Scaling and Visualization of Nucleotide Sequences

Ivan V. Stepanyan [1,*] and *Abbakar M.* Khussein [1]

[1] Mechanical Engineering Research Institute of the Russian Academy of Sciences, RU-101990, Moscow, Russia

Abstract. Algorithms for scaling and visualization of nucleotide sequences developed in this study allow identifying relationships between the biochemical parameters of DNA and RNA molecules with scale invariance, fractal clusters, nonlinear ordering and symmetry and noise immunity of visual representations in orthogonal coordinate systems. The algorithms are capable of displaying structures of the nucleotide sequences of living organisms by visualizing them in spaces of various dimensions and scales. Approximately one hundred genes (protozoa, plants, fungi, animals, viruses) were analysed and examples of visualization of the nucleotide composition of genomes of various species have been presented. The developed method contributes to an in-depth understanding of the principles of genetic coding and simplifying the perception of genetic information due to the algorithmic interpretation of the basic properties of polynucleotide fragments with visualization of the final geometric structure of the genetic code.

1 Introduction

Mathematical biology is based on computational methods and algorithms that allow to acquire scientific knowledge, model biological processes and phenomena. However, there is the problem of perceiving complex biological information, including phenomena that occur inside the cell. This problem relates to the psychophysiology of perception of any multidimensional information. It is rather difficult to imagine all processes occurring in a living cell, despite the strong theoretical foundation and the presence of a well-developed mathematical apparatus. The same applies to the task of analyzing the variability of the physicochemical parameters of long nucleotide sequences that occur in the form of DNA and RNA molecules. To analyze complex multi-parameter phenomena, which include the phenomenon of genetic coding, methods of lowering dimensions are used, including machine learning (neural network clustering, classification, deep learning, etc.).

Widely used scientific visualization using computer animation and cognitive graphics. Computer graphics methods allow to build some simplification of the studied objects in the form of computer models. A significant contribution to the development of cognitive graphics was made by A.A. Zenkin [1], who studied algorithms for the two-dimensional representation of one-dimensional processes.

For the analysis of genetic nucleotide sequences (i.e sequenced data stored as files in computers), software tools and algorithms based on statistical analysis are widely used. However, such tools do not have sufficient visibility and visual capabilities, the results of their implementation are cumbersome and not always expressive.

In this paper, we propose a new concept for studying the properties of long nucleotide sequences (and their physicochemical parameters), based on the principles of finite geometry [2]. The algorithms of cognitive computer graphics, the examples of which are presented in the paper, make it possible to simplify the perception of genetic information and biologically interpret those properties of long sequences of the type AGGCT ... in the DNA (RNA) of living organisms that cannot be seen by simply reading them or using traditional computer analysis tools. This concept is proposed as the basis of a new methodology for the development of research bioinformatics software.

When developing software algorithms, the property of the human visual analyzer was taken into account, which consists in the fact that two-dimensional objects are perceived most effectively by it: nature envisaged the use of the rich mathematical properties of two-dimensionality, which is the basis of the surface of the cerebral cortex (neocortex),that is responsible for analyzing functions and higher nervous activity. The neocortex is compactly folded in the skull, forming convolutions, the number of which (and, therefore, the area of the used two-dimensional surface) determines cognitive abilities.

The direction of bioinformatics described in this work contributes to the simplification of perception for subsequent analysis and generalization of a whole class of biophysical phenomena. In addition, the developed algorithms are promising for the analysis of biological data based on neural network technology.

2. Materials and methods

Nucleic acids of DNA and RNA are sequences of complementary nucleotide pairs that perform the

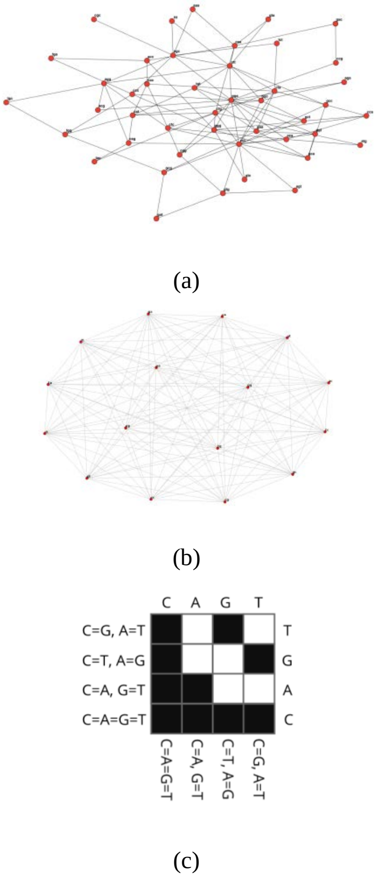
[*] e-mail: neurocomp.pro@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

functions of storing and transmitting hereditary genetic information in living organisms [3-5]. Such sequences are analyzed, as a rule, by statistical methods. They have a one-dimensional linear character and are stored on computers in the form of strings consisting of four letters of the alphabet encoding the nucleotides: cytosine (C), thymine (T), uracil (U), adenine (A), guanine (G). Thymine can be replaced by uracil upon transition from DNA to RNA.

2.1 Semantic approach to the visualization of nucleotide sequences on the example of the insulin gene

Our semantic approach to visualization is a methodology consisting in constructing such a graph whose vertices are fragments of a genetic sequence of equal length N (words or semantic units), and arcs are links that represent the neighborhood of these fragments. Arcs can be directional, showing the neighborhood on the right or left. The length of fragments is a free parameter of the algorithm, which makes it possible to construct graphs for the same nucleotide sequence at different scales. Examples of constructing the semantic network of the insulin gene when breaking the genetic sequence of nucleotides into fragments of different lengths are shown in Fig. 1.



(a)

(b)

	C	A	G	T	
C=G, A=T	■	□	■	□	T
C=T, A=G	■	□	□	■	G
C=A, G=T	■	■	□	□	A
C=A=G=T	■	■	■	■	C
	C=A=G=T	C=A, G=T	C=T, A=G	C=G, A=T	

(c)

Fig. 1. Semantic network of the insulin gene when breaking the coding nucleotide sequence into blocks of 3 (a) and 2 (b) nucleotides. Hadamard matrix (c) representing the encoding of nucleotide sub-alphabets. Shaded cells +1, white cells -1 (or vice versa, depending on the encoding method).

As can be seen from the graphs depicting the semantic relationships of the same genome, it is possible to evaluate the scale variability of its structure depending on the scale parameter, since the structure of the graph varies greatly at different scales. It can be seen that with increasing word length the graph connectivity decreases and start and stop codons can be distinguished. However, the structure of the graph is not always informative. For large genomes that occur in nature, the structure of the semantic graph will resemble a tangled ball, which will only complicate the perception of information. You can analyze the connectedness matrices corresponding to the graphs, but the absence of pronounced symmetries reflecting the nucleotide composition does not provide a fundamentally new quality. Therefore, for a more informative assessment, we decided to use the considerations developed and presented by us in [6]. These considerations are based on the ideas of semantic visualization (the presence of a free scaling parameter) and taking into account the physicochemical parameters of nucleotides.

Part of the calculations was performed on the MVS-10P supercomputer (MSC RAS) using genome database of The National Center for Biotechnology Information [7].

2.2 Encoding the physicochemical parameters of nucleotides

In [4], it was shown that each nitrogenous base of the genetic code has three variants of its binary representation. These presentation options, called S.V. Petukhov binary sub-alphabets differ in accordance with 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”; A = T = G = C (presence of a phosphate residue). Given the presence of the fourth trait, which is not oppositional, the system of genetic sub-alphabet can be presented as Hadamard matrix (fig. 1c). This matrix is symmetric, since nucleotides can be replaced by the corresponding sub-alphabets without changing the structure of the matrix. Each row and each column of this Hadamard matrix is a Walsh function [8]. Information on Hadamard symmetries and matrices in genetic coding is studied in detail in the works of S.V. Petukhov in the framework of matrix genetics [5,6,10].

2.3 Algorithm for scale-parametric modeling of physicochemical parameters of long nucleotide sequences

The presented algorithm is the author's one, was firstly presented at [6] and underlies the reconstruction of a scale-parametric model of the genome for visualization in coordinate spaces of various dimensions and topology.

1) a sequence of characters from the set {A, G, C, T} or {A, G, C, U} encoding nitrogen bases is divided into fragments of equal length N, where N is a free parameter

of the algorithm — the word length. The resulting fragments of equal length will be called N-plets [6];

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

3) the resulting binary record of fragments is their representation in the form of three sequences of decimal or other uniquely identifying values.

Converting binary N-plets to decimal numbers allows to display them in the selected coordinate system. The obtained values specify the coordinates of points in the parameter space of physicochemical parameters (hereinafter, in the visualization space or parametric space).

As a result of the application of the algorithm, a model space is defined that is parametric, finite, discrete, and three-dimensional in terms of the number of binary opposition features. The combinatorial properties of this space allow us to display any polynucleotides for an arbitrary finite N. The ordered numerical values on the coordinate axes reflect the physicochemical characteristics of N-measures, since they are uniquely determined by the properties of binary opposition sub-alphabets.

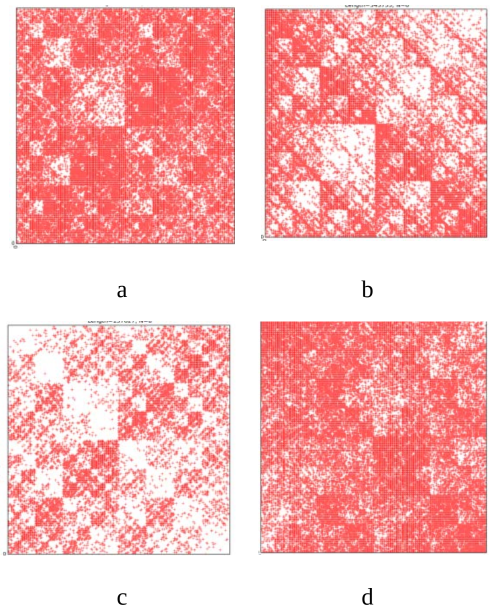
Note that general scientific methods for studying nucleic acids usually focus on those fragments that are present in them. The proposed algorithms allow us to visualize the phenomenology and features of the deficit and the presence of various types of N-measures.

The scale-parametric modeling algorithm was used to analyze various RNA and DNA molecules. During the research, about a hundred genomes (protozoa, plants, fungi, animals, viruses) were visualized. Part of the calculations was performed on the MVS-10P supercomputer (MSC RAS).

3. Results and discussion. Examples of visualization and evaluation of variability of physical and chemical parameters of nucleotide sequences

A set of three binary opposition signs can be compared with the {X, Y, Z} axes of the Cartesian coordinate system. The three-dimensional representations of nucleic acids thus obtained are not convenient for the perception and analysis of their features. However, two-dimensional projections of such three-dimensional representations are suitable for displaying the specific structure of long molecules. In the bases {X, Y}, (X, Z} and {Y, Z}, selected as Cartesian coordinate systems, we obtain three different two-dimensional projections based on pairs of the corresponding sub-alphabets of physicochemical nucleotide parameters.

Based on the property of genetic coding according to which the triple of binary oppositional sub-alphabets are interconnected by the addition operation modulo two, any pair of its binary representations is sufficient to determine an arbitrary nucleic acid. Therefore, for a two-dimensional visualization of the nucleotide composition, any pairs of coordinate axes. Fig. 2 shows examples of two-dimensional visualization of the genomes of various organisms, next to the order A, G, T, C are pairs of Walsh functions that were used to encode characters. Based on the proposed visualization algorithm, it was found that chromosomes of various species of organisms have individual structural features.



a b

c d

Fig. 2. Examples of two-dimensional visualization (abscissa axis: purine-pyrimidin, ordinate axis: amino-keto, a scaling parameter N = 8) of the genomes:

a - the bacterium Ralstonia eutropha,
b - the bacterium Candidatus Arthromitus,
c - the bacterium Actinomadura madurae,
d - Emiliania huxleyi (a species of coccolithophore found in almost all ocean ecosystems).

The biological significance of this method is evidenced by the fact that randomly generated sequences during algorithmic visualization give a pattern all of whose points are scattered randomly (for this we randomly generated sequences of nitrogen bases with a length of 100,000 nucleotides divided into N-plates of different lengths). Random visual representations are irregular, chaotic in the absence of any mosaics and symmetries across all sub-alphabets, which significantly distinguishes them from real long nucleotide sequences.

With parametric visualization and comparison of the human and monkey genomes, as well as the genomes of other species of living organisms, the similarity of two-dimensional mosaics of arbitrary fragments of genomes with the whole genome was recorded. The visualization of the genomes of various organisms can have a two-dimensional pattern, which is visually similar for all chromosomes and their arbitrary fragments, as well as for the entire organism under consideration. This demonstrates the fractal properties of the structure of nucleic acids and indicates the existence of both interspecific and intraspecific variants of the nucleotide composition. In this case, diagonal or other elements of the pattern can be directed in different directions in different organisms while maintaining the structure of the pattern. As a result of the analysis, it was also found that of the three options for two-dimensional visualization, the most informative and symmetrical are mosaics based on information about the external structure of the molecule, i.e. based on structural elements encoding the amino / keto and purine / pyrimidine traits. Such mosaics have a detailed pattern in which rectangular elements are traced. In some cases, the most pronounced and symmetrical mosaics based on the types of hydrogen bonds reflecting the internal structure of the double helix. Such mosaics are usually characterized by diagonals of the pattern and are found, for example, in the mitochondrial DNA of the arabidopsis thaliana plant. Thus, 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.

Scale invariance or scaling is a property of conservation when all distances are changed by the same number of times. Such similarity transformations form a group of scale transformations. In the task of analyzing the nucleotide composition, scale variation is the resistance of the visual pattern to scaling (changing the parameter N) and is expressed in various organisms to varying degrees. An example of genome visualization at various scale parameters is shown in Fig. 3.

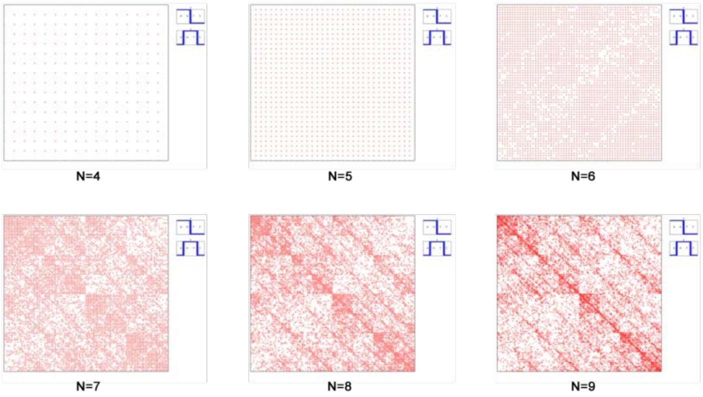
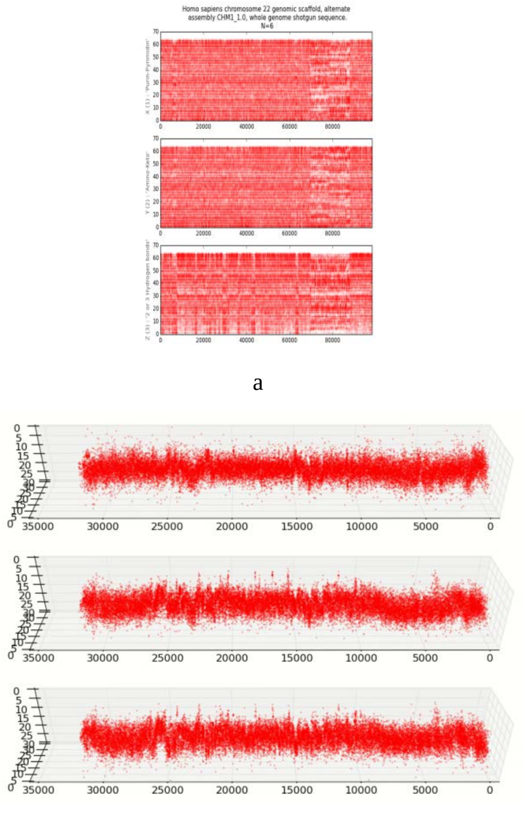


Fig. 3. Two-dimensional visualization of the nucleotide composition of the genome of the chloroplast Fistulifera sp. JPCC DA0580 with various scaling options from 4 to 9.

Changing the scaling parameter N allows the genome to be studied at various levels of detail. This option allows you to adjust the focus of the image. At a certain N, the picture becomes clear, the mosaic becomes pronounced, as a rule, fractal patterns begin to be traced. Thus, the scaling coefficient N plays the role of the resolution of geometric visualization: large N give a small number of points, small N give a small coordinate grid. This circumstance allows us to talk about multiscale analysis in multidimensional parametric spaces.

For further reasoning, we note once again that binary sub-alphabets are interconnected by the addition operation modulo two and thereby define a space with properties for which the coordinates of all points are “glued” to this operation. This follows from the properties of the Hadamard matrix in Fig. 1c. In this regard, it makes sense to consider each sub-alphabet individually as a separate dimension in which the acid is parameterized along its entire length. The abscissa axis encodes the sequence number of the N-plet in the genetic sequence, the ordinate axis displays the decimal values of the binary representation of each N-plet. We suggest calling the corresponding visualization one-dimensional due to the linear nature of the display of the molecule.

Consequently, there are three one-dimensional visualizations between which there is a real algebraic connection (a violation of this connection can lead to diseases and mutations, since it displays a violation in the molecular structure). Using the one-dimensional coordinate axes {X}, {Y} and {Z} gives a triple of mappings using the corresponding sub-alphabets. Fig. 4a shows an example of visualization of a human chromosome, on which regions with different nucleotide composition are clearly visible. These specific regions can be visualized in two-dimensional parametric spaces for their further investigation.



a

b

Fig. 4. a - visualization of the three-channel representation of the nucleotide composition of the fragment of the 1st chromosome of the apple. Each of the three rows (a) corresponds to a binary oppositional sub-alphabet. The abscissa axis encodes the sequence number of the N-plet, the ordinate axis encodes the number of units in the N-plet. The differences in the nucleotide composition for each of the subalphabets are clearly visible**.** b - 3-d visualization of integral three-channel representation of the nucleotide composition of a fragment of the 22nd chromosome of Homo Sapiens. Each of the three rows (b) corresponds to a pair of binary oppositional sub-alphabets.

Areas of the chromosome with various physicochemical parameters, which are clearly visible in Fig. 4a, when two-dimensional visualization is individual in nature due to differences in the nucleotide composition. One-dimensional visualization allows you to display the composition of the molecule in its spatial arrangement, which is closer to physical space than to parametric. In this regard, one-dimensional visualization algorithms are informative for assessing changes in the physicochemical parameters of DNA and RNA molecules along its entire length and with the possibility of scaling.

Any chromosome or other genetic nucleotide sequence can be represented as a set of two-dimensional patterns that follow each other. This requires cutting the sequence into blocks, each of which is analyzed separately in two-dimensional space. The resulting two-dimensional patterns line up one after another. The following visualization algorithm is proposed:

- three one-dimensional representations are constructed for each of the subalphabets (see the example in Fig. 4a);
- areas where there are changes in the nucleotide composition are analyzed by two-dimensional representations.

Another approach to analysis is based on the calculation of the total number of purines, pyrimidines, and other chemical characteristics in each N-measure. In accordance with the methods of the theory of sequential analysis of Harmouth [10], additional visualizations were constructed by the number of elements (zeros or ones) that were found in binary representations of N-plaits in sequences of nitrogen bases. Due to the fact that this method is based on the total number of certain parameters, the corresponding visualization spaces will be called integral. Fig. 5 shows an example of an integral-parametric representation of the nucleotide composition of a human chromosome in one of the planes of two-dimensional visualization. Integral two-dimensional visualizations of real genetic sequences usually take the form of a spot elongated horizontally or vertically. They are less informative than the corresponding two-dimensional fractal mosaics. However, they are convenient for some visualizations (Fig. 4b).

Fig. 4b is an example of visualization of a three-channel representation of the nucleotide composition of a fragment of the 22nd human chromosome. Integral visualization of the total number of units in the codes of N-measures for each of the three pairs of sub-alphabets with the cutting of the chromosome into "windows" of equal length is displayed. This is a variant of the volumetric image, which allows to evaluate the physico-chemical characteristics of the chromosome visually. An additional cut was made into the “windows” of two-dimensional visualization. Each two-dimensional window follows one after another and is an integral representation. This is necessary to assess changes in the nucleotide composition when reading a fragment of a molecule from beginning to end. The depth of recorded changes is determined by the scaling parameter N of the algorithm and the cutting step.

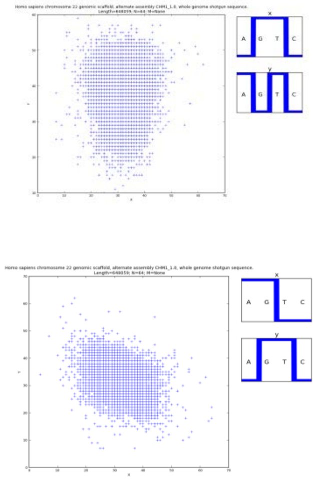


Fig. 5. Illustration of an integrated two-dimensional representation of the nucleotide composition of Homo Sapiens chromosome 22 on one of the visualization planes. A pair of Walsh functions used for parameterization is displayed on the right. The axes of abscissas and ordinates correspond to the number of units of each 64-plet using a pair of binary oppositional sub-alphabets

With integrated visualizations, the scaling parameter is saved, which allows you adjust the information content of the result. Due to the fact that different nucleotide composition is located in different parts of the molecule, rational algorithmic adjustment of the scaling parameter is a separate task and requires additional research in solving practical issues within the framework of the above approach.

In general, the ideas presented for representing genetic information in parametric spaces open up new possibilities for simplifying the perception of genomes using various metrics and spaces (cylindrical, spherical, etc.), since the physicochemical parameters of nucleic acids can be connected to any geometry based on the above algorithm and the principles of finite geometries [2]. At the same time, an algorithmic relationship with real data is preserved, which leaves the method biologically interpretable.

4. Conclusion

The described approaches to the development of research software can serve not only to simplify the perception of long polynucleotide chains and their physicochemical parameters by researchers, but also an additional criterion for the classification and identification of interspecies relationships. In this regard, modern ontologies and thesauruses for organizing and storing biological and molecular genetic data can be supplemented by visualization options for educational purposes, as well as for presenting and searching for biological information.

Fractal patterns, which are obtained by means of the described method, resemble fractal patterns of long nucleotide sequences and amino acid sequences, which were previously obtained by means of the known method "Chaos Game Representation"(CGR-method) in works [11-12] though both methods are quite different in their algorithmic essence. In particularly, CGR-method deals with representations of nucleotide sequences or other long sequences by means of four numbers 0, 1, 2, 3 but not by means of binary numbers 0, 1. In addition our new method seems to be more simple for understanding and using by biologists.

Our analysis of visualizations of the nucleotide sequences of various species of living organisms confirms that the nucleotide composition can be identical in organisms that are not related in the phylogenetic tree and different in related organisms [13-14]. It should be noted that visualization algorithms are implemented in spaces of binary-orthogonal functions. They make it possible to evaluate the types of relations between the present and absent N-measures in the genomes of various organisms and viruses (it becomes clear that these relations are characterized by a cluster structure). In this regard, the results of this study allow us to put forward the developed computer scale-parametric model of nucleic acids as an addition to the structural model of the double helix of J. Watson and F. Crick [4].

The presented concept brings the problem of computer analysis of genetic information to a whole new level. This approach contributes to the optimization of natural intelligence, enhancing its capabilities, since nucleic acids have visual parametric representations that allow one to assess the variability of physicochemical parameters. In addition, computational methods for visualizing both whole DNA and RNA molecules and their fragments substantiate the relationship of their physicochemical parameters with discrete geometry objects [2]. This circumstance can help in the study of internal symmetries, fractality, and other characteristics of nucleic acids to study the complex relationships between different types of living organisms. The emergence of sound methods for comparing viseal models with certain phenotypic traits contributes to the expansion of research methods in bioinformatics and the enrichment of research software.

References

1. A.A. Zenkin, Proc of II Int. Conf Morintech—97, **8**, 197–203 (1997)
2. L.M. Batten, *Combinatorics of Finite Geometries* (Cambridge University Press, 1997) ISBN 0521590140
3. E. Chargaff, R. Lipshitz, C. Green, J Biol Chem. 195(**1**), 155–160 (1952)
4. F.H. Crick, J.C.Wang, W.R. Bauer, J. Mol. Biol. 129(**3**), 449–57 (1979).
5. S.V. Petoukhov, M. He, *Advanced Patterns and Applications* (Medical information science reference, New York, Hershey, 2010) ISBN 978-1-60566-124-7
6. I.V. Stepanian, S.V. Petoukhov, Information, **8**, 12 (2017)
7. The National Center for Biotechnology Information, 14 September 2019, 02:53 UTC, <ftp://ftp.ncbi.nlm.nih.gov/> [accessed 15 September 2019]
8. S. Georgiou, C. Koukouvinos, J. Seberry, *Designs 2002: Further computational and constructive design theory* (Springer, Boston, MA, 2003). ISBN 1-4020-7599-5
9. S.V. Petoukhov, E.S. Petukhova, V.I. Svirin,. Advances in Intelligent Systems and Computing, **754**, 588-600 (2018)
10. H.F. Harmuth, T.W. Barrett and B. Meffert, *Modified Maxwell Equations in Quantum Electrodynamics* (World Scientific, River Edge 2001)
11. H.J. Jeffrey, Nucleic Acids Research, **18**(8), 2163-2170 (1990)
12. C. Yin, J Comput Biol., **26**(2), 143-151 (2019)
13. J.P. Townsend, Z Su,Y Tekle, Genetics. **61**(5), 835–849 (2012)
14. C.J. Norsigian, N. Pusarla, J.L. McConn, J.T. Yurkovich, A. Dräger, B.O. Palsson, Z. King, Nucleic Acids Res., **Nov 7**, gkz1054. (2019)