

Polarization- and CGR-based binary representations as identifiers of the nucleotide sequences in bioinformatics

D. A. Zimnyakov^{1,2,3}✉, M. V. Alonova¹, An. V. Skripal², M. G. Inkin², S. S. Zaitsev⁴,
V. A. Feodorova⁴

¹Yury Gagarin State Technical University of Saratov, Russia

²Saratov State University, Russia

³Institute for Problems of Precision Mechanics and Control of the Russian Academy of Sciences, Saratov, Russia

⁴Saratov State University of Genetics, Biotechnology and Engineering named after N. I. Vavilov, Russia

E-mail: ✉zimnyakov@mail.ru, alonova_marina@mail.ru, skripalav@info.sgu.ru,
sunbeam18.95@mail.ru, zaitsev-sergey@inbox.ru, feodorovav@mail.ru

Received 7.11.2023, accepted 28.02.2024, available online 28.05.2024, published 31.07.2024

Abstract. Purpose of this work is the comparative analysis of two approaches to the synthesis of two-dimensional binary identifiers of nucleotide sequences obtained using DNA sequencing of biological objects. *Methods.* One of the approaches is based on modeling the polarization-dependent diffraction of a coherent readout beam on a two-dimensional phase-modulating structure (phase screen) associated with the symbolic sequence obtained as a result of DNA sequencing. Another approach uses a two-dimensional representation of the symbolic sequence using a chaos game representation (CGR). To obtain a finite-element CGR mapping, it is fragmented into a given number of cells, ensuring acceptable sensitivity of the synthesized binary identifier to structural changes in the displayed sequence. *Results.* The comparative analysis was carried out using fragments of symbol sequences corresponding to various strains (Wuhan, Delta, Omicron) of the SarSCoV2 virus. In the course of the analysis, the correlation coefficients between the binary identifiers corresponding to various strains were obtained and compared with each other. *Conclusion.* It has been established that binary identifiers synthesized using the polarization encoding technique are characterized by significantly higher sensitivity to structural changes in the analyzed sequences and smaller sizes compared to CGR binary identifiers.

Keywords: nucleotide sequences, binary representation, polarization encoding, chaos game representation.

Acknowledgements. This work was supported by the Russian Science Foundation, grant No. 22-21-00194.

For citation: Zimnyakov DA, Alonova MV, Skripal AnV, Inkin MG, Zaitsev SS, Feodorova VA. Polarization- and CGR-based binary representations as identifiers of the nucleotide sequences in bioinformatics. *Izvestiya VUZ. Applied Nonlinear Dynamics.* 2024;32(4):439–459. DOI: 10.18500/0869-6632-003110

This is an open access article distributed under the terms of Creative Commons Attribution License (CC-BY 4.0).

Introduction

Analysis and visualization of genetic information obtained by sequencing DNA and RNA fragments of various biological objects [1] is one of the dominant areas of fundamental and applied research in modern bioinformatics. The objects of research in this case are symbolic sequences of varying length (from several hundred to several hundred thousand symbols) and their fragments, characterized by quasi-random distributions of four symbols (A, C, T, G). Symbols A, C, T, G are associated with the basic nucleotides that form the structure of the analyzed DNA fragment (A — adenine, C — cytosine, T — thymine, G — guanine [2]), and the order of their arrangement in the sequence is a unique attribute of this biological object. The quasi-randomness of nucleotide distributions in sequenced DNA fragments (and, accordingly, associated symbols in sequences) is due to both the existence of large-scale correlations between nucleotide positions in DNA and differences in the relative frequencies of basic nucleotides during random sampling of their positions in sequences.

Over the past three decades, a general strategy has been developed in bioinformatics for synthesizing mappings of DNA-associated symbol sequences in two-dimensional or three-dimensional Cartesian space. This strategy is based on the choice of a four-point basis, each point of which is related in a certain way to one of the four symbols. Currently, three types of bases are mainly used, differing in the order of the symbols when traversing the basis points clockwise: RY basis (ACGT), MK basis (AGCT) and WS basis (ACTG). The choice of one or another basis is made based on the possibility of obtaining additional information on the biochemical features of the analyzed DNA fragment; In particular, the mapping of the sequence in the RY basis allows for the frequency analysis of nucleotides by purine and pyrimidine groups. Adenine and guanine (A, G; secondary diagonal of the basis) belong to the class of pyrimidines, while cytosine and thymine (C, T; main diagonal of the basis) belong to the class of purines. Similarly, the mapping in the MK basis allows for the analysis of the distribution of nucleotides by amine and ketone groups, and the WS basis provides the opportunity to analyze the distribution of nucleotides by groups with weak and strong hydrogen bonds [3–5].

The sequential construction of sequence images is performed using various recursive procedures; as a result, either piecewise linear mappings in three-dimensional space (for example, Z-, C-, RY-, MK-, and WS-lines [6–10]) or point mappings in two-dimensional space are synthesized. A widely used two-dimensional point mapping of nucleotide sequences is the chaos game representation (CGR), first proposed by Jeffrey (H. Joel Jeffrey, [11]). According to the CGR algorithm, the synthesis of a two-dimensional point mapping of a symbol sequence is performed within a square region of unit dimensions, where the symbol A corresponds to the origin of coordinates (0.0; 0.0). The remaining points of the basis correspond to the coordinates (0.0; 1.0), (1.0; 1.0) and (1.0; 0.0). The correlation of the remaining symbols (C, G, T) and these points is made based on the selected basis (RY, MK or WS). The center of the square (0.5; 0.5) is chosen as the starting point and the mapping is synthesized in accordance with the arrangement of symbols in the sequence according to the rule that the next mapping point is located in the middle of the segment connecting the previous point and the point of the basis corresponding to the displayed symbol. This recursive procedure continues until the end of the symbol sequence. There are certain analogies between this algorithm and the algorithms for synthesizing two-dimensional fractal structures (for example, the Sierpinski carpet [12]). The existence of the above-mentioned large-scale correlations in the positions of symbols in sequences at sufficiently large lengths leads to fractal-like structures of the synthesized CGR mappings of some biological objects. A significant drawback of binary CGR maps for small lengths of symbolic sequences (from several hundred to several thousand symbols) is the low average surface density of display points, which does not allow reliable identification, for example, of the fractal nature of their

structure. In this regard, various modifications of this approach have been proposed in the last two decades, including the FCGR algorithm (frequency chaos game representation, [13–16]). The main idea of this modification is to divide the CGR mapping into equal-sized cells, count the number of display points within each cell and characterize the cells by the relative frequencies of points falling into them. Accordingly, the relative frequency as an informative parameter can be displayed in shades of gray [17].

It is obvious that changes in the structure of the displayed symbolic sequence in relation to the reference sequence, caused by mutational substitutions of nucleotides in the analyzed DNA fragment, lead to shifts of some display points in the synthesized CGR map in relation to the reference map. In other words, the synthesized map can be considered as a unique identifier of the symbolic sequence and, accordingly, of the given DNA fragment. The level of mutational changes can be quantitatively assessed by assessing the degree of mutual correlation of the analyzed and reference binary CGR structures.

In the works [17–19] an alternative approach to the synthesis of two-dimensional binary representations of nucleotide sequences was proposed based on modeling the effect of polarization modulation of a coherent light beam by a DNA-associated two-dimensional phase screen. Such modulation leads to the formation of a distribution of local polarization states of the diffracted beam in the far diffraction zone, described by three components of the normalized Stokes vector. It was shown [18] that two-dimensional binary representations formed by discrimination of spatial distributions of local values of the Stokes vector components can also be considered as identifiers of nucleotide sequences that are highly sensitive to mutational changes.

The aim of this work is a comparative analysis of two approaches to two-dimensional binary identification of nucleotide sequence fragments, one of which is based on the CGR map synthesis technique, and the second uses the principle of polarization encoding of nucleotide sequences and identification of the limiting states of polarization of the diffracted coherent light field. The analyzed objects are fragments of symbolic sequences associated with three different strains of the SARS-CoV-2 virus (covid) — «Wuhan», «Delta» and «Omicron». The SARS-CoV-2 virus continues to pose a high potential threat to humanity [20] due to its high antigenic variability [21]. Among the genes encoding SARS-CoV-2 proteins, the spike protein, or S protein, which affects the penetration of the virus into host cells, is of considerable interest for research [22, 23]. Moreover, the S protein can spread in the body separately from the virus, being released from infected viral particles, being detected in various organs and tissues, damaging the cells of the macroorganism; fragments of this protein can penetrate the blood-brain barrier [24]. It should also be noted that the nucleotide sequence encoding the biosynthesis of the S protein [25] is highly mutable compared to more conservative genes expressing the synthesis of other SARS-CoV-2 polypeptides. The nucleotide sequence of the «S» gene encoding the spike protein is located in the virus genome at positions 21563–25384 and is 3822 nucleotides (1274 amino acids) long [26]. The choice of the Wuhan, Delta and Omicron strains of the SARS-CoV-2 virus is due to the fact that the Wuhan strain is usually considered a reference strain, while the Delta and Omicron strains, as well as their sublineages, are considered to pose a greater threat compared to other strains due to their “highest contagiousness” [27].

1. Description of the analyzed symbolic sequences

The symbol sequences were taken from the open-access database GISAID (Global Initiative on Sharing All Influenza Data, open access by subscription), in which the Wuhan strain corresponds to reference [28], Delta — [29], and Omicron — [30]. The number of A, C, T, G symbols in the sequence fragments associated with the «S» gene is 3822. Accordingly, the number of triplets

Table 1. Differences in the triplet sequences for three displayed strains

Position of triplets in sequences	«Wuhan»	«Delta»	«Omicron»
19	ACA	AGA(!)	ATA(!)
95	ACT	ATT(!)	ACT
142	GGT	GAT(!)	GAT(!)
213	GTG	GTG	GGG(!)
339	GGT	GGT	GAT(!)
371	TCC	TCC	TTC(!)
373	TCA	TCA	CCA(!)
375	TCC	TCC	TTC(!)
376	ACT	ACT	GCT(!)
405	GAT	GAT	AAT(!)
408	AGA	AGA	AGC(!)
410	ATC	ATC	ATT(!)
417	AAG	AAT(!)	AAT(!)
440	AAT	AAT	AAG(!)
452	CTG	CGG(!)	CTG
477	AGC	AGC	AAC(!)
478	ACA	AAA(!)	AAA(!)
484	GAA	GAA	GCA(!)
493	CAA	CAA	CGA(!)
498	CAA	CAA	CGA(!)
501	AAT	AAT	TAT(!)
505	TAC	TAC	CAC(!)
614	GAT	GGT(!)	GGT(!)
655	CAT	CAT	TAT(!)
679	AAT	AAT	AAG(!)
681	CCT	CGT(!)	CAT(!)
764	AAC	AAC	AAA(!)
796	GAT	GAT	TAT(!)
925	AAC	AAC	AAT(!)
950	GAT	AAT(!)	GAT
954	CAA	CAA	CAT(!)
969	AAT	AAT	AAA(!)
1146	GAC	GAC	GAT(!)

(amino acids) in the analyzed fragments is 1274. Table 1 shows the differences in the triplet sequences between the three strains under consideration; the Wuhan strain is usually considered as the reference; different triplets are marked with the symbol (!).

Thus, the symbolic sequence for the Delta strain differs from the reference sequence for the Wuhan strain by 9 triplets, while the sequence for the Omicron strain has 30 different triplets. Note that all differences are due to mutational substitutions of single nucleotides in each of the different triplets.

2. Polarization coding of DNA-associated symbolic sequences and synthesis of two-dimensional binary identifiers

As noted above, the technique of virtual polarization encoding of DNA-associated symbolic sequences considered in [18] consists in representing the analyzed sequence or its fragment by a two-dimensional phase-modulating structure (a phase screen containing $2\tilde{N}_t \times 2\tilde{N}_t$ elements, where \tilde{N}_t is the number of triplets in the analyzed structure). The phase screen is read by a collimated coherent beam with a linear polarization state, and in the far diffraction zone (the focal plane of the Fourier transform lens), the spatial distributions of local polarization states of the diffracted beam are analyzed. Fig. 1 presents a physical interpretation of the simulated procedure for reading a DNA-associated phase screen.

The local polarization states are determined by the values of the components of the Stokes vector ($s_{k,m}^0 \div s_{k,m}^3$). When synthesizing the binary identifier of the coded symbolic sequence, the spatial positions in the Fourier plane of the limiting states of the fourth component of the Stokes vector $s_{k,m}^3$ are determined, characterizing the contribution of the circularly polarized component to the polarization state of the diffracted beam at the point (k, m) of the Fourier plane. Note that, within the framework of the used formalism, a discrete set of $2\tilde{N}_t \times 2\tilde{N}_t$ local phase shifts introduced by the synthesized phase screen into the reading coherent beam correspond to $4\tilde{N}_t \times 4\tilde{N}_t$ points (pixels) of the Fourier plane, uniquely determined by the two-dimensional discrete Fourier transform. The criterion for selecting the limit states $s_{k,m}^3$ is the condition $|s_{th}^3| \leq |s_{k,m}^3| < |\pm 1|$, where the threshold value s_{th}^3 is chosen close to -1 (in the case of discrimination of left-circular polarization states) or to 1 (in the case of discrimination of right-circular states). The choice of $s_{k,m}^3$ as an identification parameter is due to the fact that polarization states close to circular are characterized by maximum sensitivity to local changes in the structure of the synthesized phase screen compared to the linearly polarized components of the diffracted field determined by the components $s_{k,m}^1$ and $s_{k,m}^2$ [18].

When synthesizing a DNA-associated phase screen, it is generated as an ensemble of $\tilde{N}_t \times \tilde{N}_t$ submatrices of size (2×2) , each of which is associated with a certain triplet in the displayed

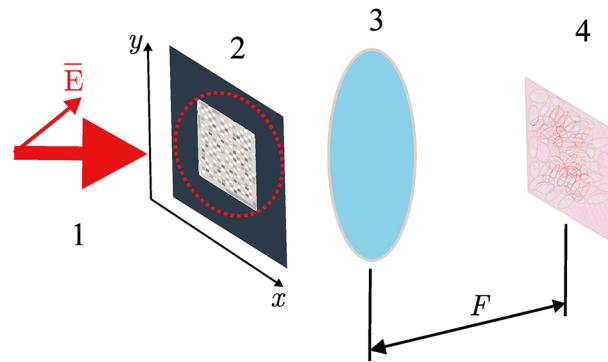


Fig 1. Physical interpretation of the procedure for polarization encoding of a DNA-associated symbol sequence and the synthesis of a two-dimensional binary identifier. The DNA-associated phase screen (2) is illuminated by a linearly polarized collimated laser beam (1) with a plane of polarization oriented at an angle of 45° to the sides of the screen (the dotted line marks the illuminated area of the screen). The polarization-dependent diffraction structure (4) is formed in the focal plane of the Fourier-transforming lens. The figure does not show the polarization-analyzing unit, located between the lens (3) and the Fourier plane (4) and used to form spatial distributions ($s_{k,m}^0 \div s_{k,m}^3$)

sequence of A, C, T, G symbols. Obviously, the size $\tilde{N}_t \times \tilde{N}_t$ of the displayed set of triplets in the sequence is determined by the maximum value of the square of an integer that does not exceed the number of triplets in the sequence. To illustrate this point, consider the symbol sequence corresponding to the «S» gene of the «Wuhan» strain of the SARS-CoV-2 virus. There are 3822 nucleotides in the sequenced DNA region (respectively, 1274 triplets, see section 1). Accordingly, the maximum possible value of \tilde{N}_t in the synthesis of the DNA-associated phase screen is 35. $\tilde{N}_t^2 = 1225$ and 49 triplets are not included in the synthesis of the phase screen; however, following the data presented in Table 1, it can be concluded that the discarded fragments of the sequences for the Wuhan, Delta, and Omicron strains do not contain any differences.

Each submatrix in the $\tilde{N}_t \times \tilde{N}_t$ set displays the properties of the set of base nucleotides in the corresponding triplet in accordance with the selected coding rule. As an example of a possible relationship between the elements of the submatrix ($b_{0,0} \div b_{2,2}$) and the base nucleotides, consider the following rule:

$$b_{0,0} \rightarrow A; b_{1,0} \rightarrow C; b_{0,1} \rightarrow T; b_{1,1} \rightarrow G. \quad (1)$$

The value of an element is determined by the number of nucleotides of a given type in a triplet; accordingly, the universal rule for all submatrices in a set is that the sum of their elements is always 3. As an illustration, Fig. 2 displays in grayscale the structure of the synthesized phase screen for the «S» gene of the «Wuhan» strain, used as a reference object.

The phase screen matrix $(a_{i,j})_{2\tilde{N}_t \times 2\tilde{N}_t}$, formed by successive row-wise and column-wise combination of submatrices $(b_{i',j'})_{2 \times 2}$, corresponding to triplets, is used for phase modulation of the reading linearly polarized beam. Since the reading beam contains x - and y -polarized components with equal values of amplitude and initial phase (the polarization plane of the reading beam forms angles of 45° with the sides of the phase screen), then, applying the phase modulation law of the following form:

$$\begin{aligned} (\Delta\varphi_{ij})_{2\tilde{N}_t \times 2\tilde{N}_t}^x &= \pi(a_{i,j})_{2\tilde{N}_t \times 2\tilde{N}_t}, \\ (\Delta\varphi_{ij})_{2\tilde{N}_t \times 2\tilde{N}_t}^y &= 0.5\pi + \pi(a_{i,j})_{2\tilde{N}_t \times 2\tilde{N}_t}, \end{aligned} \quad (2)$$

we obtain behind the screen a set of $2\tilde{N}_t \times 2\tilde{N}_t$ local states of circular polarization of the boundary light field, the differences between which are due to two possible random values of the phase of

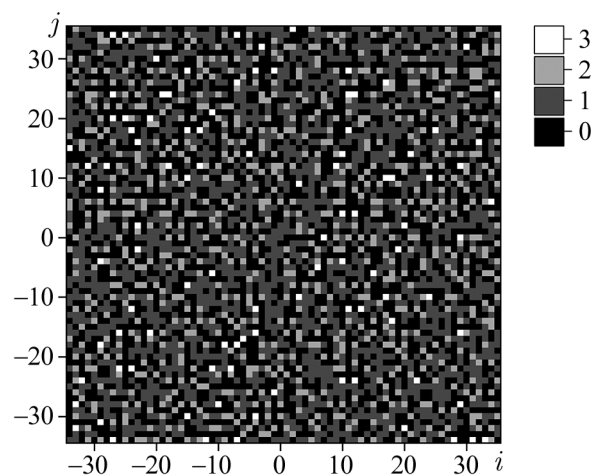


Fig 2. Structure of the DNA-associated phase screen for the «S» gene of the «Wuhan» strain

the x -polarized component — 0 (when $a_{i,j} = 0$ or 2) and π (when $a_{i,j} = 1$ or 3). The «mixing» of the partial components of the reading coherent field in the far diffraction zone that have passed through different elements of the phase screen can be described by the two-dimensional discrete Fourier transform of the x - and y -polarized components of the boundary field (see, for example, [31]):

$$E_{k,m}^{x,y} = \frac{1}{4\tilde{N}_t^2} \sum_{i=-\tilde{N}_t}^{\tilde{N}_t-1} \sum_{j=-\tilde{N}_t}^{\tilde{N}_t-1} \exp[-\tilde{j} \cdot K_{sc} \cdot \{(\pi/\tilde{N}_t)(k \cdot i + m \cdot j) - \Delta\varphi_{ij}^{x,y}\}], \quad (3)$$

where \tilde{j} is the imaginary unit, indices k, m define discrete pixel coordinates in the Fourier plane, and the coefficient K_{sc} determines the scale of displaying the spatial distribution of the diffracted readout beam in the Fourier plane. The maximum permissible value of K_{sc} is 0.5 and corresponds to the largest-scale («panoramic») display of the Fourier spectrum of the phase screen. Using values of the scale factor K_{sc} greater than 0.5 leads to distortion of the simulated spatial distributions of the amplitude of the x - and y -polarized components of the diffraction field due to the frequency substitution effect [32]. A decrease in K_{sc} corresponds to greater detail in displaying the central region of the diffracted field with a simultaneous decrease in the size of the displayed area. Note that when writing the expression (3), the assumption was made about single values of the amplitudes of the x - and y -polarized components of the reading beam, which does not violate the generality of the consideration being carried out.

For the model discrete distributions $E_{k,m}^{x,y}$ obtained in this way, the components of the Stokes vector are then calculated in accordance with the following expressions (see, for example, [33]):

$$\begin{cases} s_{k,m}^0 = (|E_{k,m}^x|^2 + |E_{k,m}^y|^2)/2, \\ s_{k,m}^1 = (|E_{k,m}^x|^2 - |E_{k,m}^y|^2)/2s_{k,m}^0, \\ s_{k,m}^2 = 2|E_{k,m}^x||E_{k,m}^y| \cos(\delta_{k,m})/2s_{k,m}^0, \\ s_{k,m}^3 = 2|E_{k,m}^x||E_{k,m}^y| \sin(\delta_{k,m})/2s_{k,m}^0, \end{cases} \quad (4)$$

where $\delta_{k,m}$ are the values of the phase differences of the x - and y -polarized components at the corresponding points of the Fourier plane. The normalized values ($s_{k,m}^1 \div s_{k,m}^0$) take values in the intervals from -1 to 1 and satisfy the fundamental relation of polarization optics:

$$(s_{k,m}^1)^2 + (s_{k,m}^2)^2 + (s_{k,m}^3)^2 = 1. \quad (5)$$

Thus, with a significant contribution of the right or left circularly polarized components to the polarization state of the diffracted beam at point k, m of the Fourier plane, the following relationships hold between the normalized components of the Stokes vector: $s_{k,m}^3 \rightarrow \pm 1$; $s_{k,m}^1, s_{k,m}^2 \rightarrow 0$.

As an example, Fig. 3 shows the distributions of $s_{k,m}^3$ values for «S»-associated symbol sequences corresponding to the «Wuhan», «Delta», and «Omicron» strains; their construction uses the scale factor K_{sc} equal to 0.1 (detailed display). To identify differences between the $s_{k,m}^3$ distributions caused by substitutions of some nucleotides (and, accordingly, symbols) in the analyzed sequences, they can be binarized according to the following rules:

$$\begin{cases} 1 \geq s_{k,m}^3 \geq s_{th}^3 \rightarrow \tilde{s}_{k,m}^3 = 1, \\ s_{k,m}^3 < s_{th}^3 \rightarrow \tilde{s}_{k,m}^3 = 0, \end{cases} \quad (6)$$

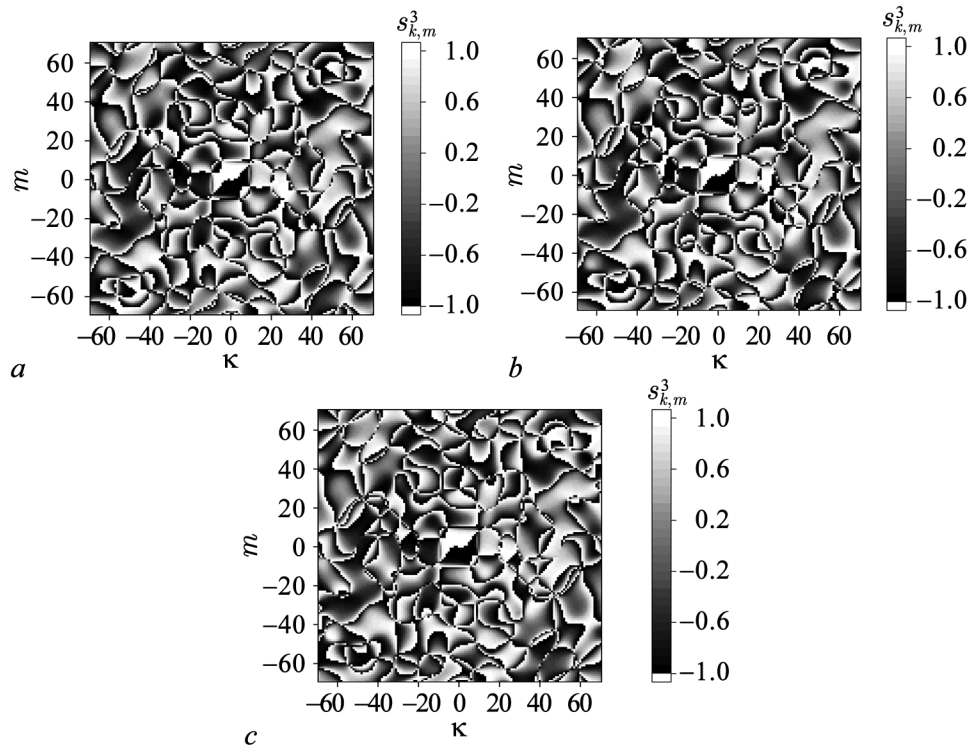


Fig 3. Detailed (small-scale) ($K_{sc} = 0.1$) representations of normalized values of the fourth component of the Stokes vector for the “Wuhan” (a), “Delta” (b) and “Omicron” (c) strains

in the case of identifying limiting states close to right circular polarization, and, accordingly,

$$\begin{cases} -1 \leq s_{k,m}^3 \leq s_{th}^3 \rightarrow \tilde{s}_{k,m}^3 = 1, \\ s_{k,m}^3 > s_{th}^3 \rightarrow \tilde{s}_{k,m}^3 = 0, \end{cases} \quad (7)$$

when identifying limiting states close to left circular polarization. The threshold value s_{th}^3 is chosen close to 1 (for right circular polarization) or to -1 (respectively, for left circular polarization).

Fig. 4 shows the results of applying a similar procedure to the distributions shown in Fig. 3 with a discrimination threshold of -0.98 . Comparison of binary distributions can be made by pixel-by-pixel logical multiplication of the reference (strain «Wuhan») and analyzed distributions. The results of this procedure are shown in Fig. 5. Quantitatively, the degree of correspondence between the analyzed and reference identifiers can be expressed using the correlation coefficient

$$R^{a,r} = \frac{\sum_{m=1}^{4\tilde{N}_t} \sum_{k=1}^{4\tilde{N}_t} a_{m,k} \times r_{m,k}}{\sum_{m=1}^{4\tilde{N}_t} \sum_{k=1}^{4\tilde{N}_t} r_{m,k}}, \quad (8)$$

where symbols $a_{m,k}$ and $r_{m,k}$ refer respectively to the pixels of the analyzed and reference identifiers.

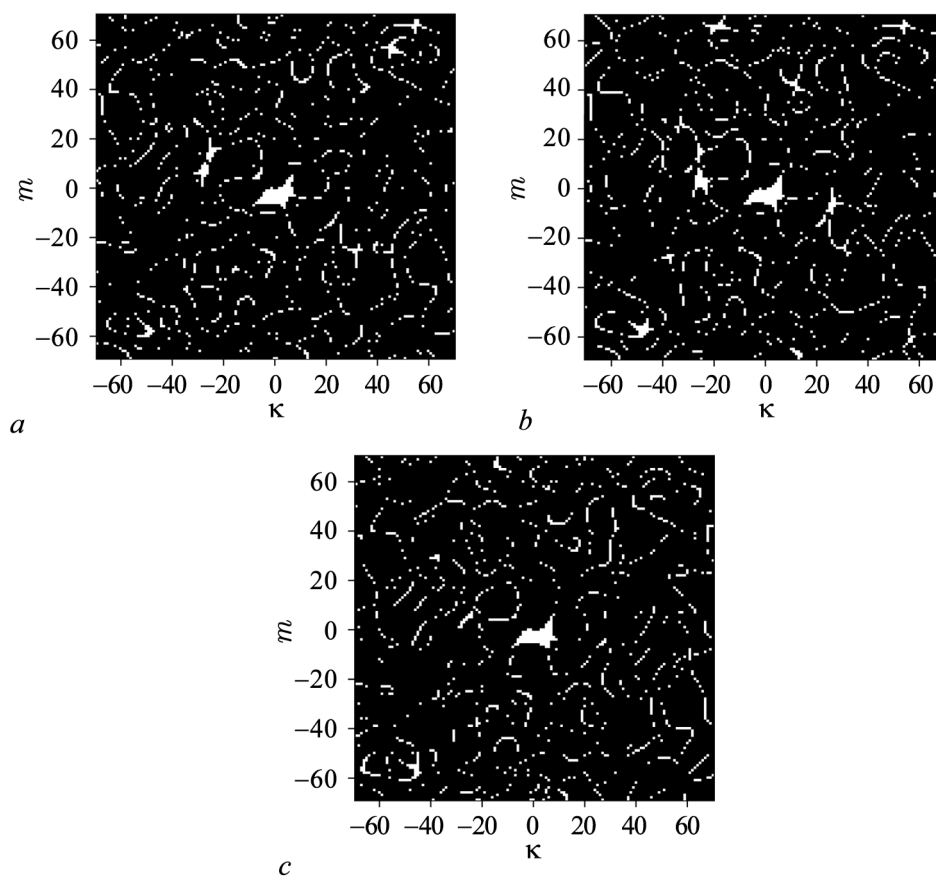


Fig 4. Binary identifiers of “S” genes of strains “Wuhan” (*a*), “Delta” (*b*) and “Omicron” (*c*), obtained using polarization encoding. Local polarization states close to the left circular polarization are shown for the discrimination threshold $s_{th}^3 = -0.98$

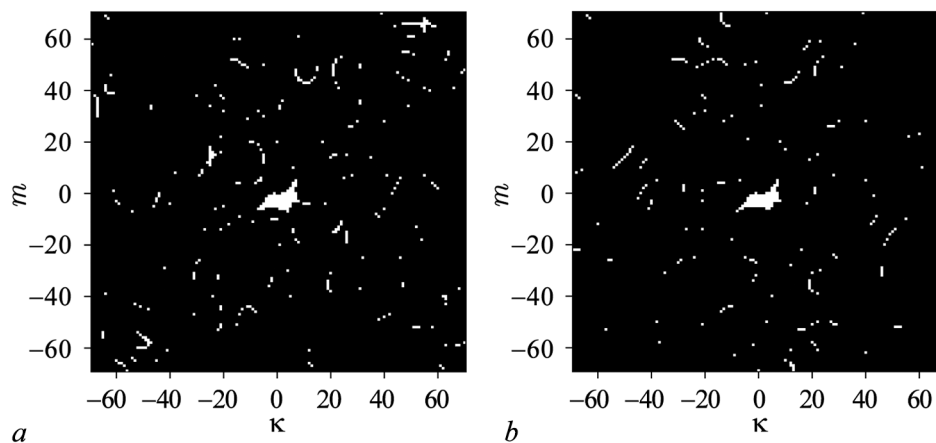


Fig 5. Results of pixel-by-pixel logical multiplication of a binary identifier for the reference sequence (“Wuhan” strain, Fig. 4, *a*) and binary identifiers for the “Delta” (*a*) and “Omicron” (*b*) strains

3. CGR mapping of DNA-associated character sequences

CGR maps as two-dimensional discrete mappings of symbolic sequences, considered as an alternative to the above-described method of polarization binary encoding, are synthesized in a square region with a unit side in accordance with the following scheme. Depending on the basis used (RY, MK or WS), the vertices of the square with Cartesian coordinates (0, 0), (0, 1), (1, 1) and (1, 0) (the direction of traversal is clockwise) are associated in a certain way with four basic nucleotides. For the RY basis, the following associations take place: A → (0,0), C → (0,1), G → (1,1), T → (1,0); for the MK basis — A → (0,0), G → (0,1), C → (1,1), T → (1,0), and for the WS basis — A → (0,0), C → (0,1), T → (1,1), G → (1,1). The starting point for construction is chosen at the center of the square ((0.5, 0.5)). The first symbol in the sequence is displayed by the point corresponding to the midpoint of the segment between the starting point and the vertex of the square corresponding to the symbol to be displayed.

The next symbol is displayed by the point corresponding to the midpoint of the segment between the previous display point and the vertex of the square associated with the symbol. This recursive procedure is repeated until the final symbol of the sequence. As an example of such a procedure, let us consider the result of generating the coordinates of the mapping points for the first 5 members of the «S»-associated symbolic sequence for the «Wuhan» strain (ATGTT...) in the RY basis: (0.5;0.5), (0.25;0.25), (0.625;0.125), (0.813;0.563), (0.906;0.281), (0.953;0.141). Formally, the algorithm for generating mapping points can be represented by the following expressions:

$$\begin{cases} x_n = \frac{1}{2^{n+1}} + \sum_{k=1}^n \frac{C_x^k}{2^{n-k+1}}, \\ y_n = \frac{1}{2^{n+1}} + \sum_{k=1}^n \frac{C_y^k}{2^{n-k+1}}, \end{cases} \quad (9)$$

where the coefficients C_x^k and C_y^k determine the coordinates of the basis vertex corresponding to the k -th nucleotide in the sequence. Obviously, the mapping points are always located inside the unit square or at least on its boundaries.

Fig. 6 shows as an example a CGR map of the symbolic sequence for the «S»-gene of the «Wuhan» strain in the RY-basis. It should be noted that the spatial distribution of the mapping points is significantly heterogeneous (the presence of both zones with high local filling density

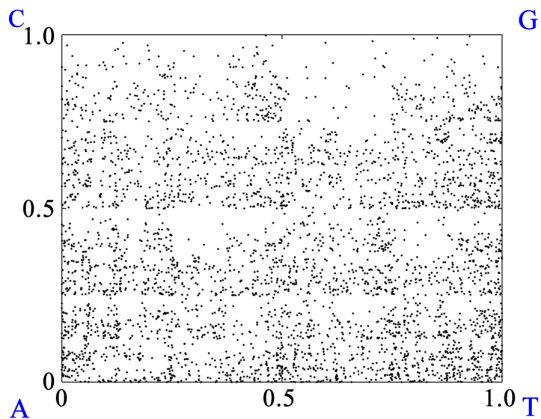


Fig 6. CGR map of the symbolic sequence for the “S” gene of the “Wuhan” strain in the RY basis

and lacunae). For some nucleotide sequences, such spatial heterogeneity allows us to speak of the "fractality" of the CGR mapping, in a certain sense analogous, for example, to the classical Sierpinski carpet. As an example, we can refer to the work of [34], which clearly illustrates the fractal nature of CGR mappings of large length (on the order of several hundred thousand nucleotides).

Taking into account the quasi-continuous nature of the distributions of the x - and y -coordinates of the mapping points in the synthesized CGR mappings of symbol sets, the analysis of the correlations of the mappings of

the reference and analyzed sequences in this case differs from the above-considered algorithm of logical multiplication of binary matrices of relatively small size, as in the case of polarization encoding (expression (8)). It should be noted that the quasi-continuous nature of the distributions of the coordinates of the mapping points in the case of CGR mappings is not consistent with the concept of synthesizing two-dimensional binary identifiers of relatively small sizes for DNA-associated symbol sequences. The solution to this problem is reduced to fragmentation («coarsening») of CGR maps, which is to a certain extent similar to the above-mentioned FCGR algorithm, although it differs significantly from it. At a qualitative level, this solution is formulated as follows: it is necessary to determine the minimum acceptable level of partitioning the CGR map into equal-sized cells, at which the probability of two or more display points falling into an arbitrarily selected fragment is less than a given threshold value. In other words, each fragment may either contain a single display point or not.

Fig. 7 shows the results of statistical analysis of the process of «coarsening» of CGR maps for the «Wuhan» strain in RY, MK and WS bases depending on the parameter N/N_n (N is the number of coordinate axes partitions, respectively, the number of CGR map fragments is N^2 ; N_n is the number of symbols in the displayed sequence). Curves 1–3 display the values of the probability of finding several (more than 1) display points P_s in the cells of RY, MK and WS CGR maps, allocated by random sampling, depending on N/N_n . For comparison, a similar dependence (4) is also shown for the case of a uniform distribution of display points over the CGR map. Note the significantly larger values of W for the analyzed symbolic sequence in comparison with the uniform distribution, due to the significantly non-uniform spatial distributions of the display points in the synthesized RY, MK and WS CGR maps (see Fig. 6 for the RY map).

Fig. 8 presents the calculated values of the correlation coefficient $R^{a,r}$ (see expression (8)) between the reference (strain «Wuhan») and analyzed (strains «Delta» (1) and «Omicron» (2)) fragmented CGR maps depending on the ratio N/N_n . It should be noted that the $R^{a,r} = f(N/N_n)$ dependencies for N/N_n exceeding 0.3 demonstrate virtually constant values of the correlation coefficient, which allows us to propose this fragmentation level as the minimum for synthesizing binary identifiers based on fragmented CGR maps. The probability of detecting

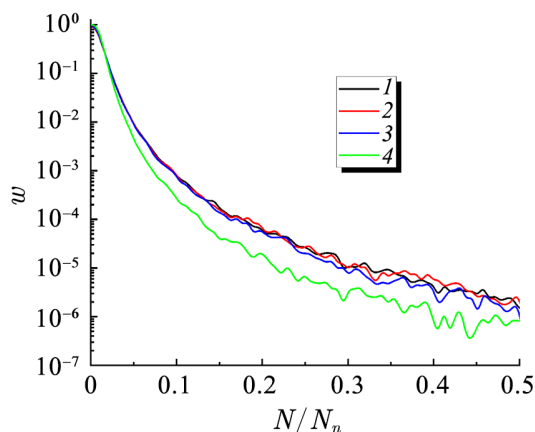


Fig 7. Model values of the probability of detecting a CGR map cell containing more than one mapping point, depending on the level of map fragmentation N/N_n . 1 – RY basis; 2 – MK basis; 3 – WS basis; 4 – uniform distribution of mapping points over the map area (color online)

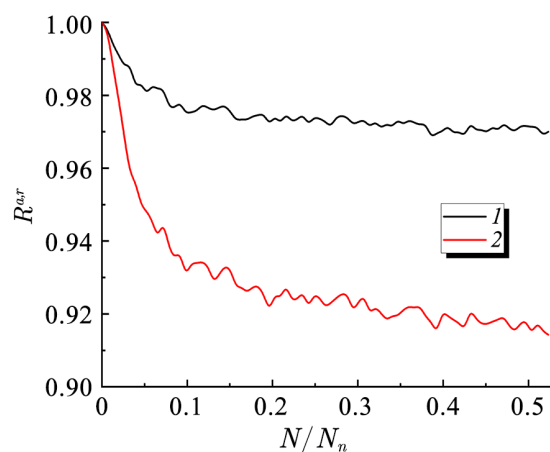


Fig 8. Correlation coefficients between binary identifiers of “S”-associated reference (strain “Wuhan”) and analyzed sequences (strain “Delta” (1); strain “Omicron” (2)) depending on N/N_n (RY basis) (color online)

cells with more than one display point, which are assigned single values during fragmentation, is small and amounts to approximately 2×10^{-5} . Taking into account the number of symbols in the analyzed sequences (3822), the number of partitions of each of the 2 coordinate axes of the CGR map is at least 1147. Accordingly, the total number of pixels in the binary identifier synthesized in this way is at least 1315609. Note that the use of other bases (MK, WS) for constructing the original CGR maps leads to similar results.

4. Discussion of results

It is advisable to conduct a comparative analysis of the two approaches considered in relation to the synthesis of two-dimensional binary identifiers of DNA-associated symbolic sequences by comparing two key parameters: the number of elements (pixels) in the identifier and the sensitivity of the binary structure of the identifier to structural changes in the analyzed symbolic sequence with respect to the reference sequence. As the last parameter, it is advisable to consider the correlation coefficient $R^{a,r}$, defined by the expression (8). In the case of polarization encoding of symbolic sequences, the number of pixels in the synthesized identifiers is determined by the number of used triplets \tilde{N}_t^2 and is equal to $16\tilde{N}_t^2$. At the same time, the correlation coefficient $R^{a,r}$ significantly depends on the used discrimination threshold s_{th}^3 of the limiting states of polarization.

In the case of synthesizing binary identifiers based on fragmented CGR maps, both the number of pixels and the correlation coefficients between the reference and analyzed identifiers depend on the fragmentation level used. These features are illustrated by the data presented in Table 2 and 3.

Table 2. Parameters of binary identifiers synthesized on the base of polarization encoding

Threshold of discrimination s_{th}^3	Number of pixels in binary identifier	Correlation coefficient $R^{a,r}$ for the pair «Wuhan – Delta»	Correlation coefficient $R^{a,r}$ for the pair «Wuhan – Omicron»
-0.92	19600	≈ 0.49	≈ 0.41
-0.94	19600	≈ 0.45	≈ 0.36
-0.96	19600	≈ 0.41	≈ 0.31
-0.98	19600	≈ 0.34	≈ 0.25
-0.999	19600	≈ 0.21	≈ 0.13

Table 3. Parameters of binary identifiers synthesized on the base of fragmentation of the CGR maps

Level of fragmentation N/N_n	Number of pixels in binary identifier	Correlation coefficient $R^{a,r}$ for the pair «Wuhan – Delta»	Correlation coefficient $R^{a,r}$ for the pair «Wuhan – Omicron»
0.05	36864	≈ 0.982	≈ 0.951
0.1	146689	≈ 0.977	≈ 0.935
0.2	585225	≈ 0.973	≈ 0.926
0.3	1315609	≈ 0.972	≈ 0.926
0.4	2337841	≈ 0.972	≈ 0.919
0.5	3651921	≈ 0.972	≈ 0.918

A comparison of the presented data sets allows us to conclude that binary identifiers synthesized using polarization encoding have significant advantages over identifiers obtained by fragmenting CGR mappings. An increase in the absolute value of the discrimination threshold in the first case leads not only to an increase in the sensitivity of the correlation coefficient to the substitutions of individual nucleotides in the analyzed sequence relative to the reference one, but also to a decrease in the number of pixels with single values in the identifier. Thus, at a discrimination threshold of -0.92 , the number of pixels with single values is 2256, while for a discrimination threshold of -0.999 , this number is 266. Earlier [18] it was noted that the polarization encoding method should be used to compare nucleotide sequences that differ in relatively small numbers of nucleotides (from 1 to 10). In particular, with artificial random substitutions of one of the symbols in the considered fragment of the sequence for the “Wuhan” strain (3822 symbols) and a discrimination threshold of -0.99 , the value of the correlation coefficient $R^{a,r}$ between the binary polarization identifiers of the original and modified fragments is $\approx 0.556 \pm 0.036$ [18]. The average value of $R^{a,r}$ and the confidence interval corresponding to the significance level of 0.9 were obtained for a sample of 30 randomly modified fragments. Note that an almost twofold decrease in the correlation coefficient (from 1.0 to 0.556) occurs with the replacement of only one symbol associated with one of the 4 base nucleotides (i.e., with a difference in the structures of the original and modified sequences of about 0.026%). According to the data in Table 2, with an increase in the absolute value of the discrimination threshold to values close to 1, the sensitivity of $R^{a,r}$ to changes in the sequence structure increases significantly. On the other hand, the considered fragment of the sequence corresponding to the S protein in the SARS-CoV-2 genome is characterized by a relatively short length. The issue of the maximum sensitivity of binary polarization identifiers to small changes in the structure of long sequences (of the order of several tens of thousands of elements or more) will be studied further.

Based on the developed algorithms for the synthesis of binary polarization and CGR identifiers of DNA-associated symbolic sequences, trial program texts were implemented in C++. The programs provided reading of the initial symbol sequences of arbitrary length from a text file (*.txt), conversion of symbol (char) data into integer values, processing of the obtained arrays of integer data with their simultaneous conversion to double-precision floating-point format (double) and output to a text file (*.dat) of synthesized binary identifiers, which are two-dimensional arrays of integer data. In accordance with the above-described methods of synthesizing binary identifiers, the size of the formed square matrix is determined by the length of the initial symbol sequence, and the matrix elements are single-bit values, taking the values of either 1 or 0. The synthesis of polarization identifiers was carried out in 2 stages: conversion of the symbol sequence into a phase-modulating matrix (1, the original program convert.cpp, the executable file convert.exe); synthesis of the corresponding binary distribution of limit states of circular polarization in the Fourier plane (2, source program pol_map.cpp, executable file pol_map.exe). When synthesizing CGR identifiers, a single executable file cgr_map.exe was used (corresponding source text cgr_map.cpp). Note that the programs are pilot versions intended only for verification of the methods of binary mapping of nucleotide sequences discussed in this work. It is possible to optimize them in order to save computing resources (the amount of RAM used and computing time), but these issues are beyond the scope of this study and are planned for future work.

Nevertheless, the time costs for synthesizing binary polarization and CGR identifiers were estimated using the executable files convert.exe, pol_map.exe, and cgr_map.exe. The initial data were fragments of character sequences of various lengths (in the range from 3822 to 492 characters), obtained as a result of step-by-step exclusion of terminal groups of 333 characters (111 triplets) from the original «S»-associated sequence for the Wuhan strain. The executable files

were tested on a personal computer with an AMD Ryzen 9 3900X processor (12 cores, 3.79 GHz) and 32 GB of RAM. Fig. 9 shows the execution times of the files `convert.exe`, `pol_map.exe`, `cgr_map.exe` and the sizes of the synthesized polarization and CGR identifiers depending on the length of the character sequence. It should be noted that for relatively short symbolic sequences (with the number of elements up to ≈ 1000) the execution times for all programs are comparable and are determined mainly by the duration of the procedures for loading the original symbolic data from the hard disk into RAM and outputting the synthesized binary identifiers from RAM to the hard disk. With an increase in the number of symbols in the sequences, the execution time of the `pol_map.exe` program increases significantly compared to the execution time of `cgr_map.exe` (approximately a fivefold difference for 3822 symbols). This is due to the use of a fairly lengthy procedure for calculating the two-dimensional discrete Fourier transform when synthesizing polarization identifiers compared to a significantly faster distribution of CGR mapping points over the cells of binary CGR identifiers. However, it should be borne in mind that, despite the more time-consuming process, binary polarization identifiers are characterized by a significantly higher sensitivity to structural changes in the analyzed sequences (see Table 2, 3) and significantly smaller sizes (Fig. 9) compared to binary CGR identifiers. Binary polarization and CGR identifiers were also synthesized for the complete genome of the Wuhan strain, described by 29891 symbols [28]. In the first case, the synthesis time was 849 s with an identifier size of 396×396 pixels; in the case of the CGR identifier, the synthesis time was 24.85 s with a size of 14945×14945 pixels. It should be noted that the tested file `pol_map.exe` uses a fairly simple “linear” procedure for processing the read symbolic data, which can be optimized to reduce the computation time (for example, by using the fast Fourier transform (FFT) algorithm instead of the usual discrete Fourier transform, parallelizing the computational process, etc.).

One of the key problems in applying the method for synthesizing binary polarization identifiers considered in the work is the partial loss of information about the structure of the sequence when it is transformed into a square phase-modulating matrix with a smaller number of elements than the number of triplets of the analyzed sequence. In the considered examples using symbolic sequences for the Wuhan, Delta, and Omicron strains of the SARS-CoV-2 virus,

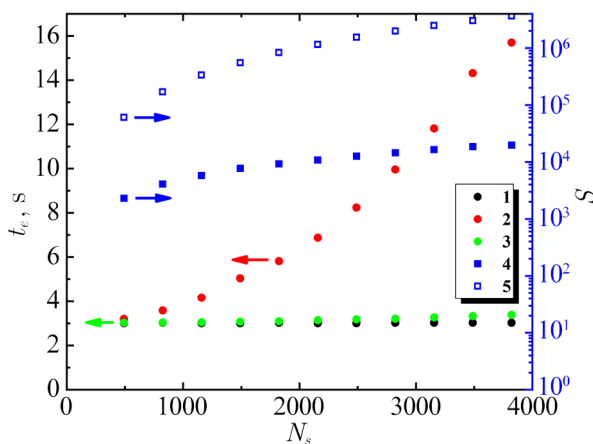


Fig 9. The execution time t_e for the files `convert.exe` (1), `pol_map.exe` (2), `cgr_map.exe` (3) and the sizes S of binary polarization (4) and CGR (5) identifiers depending on the number of symbols in the processed sequence. Arrows show the assignment of data to axes (color online)

this problem is not important, since all differences in the structures of the strains appear within the allocated section containing 1225 triplets (3675 symbols). However, for other objects, there may be a certain probability of structural inconsistencies falling into the sequence fragments discarded during the transformation. In our opinion, there are two possible ways to circumvent this problem in further studies. The first of these is to associate with the analyzed sequence not one, but two binary polarization identifiers of the same size, one of which is synthesized when reading a fragment of the symbolic sequence, starting with the first symbol in the forward direction, and the second when reading a fragment of the same length in the reverse direction, starting with the last symbol. Thus, the features of the sequence structure will be

fully displayed by this pair of binary identifiers. Another approach can be based on representing the sequence structure in the form of a rectangular rather than a square phase-modulating matrix, the dimensions of which are determined as a result of representing the number of triplets in the sequence as a product of two integers close to each other in magnitude. Accordingly, the binary identifier synthesized in this way will also represent a rectangular matrix. In the limiting case, when the number of triplets in the symbolic sequence is prime, it is possible to construct a "degenerate" polarization identifier in the form of a matrix of size $4 \times 4N_T$, where N_T is the total number of triplets in the sequence. The synthesis time of such an identifier will be significantly less compared to the time of formation of a two-dimensional identifier. Verification and comparative analysis of these approaches, as well as optimization of algorithms and programs for synthesizing binary polarization identifiers, are the objects of further research.

The procedure for synthesizing binary polarization identifiers of nucleotide sequence fragments can be implemented not only using only computer data processing, but also within the framework of a hybrid (instrumental and software approach). In this case, in accordance with Fig. 1, the reading of the DNA-associated phase-modulating matrix is carried out by a collimated beam of linearly polarized continuous laser radiation, the source of which can be, for example, a single-mode helium-neon laser GN-5P. A computer-controlled liquid crystal spatial phase modulator of the LS2012 type with a resolution of 1024×768 pixels and an 8-bit representation of the encoded information (manufactured by Holoeye Photonics AG, Germany) can be used as a phase-modulating screen 2. Reading of DNA-associated diffraction patterns in the focal plane of the Fourier transform lens 3 can be carried out using a polarization-sensitive CMOS camera Kiralux CS505MUP1 (resolution — 2448×2448 pixels, 12-bit representation of read data, official distributor — Thorlabs, USA). Pixel size of the liquid crystal matrix LS2012 equals 36 microns; camera Kiralux CS505MUP1 has a pixel size of $3.45 \mu\text{m}$. In the case of a discrete Fourier transform of a structure with an element size of Δ , the corresponding size of the elements in the Fourier transform of the structure Δ_r , scale factor K_{sc} , focal length F of lens 3 and wavelength of laser light λ are related to each other by the ratio $K_{sc} = \Delta\Delta_r/F\lambda$. Let us consider the case of instrumental implementation of the polarimetric system (Fig. 1 using the LS2012 liquid crystal modulator, camera Kiralux CS505MUP1 and helium-neon laser ($\lambda = 0.63 \text{ мкм}$)). In polarization display of symbolic sequences for the «Wuhan», «Delta» and «Omicron» strains of the SARS-CoV-2 virus (the number of elements of the phase-modulating matrix 70×70) the working area of the phase-modulating screen in its central part occupies dimensions of $2.52 \text{ mm} \times 2.52 \text{ mm}$. Let us assume that the analyzed polarization-sensitive diffraction patterns contain 140×140 elements in the case of representation by discrete Fourier transform are displayed on the entire working area of the camera in size $8.45 \text{ mm} \times 8.45 \text{ mm}$. With a scaling factor of $K_{sc} = 0.1$, the focal length of the Fourier transform lens should be equal to $\approx 34 \text{ mm}$. Note that the area of the Fourier plane corresponding to one element of the discrete Fourier image is overlapped by approximately three hundred camera pixels. The aperture of lens 3 should exceed the dimensions of the working area of the phase-modulating screen; accordingly, one possible solution may be to use commercially available lenses supplied by Thorlabs (for example, LA1700 with a diameter of 6 mm and a focal length of 30 mm) or Russian analogues. A shorter focal length will lead to a slight increase in the scaling factor to 1.13, which will not have a significant effect on the functioning of the system. It should also be noted that when using single-mode helium-neon lasers as sources of reading radiation, traditionally used in interferometric, diffractometric and polarimetric measurements, any influence of the amplitude-phase noise of the reading beam on the polarization-dependent diffraction mapping of symbol sequences can be neglected.

Conclusion

Thus, the considered method of polarization encoding demonstrates high efficiency in terms of synthesizing two-dimensional binary objects that uniquely display the structure of DNA-associated symbolic sequences. It should be noted that the capabilities of this approach are not limited to the algorithm for transforming the structure of a symbolic sequence into the structure of a phase-modulating matrix (phase screen) considered in this paper. For example, the algorithm for coding submatrices considered in section 2 based on the content of basic nucleotides in the corresponding triplets (see expression (1)) under the condition of traversing the submatrices clockwise corresponds to the RY basis. Changing the associations of submatrix elements during polarization encoding in accordance with the rules: $b_{0,0} \rightarrow A$; $b_{1,0} \rightarrow G$; $b_{0,1} \rightarrow C$; $b_{1,1} \rightarrow T$ and $b_{0,0} \rightarrow A$; $b_{1,0} \rightarrow C$; $b_{0,1} \rightarrow G$; $b_{1,1} \rightarrow T$ (respectively, binding the MK and WS bases to coding) will expand the functionality of the approach under consideration. In particular, the analysis of the influence of the choice of basis during polarization coding is the object of further research in this direction. Another possible direction of research is the structural analysis of the synthesized binary mappings for discrimination threshold values close to ± 1 (similar to those shown in Fig. 4). For example, as elements characterizing such structures, one can consider the points of breaks and branches of the lines of limiting states of polarization, which appear in the binary mappings of Fig. 4.

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*. 2016;17(6):333–351. DOI: 10.1038/nrg.2016.49.
2. Neidle S, Sanderson M. *Principles of Nucleic Acid Structure*. Academic Press; 2021. 454 p.
3. Randić M, Vracko M, Lers N, Plavšić D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chemical Physics Letters*. 2003;368(1–2): 1–6. DOI: 10.1016/S0009-2614(02)01784-0.
4. Randić M, Vracko M, Nandy A, Basak SC. On 3-D graphical representation of DNA primary sequence and their numerical characterization. *Journal of Chemical Information and Computer Sciences*. 2000;40(5):1235–1244. DOI: 10.1021/ci000034q.
5. Xie G, Mo Z. Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. *Journal of Theoretical Biology*. 2011;269(1):123–130. DOI: 10.1016/j.jtbi.2010.10.018.
6. Jafarzadeh N, Iranmanesh A. A novel graphical and numerical representation for analyzing DNA sequences based on codons. *Match-Communications in Mathematical and Computer Chemistry*. 2012;68(2):611–620.
7. Jafarzadeh N, Iranmanesh A. C-curve: A novel 3D graphical representation of DNA sequence based on codons. *Mathematical Biosciences*. 2013;241(2):217–224. DOI: 10.1016/j.mbs.2012.11.009.
8. Hamori E, Ruskin J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *Journal of Biological Chemistry*. 1983;258(2): 1318–1327. DOI: 10.1016/S0021-9258(18)33196-X.
9. Zhang CT, Zhang R, Ou HY. The Z-curve databases: A graphic representation of genome sequence. *Bioinformatics*. 2003;19(5):593–599. DOI: 10.1093/bioinformatics/btg041.
10. Yu ZG, Wang B. A time series model of CDS sequences in complete genome. *Chaos Solitons Fractals*. 2001;12(3):519–526. DOI: 10.1016/S0960-0779(99)00208-8.

Zimnyakov D. A., Alonova M. V., Skripal An. V., Inkin M. G., Zaitsev S. S., Feodorova V. A.

11. Jeffrey HJ. Chaos game representation of gene structure. *Nucleic Acids Research*. 1990;18(8): 2163–2170. DOI: 10.1093/nar/18.8.2163.
12. Anitas EM. Small-angle scattering and multifractal analysis of DNA sequences. *International Journal of Molecular Sciences*. 2020;21(13):4651. DOI: 10.3390/ijms21134651.
13. Burma PK, Raj A, Deb JK, Brahmachari SK. Genome analysis: a new approach for visualization of sequence organization in genomes. *Journal of Biosciences*. 1992;17(4): 395–411. DOI: 10.1007/BF02720095.
14. Huynen MA, Konings DAM, Hogeweg P. Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure. *Journal of Molecular Evolution*. 1992; 34(4):280–291. DOI: 10.1007/BF00160235.
15. Hill KA, Schisler NJ, Singh SM. Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. *Journal of Molecular Evolution*. 1992;35(3):261–269. DOI: 10.1007/BF00178602.
16. Almeida JS, Carrico JA, Marezek A, Noble PA, Fletcher M. Analysis of genomic sequences by chaos game representation. *Bioinformatics*. 2001;17(5):429–437. DOI: 10.1093/bioinformatics/17.5.429.
17. Zimnyakov DA, Alonova MV, Skripal AnV, Zaitsev SS, Feodorova VA. Polarization analysis of gene sequence structures: Mapping of extreme local polarization states. *Journal of Biomedical Photonics & Engineering*. 2022;8(4):040302. DOI: 10.18287/JBPE22.08.040302.
18. Zimnyakov DA, Alonova MV, Skripal AnV, Dobdin SY, Feodorova VA. Quantification of the diversity in gene structures using the principles of polarization mapping. *Current Issues in Molecular Biology*. 2023;45(2):1720–1740. DOI: 10.3390/cimb45020111.
19. Ulyanov SS, Ulianova OV, Zaytsev SS, Saltykov YV, Feodorova VA. 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*. 2018;15: 045601. DOI: 10.1088/1612-202X/aaa11c.
20. Rak A, Isakova-Sivak I, Rudenko L. Overview of Nucleocapsid-Targeting Vaccines against COVID-19. *Vaccines*. 2023;11(12):1810. DOI: 10.3390/vaccines11121810.
21. Telenti A, Hodcroft EB, Robertson DL. The Evolution and Biology of SARS-CoV-2 Variants. *Cold Spring Harbor Perspectives in Medicine*. 2022;12:a041390. DOI: 10.1101/cshperspect.a041390.
22. Bergmann CC, Silverman RH. COVID-19: coronavirus replication, pathogenesis, and therapeutic strategies. *Cleveland Clinic Journal of Medicine*. 2020;87:321–327 DOI: 10.3949/ccjm.87a.20047.
23. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, Li F. Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences*. 2020;117:11727–11734. DOI: 10.1073/pnas.2003138117.
24. Grobbelaar LM, Venter C, Vlok M, Ngoepe M, Laubscher GJ, Lourens PJ, Steenkamp J, Kell DB, Pretorius E. SARS-CoV-2 spike protein S1 induces fibrin (ogen) resistant to fibrinolysis: implications for microclot formation in COVID-19. *Bioscience Reports*. 2021; 41(8):BSR20210611. DOI: 10.1042/BSR20210611.
25. Singh D, Yi SV. On the origin and evolution of SARS-CoV-2. *Experimental & Molecular Medicine*. 2021;53:537–547. DOI: 10.1038/s12276-021-00604-z.
26. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270–273. DOI: 10.1038/s41586-020-2012-7.

27. Chakraborty C, Bhattacharya M, Chopra H, Bhattacharya P, Islam MA, Dhama K. Recently emerged omicron subvariant BF.7 and its R346T mutation in the RBD region reveal increased transmissibility and higher resistance to neutralization antibodies: need to understand more under the current scenario of rising cases in China and fears of driving a new wave of the COVID-19 pandemic. *International Journal of Surgery*. 2023;109(4):1037–1040. DOI: 10.1097/JS9.000000000000219.
28. GISAID [Electronic resource]: Official hCoV-19 Reference Sequence. Acc. ID: EPI_ISL_402124. Available online: <https://gisaid.org/wiv04/>.
29. GISAID [Electronic resource]: Official hCoV-19 Reference Sequence. Acc. ID: EPI_ISL_2552101. Available online: <https://gisaid.org/wiv04/>.
30. GISAID [Electronic resource]: Official hCoV-19 Reference Sequence. Acc. ID: EPI_ISL_9991311. Available online: <https://gisaid.org/wiv04/>.
31. Goodman JW. *Introduction to Fourier Optics*, 4th ed. New York: Macmillan Learning; 2017. 491 p.
32. Bracewell R. *The Fourier Transform and Its Applications*. New York: McGraw Hill; 1986. 474 p.
33. Chipman R, Lam WST, Young G. *Polarized Light and Optical Systems (Optical Sciences and Applications of Light)*. Boca-Raton: CRC Press; 2018. 1036 p.
34. Anitas EM. Fractal analysis of DNA sequences using frequency chaos game representation and small-angle scattering. *International Journal of Molecular Sciences*. 2022;23(3):1847. DOI: 10.3390/ijms23031847.