The inherited information geometric analysis and the genetic sub-alphabets invariant system

IOP Conf. Series: Materials Science and Engineering **489** (2019) 012049 doi:10.1088/1757-899X/489/1/012049

The inherited information geometric analysis and the genetic sub-alphabets invariant system

I Stepanyan[1,2,3] and V Svirin[1,3]

[1]Mechanical Engineering Research Institute of RAS, 4, M. Kharitonyevskiy Lane, Moscow, 101990, Russia
[2]Izmerov Research Institute of Occupational Health of the RAS, 31, Prospect Budennogo, Moscow, 105275, Russia
[3]Moscow State Conservatory, 13/6 Bolshaya Nikitskaya Street, Moscow, 125009, Russia

neurocomp.pro@mgail.com

Abstract. Examples of invariants in the system of molecular genetic coding are given. Models and methods of algebraic biology based on matrix genetics (Petukhov, 2008) with the transition from matrix algebra to discrete geometry and visualization are presented. Algorithms based on genetic sub-alphabets allow displaying geometrically the biochemical composition of polynucleotide chains in spaces of different dimensions using Walsh orthogonal functions of physicochemical parameters. The visualization method makes it possible to substantiate the relationship between the parameters of DNA and RNA molecules with fractal clusters and geometric mosaics, reveals the orderliness and symmetry of polynucleotides and the noise immunity of their visual representations in the orthogonal coordinate system. A scale-parametric model of nucleic acid visualization is proposed. The developed methods can serve to simplify the perception by researchers of polynucleotide chains by visualizing them in spaces of various dimensions. Examples of visualization of the nucleotide composition of the various species of living organisms genomes are given

1. Introduction

Many papers are devoted to symmetries in genetic coding. In recent years, interest in symmetries in genetic coding has increased due to the intensive study of CRISPR-C as systems. Symmetries are closely related to invariance.

In [1] it is applied to study molecular evolution in vivo and in vitro. A representation of the genetic code as a six–dimensional Boolean hypercube is described. This structure is the result of the interaction energies hierarchical order of the bases in codon– anticodon recognition. Based on dynamical systems theory convergence properties of genetic algorithms are investigated at [2]. A classification procedure is proposed for genetic algorithms based on a conjecture: the entropy and the fractal dimension of trajectories produced by them are quantities that characterize the classes of the algorithms. The role of these quantities as invariants of the algorithm classes is presented.

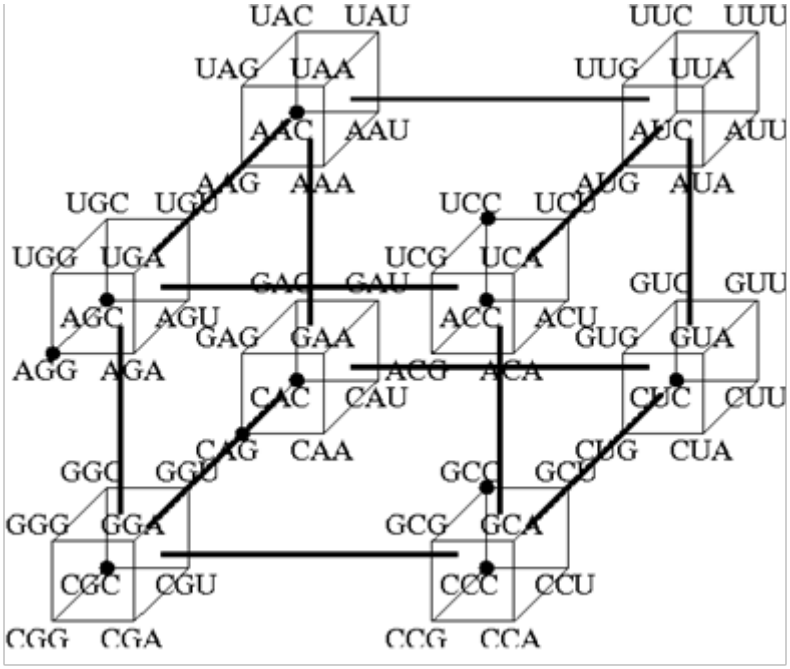


Figure 1. The hypercube representation of the genetic code from [3]

The aim of [3] was to present and analyze the probabilistic models of mathematical phylogenetics. In other cases new objects appear, such as the remarkable quintic `squangle' invariants for quartet tree discrimination and DNA data, with their own unique interpretation in the phylogenetic modelling context. The hypercube representation of the genetic code from [3] is shown at figure 1.

This paper considers geometrization as the methodological principle for the development of bioinformatics. The emergence of sound methods for comparing geometric models of genotypes with certain phenotypic traits contributes to the expansion of the circle of researchers in the field of molecular genetics.

2. Materials and methods. Invariant coding of physicochemical parameters of nucleotides

In [5], it was demonstrated that each nitrogenous base of the genetic code has three variants of its binary representation. These variants of representations, called by S.V. Petuhov the binary sub-alphabets are distinguished according to the types of binary oppositional properties in the set of nitrogenous bases:

- G = C "3 hydrogen bonds" / A = T "2 hydrogen bonds";
- C = T "pyrimidines" / A = G "purines";
- A = C "amino" / G = T "keto" [24];
- A = T = G = C (presence of phosphate residue).

Taking into consideration the additional fourth feature, which is not oppositional, the system of genetic sub-alphabets can be represented in the form of the Hadamard matrix, demonstrated in figure 2.

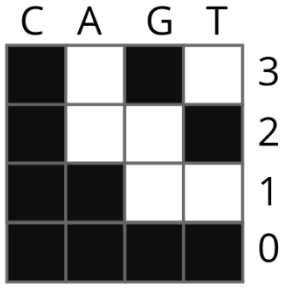


Figure 2. The variant of the Hadamard matrix displaying the encoding of nucleotide sub-alphabets. Shaded cells +1, white cells -1 (or vice versa, depending on the encoding method). Numbers of sub-alphabets are denoted as 1, 2, 3 and 0.

This matrix is symmetric, since nucleotides can be replaced by corresponding sub-alphabets without changing the structure of the matrix [6].

Each row and column of the resulting symmetric Hadamard matrix is a Walsh function [7]. The Walsh functions are a complete set of orthogonal functions that can be used to represent any discrete function by analogy with the use of trigonometric functions in Fourier analysis [10]. They are widely used in digital computing, in coding noise-resistant communications, as well as in quantum mechanics and quantum computer science. Information on Hadamard symmetries and matrices in genetic coding is described in detail in [5,13,18].

Physical holography, which possesses the highest properties of noise-immunity, is based on a record of standing waves from two coherent physical waves of the object beam and of the reference beam. But physical waves can be modelled digitally.

Correspondingly noise-immunity and other properties of optical and acoustical holography can be digitally modelled, in particular, with using Walsh functions and

logic operations concerning dyadic groups of binary numbers because Walsh transforms are Fourier transforms on the dyadic groups. This can be made on the base of discrete electrical or other signals without any application of physical waves. In this digital Walsh-holography, objects, whose digital holograms should be made, are represented in forms of 2 n -dimensional vectors.

3. Algorithm of scale-parametric modelling of nucleic acids

1) a sequence of single-letter symbols encoding nitrogen bases in a nucleic acid is divided into fragments of equal length N, where N is a free parameter of the algorithm. The resulting fragments of equal length will be called N-measures or N-lashes [8];

2) taking into account the system of genetic sub-alphabets, the sequence of nitrogenous bases can be represented in the form of three binary sequences consisting of zeros and ones. The choice of coding method (which is considered to be zero or one) affects the turns and other transformations of the final visualization;

3) the resulting binary record of these fragments is a representation in the form of three sequences of decimal or other uniquely identifying values. Converting binary N-plets to decimal numbers allows their use in one or another coordinate system. These values define the coordinates of points in the parameter space, taking into account the chosen coding method (hereinafter, parametric space or visualization space).

Steps 1 and 2 can be rearranged (first parameterization, then cutting), which can significantly affect the computational load when calculating long sequences.

The given parametric space is finite, discrete, and three-dimensional in terms of the number of binary opposition signs. Its combinatorial properties allow you to display any polynucleotides for an arbitrary finite N. Ordered numerical values on the coordinate axes uniquely reflect the physicochemical characteristics of N-mers, since they are precisely determined by the properties of binary opposition sub-alphabets.

4. Results and discussion.

The set of three binary opposition signs can be compared with the {X, Y, Z} axes of the Cartesian coordinate system. The N coefficient performs the functions of the resolution of geometric visualization: large N gives a small number of points, small N gives a small coordinate grid. This circumstance allows us to speak of multi-scale analysis in multidimensional parametric spaces.

In accordance with the methods of Harmut's theory of sequential analysis [9], additional visualizations were constructed by the number of elements (zeros or ones) that were encountered in binary representations of N-plets in sequences of nitrogenous bases. Due to the fact that such a visualization method is based on the total number of certain parameters in an N-plet, the corresponding visualization spaces will be called integrals.

Visualization algorithms can be applied to any nucleic acids, including RNA and chromosomes of living organisms. In this work, genomes from the NCBI bioinformation database [17] were used for visualization, as well as materials kindly provided by the laboratory of prof. N.S. Zenkina Center of Bacterial Cell Biology of the Newcastle University (UK).

On the basis of the developed computer program, it was found that the chromosomes of various types of organisms have individual structural features in various spaces. Genomic imaging of various organisms may have a pronounced fractal character, which is identical for all chromosomes of the organism under study, as well as for arbitrary fragments of these chromosomes.

To visualize nucleic acids in the three-dimensional parameter space, we use the orthogonal basis {X, Y, Z}, chosen as the three-dimensional Cartesian coordinate system. It gives the visualization, the example of which is shown in Fig. 3. Each point represents the characteristic of the binary opposition of the corresponding fragment of the sequence, summarized in accordance with the algorithm, which allows one to display the nucleotide composition of the molecule.

Figure 3. Illustration of a three-dimensional representation of the nucleotide composition on the example of the chromosome of a living organism in various projections. The X, Y, and Z axes correspond to the ordered decimal representations of the binary coding of N-plets based on three binary opposition sub-alphabets. Each point of the figure corresponds to N-measures, the coordinate of which is given by its proton-numerical features.

The resulting geometric shape, which is the three-dimensional analogue of the Sierpinski triangle, is typical for the three-dimensional visualization of any long nucleotide sequences. The shape of the resulting figure is determined by the properties of binary sub-alphabets. The coordinates of each point in the three-dimensional visualization space are completely defined by any two coordinates, since the

third coordinate is calculated by modulo two remaining coordinates. This algebraic feature is associated with the redundancy of binary sub-alphabets used by nature for storing and transmitting genetic information in the chains of generations. Fig.4. demonstrates an integral three-dimensional representation of the nucleotide composition.

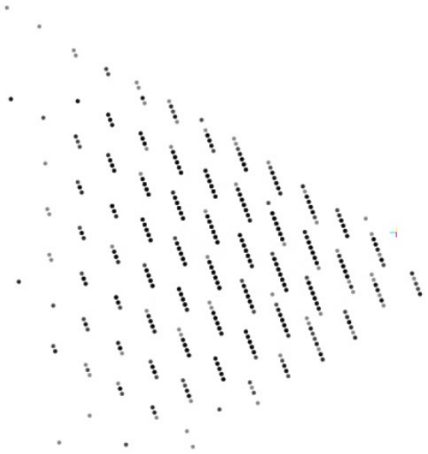
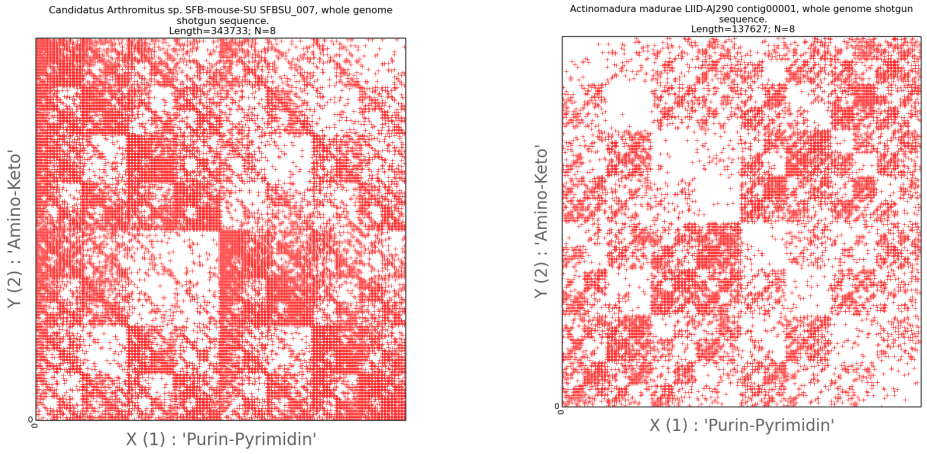


Figure 4. Integral-three-dimensional representation of the nucleotide composition. The X, Y, and Z axes correspond to the number of units in the decimal binary encoding of each N-plet using three binary opposition sub-alphabets.

An example of the nucleotide composition integral three-dimensional representation of the living organism chromosome is presented in Figure 3. This is a discrete pyramid, each point of which corresponds to a set of N-measures of nucleic acid, combined by the number of units in binary coding for each of the signs.

As can be seen, in connection with the geometric properties of a given space, three-dimensional visualizations are not very convenient for perceiving and analyzing the characteristics of a particular nucleic acid. However, the available three two-dimensional projections of these visualizations are well suited to display the specificity of the structure of long polynucleotide chains. In the bases {X, Y}, (X, Z} and {Y, Z}, chosen as Cartesian coordinate systems, three-dimensional visualization gives three different two-dimensional projections based on the corresponding sub-alphabets of physicochemical signs of nucleotides. Based on the property of the genetic coding, according to which a triple of binary opposition sub-alphabets is connected with each other by addition modulo two, to determine an arbitrary nucleic acid, any pair of its binary representations is sufficient. Therefore, for two-dimensional visualization of a nucleotide composition it is sufficient to have any axes pair.

The question of determining the most informative pair of coordinate axes and, accordingly, the parameters taken into account may depend on the type of organism being analyzed and requires additional research. As a result of the analysis of some genomes, it was found that the three variants of two-dimensional visualization, mosaics based on information about the external structure of the molecule, i.e. built on the elements of the structures encoding the signs of amino / keto and purine / pyrimidine. Such mosaics had a detailed fractal pattern consisting of squares or rectangles. In some genomes, mosaics based on types of hydrogen bonds appeared to be the most symmetrical. They are usually characterized by triangular elements or pattern diagonals and are found in plant DNA. At the same time, diagonal or other elements of the pattern can be directed in various directions in different organisms while maintaining the structure of the pattern (this feature can also be modelled by reading a complementary strand of DNA).



a) b)

Figure 5. a) Illustration of a two-dimensional representation of the nucleotide composition of genomes: Candidatus Arthromitus bacterium; b) the genus Actinomadura (one of the four genes). The abscissa and ordinate axis correspond to the decimal representations of the binary encoding of each 8-plet.

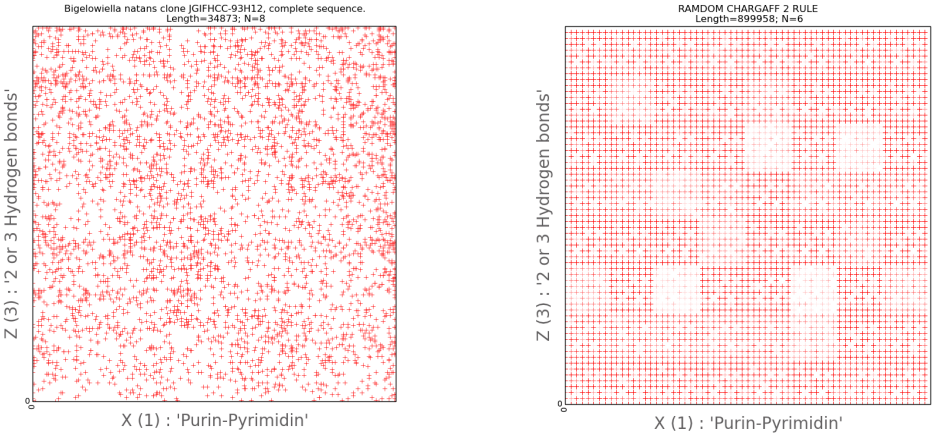
General scientific studying methods of nucleic acids usually focus their attention on those fragments that are present in them. The proposed algorithms allow visualization of the phenomenology and features of the deficiency and presence of various types of N-mers.

Two-dimensional models, which are obtained using the method described, sometimes resemble the fractal patterns of nucleic acids, which in [11] were obtained using the CGR method. However, these methods are quite different in their algorithmic nature. In particular, the CGR method is missing information about binary sub-alphabets.

In parametric visualization and comparison of the human and monkey genomes, as well as the genomes of other species of living organisms, two-dimensional mosaics of arbitrary fragments of the genomes with the whole genome were observed. This dismantles the fractal properties of the structure of nucleic acids and indicates the existence of both interspecific and intraspecific variants of nucleotide composition.

The preliminary results of applying two-dimensional visualization algorithms allow us to conclude that the final mosaics are highly stable when the original sequence is noisy, including when shifting the reading frame of the sequences, when arbitrary fragments of the sequence are removed (distortion), when reversing the entire analyzed circuit or its fragments permutations of N-mers and nucleotides (in some cases up to a complete permutation of all nucleotides in the sequence).

In particular, we observed the stability of mosaic patterns in cases of every second nucleotide deletion, every third nucleotide, etc. At the same time, the visualization of nucleic acids in two-dimensional spaces in some cases is characterized by pronounced symmetries and stability not only to different noise in the original data, but also to different values of the scale parameter N within a certain range.



a) b)

Figure 6. .a) Illustration of a two-dimensional representation of the nucleotide composition of the organism bigelowiella natans (Eukaryota). Natans are a key resource for the supergroup of mostly unicellular eukaryotes. The abscissa and ordinate axis correspond to the decimal representations of the binary encoding of each 8-plet; b) illustration of a two-dimensional representation of the nucleotide composition of an algorithmically generated chromosome in compliance with the second Chargaff rule. The abscissa and ordinate axis correspond to the decimal representations of the binary encoding of each 6-plet.

For further studies, sequences of nitrogenous bases 100,000 nucleotides in length were randomly generated using a computer program and divided into N-laps on 8, 16, and 28. Randomly generated sequences in visualization give a pattern, all points of which are scattered randomly. Such visual representations are irregular, chaotic in the complete absence of any mosaics and patterns in all sub-alphabets, which significantly distinguishes them from real genomes.

Visual representations of the genomes of various penicillin kinds were constructed. The results suggest that the genomes of this group of antibiotics usually generate high-density mosaics that resemble mosaics of random sequences, which indicates a high diversity of their nucleotide composition. Perhaps the medical value of penicillin kinds is associated with this their feature.

Pseudorandom nucleic acids are modelled following the rules of Chargaff, valid for each of the two DNA strands [4,14]. A special type of patterns for these generated sequences was visualized at N = 6 in Fig. 6b. Some fractals of real genomes were also modelled by generating pseudo-random nucleic acids with selection of probability coefficients.

Two-dimensional visualization algorithms appear to be useful for studying hidden patterns in chromosomes, as well as for the classification and comparative analysis of various genomes with possible applications in biotechnology and medicine. Examples of genetic mosaics built in a nonpositional number system are given in [12].

As a result of the analysis, we can conclude that the two-dimensional visualization of the nucleotide composition of both algorithmically generated and real genomes makes it possible to display the implementation options for the quantitative rules of Chargaff [4] using elements of combinatorial final geometry [25]. This fact can help in the study of internal symmetries and other characteristics of nucleic acids, transferring their analysis from the field of statistics to the field of geometry to study the complex relationships between biological objects.

As noted, binary sub-alphabets are interconnected by modulo-two addition and de-fine a parametric space with properties at which the coordinates of all points are connected or “glued” together. In this

regard, for a one-dimensional analysis, it makes sense to consider each dimension separately and independently

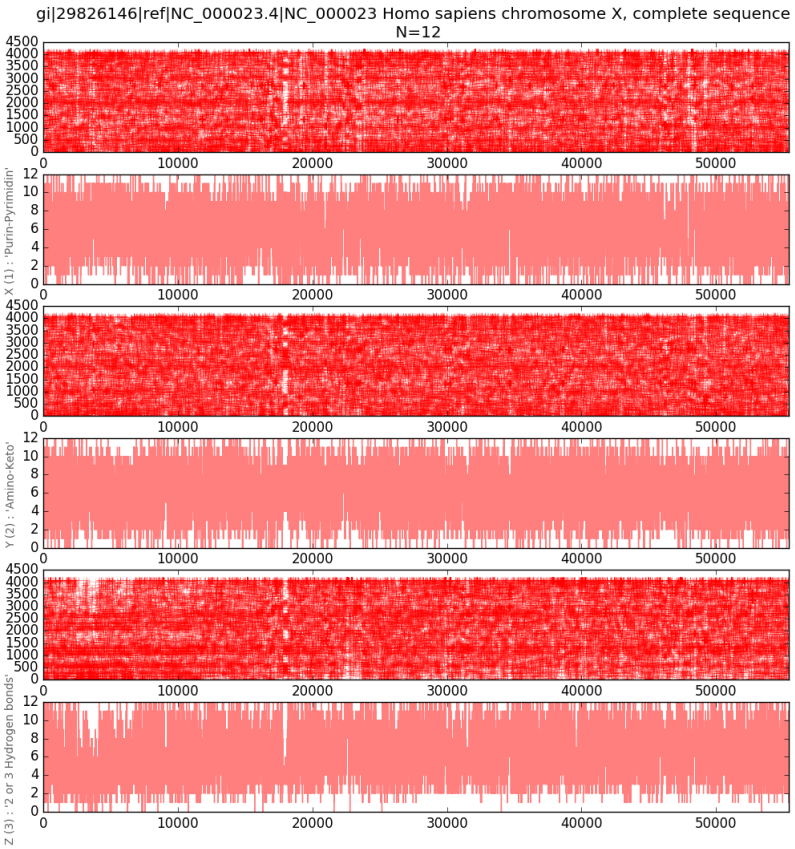


Figure 7. One-dimensional visualization of the three-channel representation of the nucleotide composition of a fragment of the 22nd chromosome of Homo Sapiens. Each of the three rows corresponds to a binary opposition sub-alphabet. In each channel, the abscissa axis encodes the ordinal number of the N-plet, the ordinate axis encodes the ordered ascending decimal values of the binary representation of the N-plets.

Taking into consideration the presence of three binary opposition signs, there are three one-dimensional visualization spaces. At the same time, the mathematical relationship between their coordinate axes is preserved. Thus, the use of one-dimensional coordinate axes {X}, {Y} and {Z} gives a triplet of various mappings using the corresponding sub-alphabets.

In Fig. 7 in the 22nd chromosome of Homo Sapiens, sub-alphabets clearly show areas with different nucleotide composition. These specific regions can be visualized in two-dimensional parametric spaces for further comparison and analysis.

The obtained graphs allow us to estimate changes in the nucleotide composition when reading a fragment of a molecule from beginning to the end. The depth of the recorded changes is determined by the scaling parameter N of the algorithm.

5. Conclusions

Parametric visualization of both fragments and whole DNA and RNA molecules made it possible to substantiate their connection with fractal and other mosaics. The correlation of molecular genetic systems with binary code and Hadamard matrices are demonstrated. The stated principles of visualization allow visual assessment of the correlation types between the N-measures present and absent in the genomes of various organisms and viruses, which in some cases are characterized by fractal-cluster organization in spaces of binary-orthogonal functions of their physical and chemical parameters. The obtained geometric patterns are associated with the quantitative phenomenological rules of Chargaff [4,14,22].

The scaling parameter N allows exploring the genome at a variety of different levels of detail to search for hidden symmetries and patterns when visualizing in spaces of different dimensions.

The proposed methods can serve to simplify the perception by researchers of long polynucleotide chains, and also serve as an additional criterion for the classification and identification of interspecific interactions. Modern ontologies and thesauri for organizing and storing biological and molecular genetic data can be supplemented with visualization options for educational purposes, as well as for the presentation and retrieval of biological information.

The results obtained are related to the theory of number systems, self-organizing systems and correlate with the theory of dynamic systems. Parametric modelling and visualization of nucleotide composition can contribute to an in-depth understanding of genetic phenomena, not only by simplifying the mechanisms of perception, but also by the possibility of using adaptive neural network technologies, since the structure of chromosomes of living organisms, represented in binary form, fits well with the format of binary artificial neural networks [16] . In addition, DNA geometrization is associated with logical holography.

As known, holographic methods in engineering allow quick detecting of individual elements in a huge image. The theory of genetic logical holography allows assuming that one of secrets of noise-immunity of genetic informatics is based on the similar possibilities of genetic logical holography. In an appropriate modelling approach, if one of DNA molecules mutates, the genetic logical holography – by analogy with classic physical holography - allows quick detecting of the mutated DNA in the whole logical hologram of a set of DNA molecules. In the result, the genetic information of this individual DNA molecule could be found to be incorrect for further using in organism automatically.

Conducted studies on the comparative analysis of nucleotide sequences visualization of various types of living organisms (protozoa, plants, fungi, animals, viruses) confirm that the nucleotide composition may be identical in organisms that are not related in the phylogenetic tree and different in related organisms [15].

It opens up the possibility of modelling pseudorandom nucleotide sequences with the observance of the phenomenological rules of Chargaff for their visualization and further research in parametric spaces of various dimensions. The proposed tool for visualization of nucleic acids can help advance the understanding of the immune system functioning principles in recognizing the nucleotide composition of viruses, parasite DNA, as well as in food chains and ecosystems.

Part of the calculations was performed on the MVS-10P supercomputer (MSC RAS).

References

[1] Jiménez-Montaño M A, Mora-Basánez C R and Pöschel T 1996 *Biosystems* **39(2)** 117-25.

[2] Kotowski S., Kosiński W., Michalewicz Z., Nowicki J., Przepiórkiewicz B. 2008 Fractal Dimension of Trajectory as Invariant of Genetic Algorithms. In: Rutkowski L., Tadeusiewicz R., Zadeh L.A., Zurada J.M. (eds) Artificial Intelligence and Soft Computing – ICAISC 2008. ICAISC 2008. Lecture Notes in Computer Science, vol 5097. Springer, Berlin, Heidelberg

[3] Jarvis P D , Sumner J G 2018 Systematics and symmetry in molecular phylogenetic modelling: perspectives from physics *Preprint* 1809.03078

[4] Chargaff E, Lipshitz R, Green C 1952 *J Biol Chem.***195 (1)** 155–60

[5] Petoukhov S.V. and He M 2010 *Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications* (Hershey, USA: IGI Global)
[6] Balonin N A, Balonin Y N, Djokovic D Z, Karbovskiy D A, Sergeev M B 2017 Construc-tion of symmetric Hadamard matrices *Preprint* 1708.05098
[7] Georgiou S, Koukouvinos C, Seberry J 2003 *Further computational and con-structive design theory* (Boston: Kluwer.) pp 133–205. ISBN 1-4020-7599-5
[8] Stepanian I V, Petoukhov SV 2013 The matrix method of representation, analysis and classification of long genetic sequences *Preprint* 1310.8469
[9] Harmuth H 2016 *Applying of Methods of Theory of Information in phisics* (Moscow.: Mir)
[10] 10.Ferleger, Sergei V. (March 1998). RUC-Systems In Non-Commutative Symmetric Spaces (Technical report). MP-ARC-98-188.
[11] Jeffrey H J 1990. *Nucleic Acids Re-search* **18(8)** pp 2163-70
[12] Feldman D P 2012 *Chaos and Fractals: An Elementary Introduction* (Oxford: OUP) pp 178–80. ISBN 9780199566440
[13] 13.Darvas G, Koblyakov A A, Petoukhov S V, Stepanyan I V. 2012 *Symmetry: Culture and Science* **23(3-4)** 343-75
[14] Rudner R, Karkas J D and Chargaff E 1968 Separation of B. subtilis DNA into complementary strands. 3. Direct analysis *PNAS* **60(3)** 921-22; https://doi.org/10.1073/pnas.60.3.921
[15] Townsend JP, Su Z, Tekle Y 2012 *Systematic Biology* **61(5)** 835–49. https://doi.org/10.1093/sysbio/sys036
[16] Stepanyan I V, Ziep N N 2018 *Neurocomputers: development, application* **5** pp 4-11
[17] ftp://ftp.ncbi.nlm.nih.gov/
[18] Petukhov S 2008 *Matrix Genetics, Algebra of Genetic Code, Noise Immunity* (Moscow: RHD)
[19] Takuyuki I, Wakana T, Masafumi M, Masaki E, Hiro-Yuki H 2015 *Genes & Genetic Systems* **90(4)** 231-35. https://doi.org/10.1266/ggs.15-00030
[20] Crick F H C, Wang J C and Bauer W R 1979 *jmb* **129(3)** 449-61. https://doi.org/10.1016/0022-2836(79)90506-0
[21] Wilkins M H F, Stokes A R and Wilson H R 1953 *Nature* **171** 738–40. https://doi.org/10.1038/171738a0
[22] Chargaff E, Lipshitz R and Green C 1952 *J Biol Chem.***195 (1)** 155–60
[23] Sharov A. 1992 *Biosemiotics: The Semiotic Web 1991.* Ed T.A.Sebeok and J.Umiker-Sebeok. (New York :Mouton de Gruyter) pp 345-373
[24] 24.Waterman M S 1995 *Introduction to Computational Biology. Map, Sequences and Genomes* (London: Chapman & Hall)
[25] Batten L M 1997 *Combinatorics of Finite Geometries* (Cambridge: CUP) ISBN 0521590140