlation coefficient for the pairs "Wuhan strain/Omicron strain" and "Wuhan strain/Delta strain" [62] are displayed by the square and triangular markers.

It should be noted that possibilities of the method of polarization mapping of DNA-associated symbol sequences are not limited to the considered algorithm of polarization encoding of initial sequences, as well as to visualization and analysis of distributions of extreme polarization states in the readout plane. In particular, it is possible to provide the frequency analysis of the sequences using the multistage polarization modulation of a readout laser beam by the synthesized phase-modulating matrix $\{\tilde{a}_{km}\}$ [60].

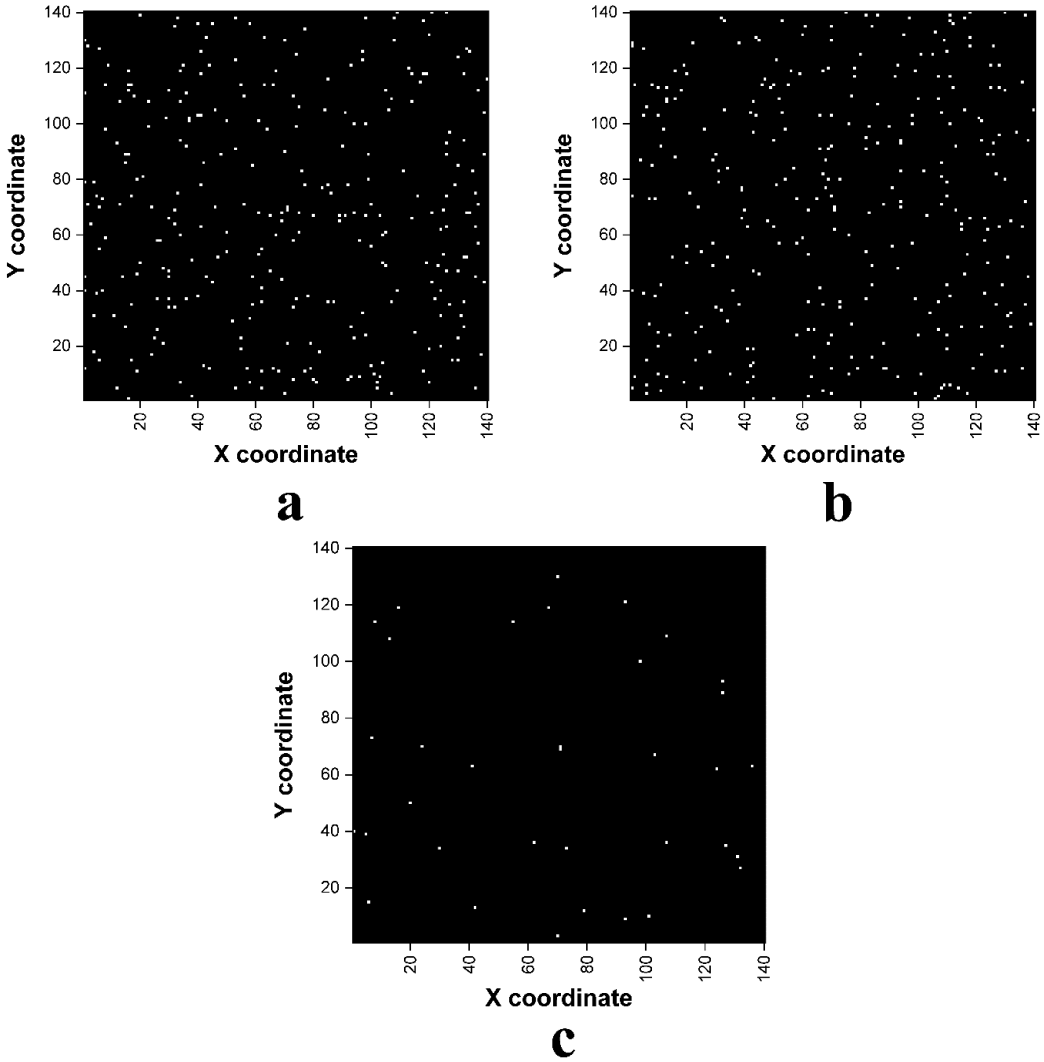


Fig. 8 Binary maps of the extreme polarization state (the predominating left circular polarization with $-0.999 > s^3_{p,t} \geq -1$) in the cases of (a) Wuhan and (b) Omicron strains of the SARS-CoV-2 virus. So high threshold level (−0.999) was chosen to highlight differences between two binary maps. Panel (c) is the result of the pixel-by-pixel conjunction of panels (a) and (b).

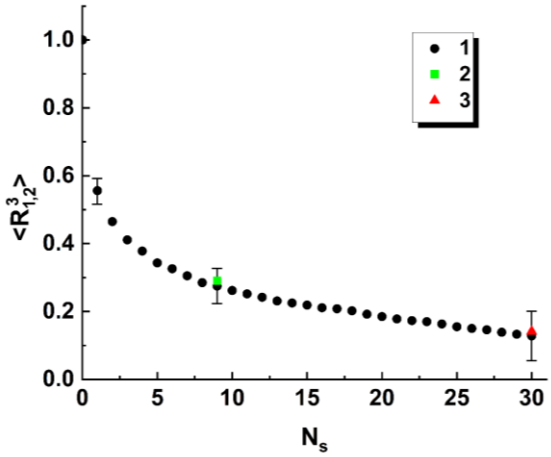


Fig. 9 The correlation between the reference and changed binary distributions $\tilde{s}^3_{(p,t)}$ depending on the number of symbol substitutions to the reference (A, C, T, G) sequence (Wuhan strain of the SARS-CoV-2 virus). The threshold level is equal to −0.99. (1) model data; (2) "Wuhan/Delta" correlation; (3) "Wuhan/Omicron" correlation. Selectively shown error bars correspond to the confidence level of 0.9.

7 Conclusions

Thus, abundance of the methods for graphical representation of nucleotide sequences developed in bioinformatics over the past three decades makes it possible to analyze various features of the genetic structure of biological objects, based on the results of the sequencing DNA fragments. At the same time, it should be noted that none of these methods has sufficient universality in terms of characterization of the genetic structure, which would make it possible to displace other approaches from the field of practical applications in bioinformatics. Judging by the number of references, one of the most popular methods is the chaos game representation (CGR) and its modifications, that make it possible to reveal the presence of large-scale correlations of the symbols and their groups in DNA-associated (A, C, T, G) sequences. In addition, a comparative analysis of the patterns synthesized in RY, MK, and WS bases provides a lot of additional information about the biochemical features of the sequences. However, such analysis requires further development of the methods for the quantitative analysis of point patterns with a fractal structure in terms of identifying the features of initial sequences that are important for bioinformatics. Another disadvantage of the CGR technique in its classical form is the need to use sufficiently long character sequences (from 10000 bases and over) in order to obtain sufficiently representative patterns. The FCGR modification is free of this disadvantage.

The sequence logo technique is good for identifying and visualizing the most conserved bases in the sequences during mutational changes; however, for its implementation, it requires a fairly large volume of collections of the aligned sequences. Numerous attempts at polyline representation of nucleotide sequences in 2D and 3D spaces (H-curve, C-curve, RY-curve, etc.), varying in the methods of choosing the base points in the Cartesian coordinate system and recursive algorithms for generating polyline node coordinates, have various particular advantages though they are by no means universal.

The coherent-optical methods of representing the DNA-associated symbol sequences considered here (the GB speckle technique and polarization mapping) can be classified as hybrid methods in which computer processing of the information is supplemented by integral transformations of optical information in the coherent-optical processor.

Acknowledgments

This research was funded by the Russian Science Foundation, grant number 22-21-00194.

Disclosures

The authors declare no conflict of interest.

References

1. S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," Nature Reviews Genetics 17, 333–351 (2016).
2. S. Neidle, M. Sanderson, Principles of nucleic acid structure, 2nd ed., Academic Press (2021). ISBN: 9780128196786.
3. J. F.-W. Chan, S. Yuan, K.-H. Kok, K. K.-W. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C.-Y. Yip, R. W.-S. Poon, H.-W. Tsoi, S. K.-F. Lo, K.-H. Chan, V. K.-M. Poon, W.-M. Chan, J. D. Ip, J.-P. Cai, V. C.-C. Cheng, H. Chen, C. K.-M. Christopher Kim-Ming Hui, and K.-Y. Yuen, "A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster," Lancet 395(10223), 514–523 (2020).
4. "Official hCoV-19 Reference Sequence, hCoV-19/Wuhan/WIV04/2019", GISAID (accessed on 10 April 2023). [https://www.epicov.org/epi3/frontend#541d4].
5. "Influenza A virus (A/Weiss/43 (H1N1)) neuraminidase (NA) gene, complete cds," National Library of Medicine (accessed on 10 April 2023). [https://www.ncbi.nlm.nih.gov/nuccore/AF250365.2].
6. "Variola virus, complete genome," National Library of Medicine (accessed on 10 April 2023). [https://www.ncbi.nlm.nih.gov/nuccore/NC_001611.1].

7. A. S. Borovik, A. Yu. Grosberg, and M. D. Frank-Kamenetskii, "Fractality of DNA texts," Journal of Biomolecular Structure and Dynamics 12(3), 655–669 (1994).
8. C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, "Long-range correlations in nucleotide sequences," Nature 356(6365), 168–170 (1992).
9. W. Li, K. Kaneko, "Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence," Europhysics Letters, 17(7), 655–660 (1992).
10. W. Li, "Generating nontrivial long-range correlations and 1/F spectra by replication and mutation," International Journal of Bifurcation and Chaos 2(1), 137–154 (1992).
11. R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," Physical Review Letters 68(25), 3805–3808 (1992).
12. B.-L. Hao, "Fractals from genomes–exact solutions of a biology-inspired problem," Physica A 282(1–2), 225–246 (2000).
13. S. Buldyrev, N. Dokholyan, A. Goldberger, S. Havlin, C.-K. Peng, H. Stanley, and G. Viswanathan, "Analysis of DNA sequences using methods of statistical physics," Physica A: Statistical Mechanics and its Applications 249(1–4), 430–438 (1998).
14. A. N. Kolmogorov, "Three approaches to the quantitative definition of information," Problems of Information Transmission 1(1), 3–11 (1965).
15. J. Ziv, A. Lempel, "On the complexity of finite sequences," IEEE Transactions on Information Theory 22(1), 75–81 (1976).
16. S. Grumbach, F. Taxi, "A new challenge for compression algorithms: genetic sequences," Information Processing & Management 30(6), 875–885 (1994).
17. L. Alison, T. Edgoose, and T. I. Dix, "Compression of strings with approximate repeats," Proceedings on Intelligent Systems for Molecular Biology, 8–16 (1998).
18. V. D. Gusev, L. A. Nemytikova, and N. A. Chuzhanova, "On the complexity measures of genetic sequences," Bioinformatics 15(12), 994–999 (1999).
19. O. G. Troyanskaya, O. Arbell, Y. Koren, G. M. Landau, and A. Bolshoy, "Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity," Bioinformatics 18, 679–688 (2002).
20. G. Gordon, "Multi-dimensional linguistic complexity," Journal of Biomolecular Structure & Dynamics 20(6), 747–750 (2003).
21. Yu. L. Orlov, R. Te Brokherst, and I. I. Abnizova, "Statistical measures of the structure of genomic sequences: entropy, complexity, and position information," Journal of Bioinformatics and Computational Biology 4(2), 523–536 (2006).
22. J. Riolo, A. J. Steckl, "Comparative analysis of genome code complexity and manufacturability with engineering benchmarks," Scientific Reports 12(1), 2808 (2022).
23. M. Anisimova, P. Joseph, J. P. Bielawski, and Y. Ziheng, "Accuracy and power of Bayes prediction of amino Acid sites under positive selection," Molecular Biology and Evolution 19(6), 950–958 (2002).
24. E. Rivas, S. R. Eddy, "Noncoding RNA gene detection using comparative sequence analysis," BMC Bioinformatics 2, 1–19 (2001).
25. I. Abnizova, K. Walter, R. Te Boekhorst, G. Elgar, and W. R. Gilks, "Statistical information characterization of conserved non-coding elements in vertebrates," Journal of Bioinformatics and Computational Biology 5(02b), 533–547 (2007).
26. S. R. Eddy, "A model of the statistical power of comparative genome sequence analysis," PLoS Biology 3(1), e10 (2005).
27. D. G. Hwang, P. Green, "Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution pat-terns in mammalian evolution," proceedings of the National Academy of Sciences 101(39), 13994–14001 (2004).
28. A. J. Pinho, S. P. Garcia, D. Pratas, and P. J. S. G. Ferreira, "DNA sequences at a glance," PLoS ONE 8(11), e79922 (2013).
29. J. T. Machado, A. M. Lopes, "Multidimensional scaling and visualization of patterns in prime numbers," Communications in Nonlinear Science and Numerical Simulation 83, 105128 (2020).
30. J. A. T. Machado, J. M. Rocha-Neves, F. Azevedo, and J. P. Andrade, "Advances in the computational analysis of SARS-COV2 genome," Nonlinear Dynamics 106(2), 1525–1555 (2021).
31. E. Hamori, J. Ruskin, "H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences," Journal of Biological Chemistry 258(2), 1318–1327 (1983).
32. M.A. Gates, "A simple way to look at DNA," Journal of Theoretical Biology 119(3), 319–328 (1986).
33. A. Nandy, "A new graphical representation and analysis of DNA sequence structure. I: methodology and application to globin genes," Current Science 66, 309–314 (1994).
34. P. M. Leong, S. Morgenthaler, "Random walk and gap plots of DNA sequences," Bioinformatics 11(5), 503–507 (1995).

35. M. Randic, M. Vracko, N. Lers, and O. Plavsic, "Novel 2-D graphical representation of DNA sequences and their numerical characterization," Chemical Physics Letters 368(1–2), 1–6 (2003).
36. M. Randic, M. Vracko, N. Lers, and O. Plavsic, "On 3-D graphical representation of DNA primary sequence and their numerical characterization," Journal of Chemical Information and Computer Sciences 40(5), 1235–1244 (2000).
37. Z. G. Yu, B. Wang, "A time series model of CDS sequences in complete genome," Chaos Solitons Fractals 12(3), 519–526 (2001).
38. G. Xie, Z. Mo, "Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications," Journal of Theoretical Biology 269(1), 123–130 (2011).
39. N. Jafarzadeh, A. Iranmanesh, "A novel graphical and numerical representation for analyzing DNA sequences based on codons," Match-Communications in Mathematical and in Computer Chemistry 68(2), 611–620 (2012).
40. N. Jafarzadeh, A. Iranmanesh, "C-curve: A novel 3D graphical representation of DNA sequence based on codons," Mathematical Biosciences 241(2), 217–224, (2013).
41. H. J. Jeffrey, "Chaos game representation of gene structure," Nucleic Acids Research 18(8), 2163–2170 (1990).
42. P. K. Burma, A. Raj, J. K. Deb, and S. K. Brahmachari, "Genome analysis: a new approach for visualization of sequence organization in genomes," Journal of Biosciences 17(4), 395–411 (1992).
43. M. A. Huynen, D. A. M. Konings, and P. Hogeweg, "Equal g and c contents in histone genes indicate selection pressures on mrna secondary structure," Journal of Molecular Evolution 34(4), 280–291 (1992).
44. K. A. Hill, N. J. Schisler, and S. M. Singh, "Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species," Journal of Molecular Evolution 35(3), 261–269 (1992).
45. J. S. Almeida, J. A. Carrico, A. Maretzek, P. A. Noble, and M. Fletcher, "Analysis of genomic sequences by chaos game representation," Bioinformatics 17(5), 429–437 (2001).
46. S. V. Korolev, V. G. Tumanyan, "Fractal dimensions of oligonucleotide compositions of dna sequences," Bioinformatics, Supercomputing and Complex Genome Analysis 635–638 (1993).
47. V. V. Solovyev, H. A. Lim, L. Milanesi, and C. Charles, "Application of fractal representation of genetic texts for recognition of genome functional and coding regions," Bioinformatics, Supercomputing and Complex Genome Analysis 609–622 (1993).
48. P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil, "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences," Molecular Biology and Evolution 16(10), 1391–1399 (1999).
49. Z. Sun, S. Pei, L. H. Rong, and S. S.-T. Yau, "A novel numerical representation for proteins: Three-dimensional chaos game representation and its extended natural vector," Computational and Structural Biotechnology Journal 18, 1904–1913 (2020).
50. R. Touati, S. Haddad-Boubaker, I. Ferchichi, I. Messaoudi, A. E. Ouesleti, H. Triki, Z. Lachiri, and M. Kharrat, "Comparative genomic signature representations of the emerging covid-19 coronavirus and other coronaviruses: High identity and possible recombination between bat and pangolin coronaviruses," Genomics 112(6), 4189–4202 (2020).
51. D. C. Sengupta, M. D. Hill, K. R. Benton, and H. N. Banerjee, "Similarity studies of corona viruses through chaos game representation," Computational Molecular Bioscience 10(3), 61 (2020).
52. G. S. Randhawa, M. P. M. Soltysiak, H. El Roz, C. P. E. de Souza, K. A. Hill, and L. Kari, "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study," PLoS ONE 15(4), e0232391 (2020).
53. T. Paul, S. Vainio, and J. Roning, "Detection of intra-family coronavirus genome sequences through graphical representation and artificial neural network," Expert Systems With Applications 194, 116559 (2022).
54. E. M. Anitas, "Small-angle scattering and multifractal analysis of DNA sequences," International Journal of Molecular Sciences 21(13), 4651 (2020).
55. E. M. Anitas, "Fractal analysis of DNA sequences using frequency chaos game representation and small-angle scattering," International Journal of Molecular Sciences 23, 1847 (2022).
56. T. D. Schneider, R. M. Stephens, "Sequence logos: a new way to display consensus sequences," Nucleic Acids Research 18(20), 6097–6100 (1990).
57. S. S. Ulyanov, O. V. Ulianova, S. S. Zaytsev, Y. V. Saltykov, and V. A. Feodorova, "Statistics on gene-based laser speckles with a small number of scatterers: implications for the detection of polymorphism in the Chlamydia trachomatis omp1 gene," Laser Physics Letters 15(4), 045601 (2018).
58. J. W. Goodman, Statistical Optics, 2nd ed, John Wiley & Sons (2015).
59. D. A. Zimnyakov, M. V. Alonova, An. V. Skripal, S. S. Zaitsev, and V. A. Feodorova, "Polarization analysis of gene sequence structures: mapping of extreme local polarization states," Journal of Biomedical Photonics & Engineering 8(4), 040322 (2022).
60. D. Zimnyakov, M. Alonova, An. Skripal, S. Dobdin, and V. Feodorova, "Quantification of the diversity in gene structures using the principles of polarization mapping," Current Issues in Molecular Biology 45(2), 1720–1740 (2023).

61. “Official hCoV-19 Reference Sequence, hCoV-19/Italy/LOM_ASSTMonza_6000027610_20220207092700/2022,” GISAID (accessed on 10 April 2023). [https://www.epicov.org/epi3/frontend#c8999].
62. “Official hCoV-19 Reference Sequence, hCoV-19/England/MILK-1681E7D/2021,” GISAID (accessed on 10 April 2023). [https://www.epicov.org/epi3/frontend#6379b4].