Cognitive ergonomics of DNA-algorithms

I V Stepanyan
Mechanical Engineering Research Institute of the Russian Academy of Sciences, 4,M. Kharitonyevskiy Lane, Moscow, 101990, Russia
neurocomp.pro@gmail.com

Abstract. To improve the ergonomic qualities of graphical interfaces and optimize the visual perception of large data presented in text form, a method of information processing based on DNA algorithms is proposed. These algorithms were developed by the author in the study of the molecular genetic and structural properties of long nucleotide sequences. The result of processing information presented in text form is an object of discrete geometry, two-dimensional projections of which can serve for visual identification of the internal structure of information, as well as for interpretation as characteristic symbols. At the same time, the considered transformations have the property of visualizing the characteristic cluster structure of information. It is shown that genetic nucleotide sequences, as well as texts in natural language, lend themselves to visual classification, which contributes to an increase in the productivity of mental work when analyzing big data. It is shown that DNA algorithms can reduce the load on the visual analyzer during the initial examination of data presented in text form. This makes it possible to apply this method when operators interact with various information streams and text data of arbitrary nature, as well as to improve the ergonomic qualities of specialized graphical interfaces.

1. Introduction
The quality of information perception depends on the way it is presented. Any information stream can be represented as a sequence of characters of a finite alphabet. A person's ability to perceive information through a visual analyser is closely related to the ability to visualize data. Despite progress in the field of machine learning, the problem of analyzing and visualizing big data has not yet been completely solved. In this regard, in this work, the task is to develop such a computer system that allows to visualize large amounts of information that are not available for quick and primary analysis in their original form due to their significant volume.

This article presents the author's understanding of Gaussian numbers [1] from the standpoint of genetic coding. Gaussian numbers (complex numbers whose real and imaginary parts are integers) have long attracted the attention of researchers. The author initially set the task to represent the DNA molecule with exactly two-dimensional (complex) numbers. This turned out to be possible due to the system of orthogonal Walsh functions encoding the physico-chemical properties of nucleotides when displaying the molecular genetic sequence.

Earlier, to optimize mental work, we developed an algorithm for automatically generated cognitive graphics based on the neurosemantic approach [2-3] for transforming a linear information flow into a hierarchical graph using the mechanism of auto-structuring [4]. This contributes to the measurement,
assessment and regulation of information, the achievement of the adequacy of mental stress and prevention of neuro-emotional overstrain of operators interacting with large amounts of information.

In 1990, John Archibald Wheeler proposed the “it from bit” doctrine, according to which information is a fundamental concept in physics [5]. The Hawking paradox about the disappearance of information in black holes is known [6]. At the same time, information is one of the occupational health risk factors. In [7] was mentioned that from 64% to 90% of computer users experience computer syndrome as a reaction of the body to work with a computer (in particular, it is pain in the back, neck and arms). At the same time, new methods of professional ergonomics are being developed for human interaction with various information systems [8]. In [9], data on clinical, pathogenetic, prognostic and therapeutic aspects of computer syndrome were studied. In [10-12], cognitive factors that must be taken into account when developing decision support systems based on large amounts of data are identified, in particular, cognitive limitations of a person, manifested in the context of data analysis, are considered. Thus, the principles of cognitive ergonomics can help to reduce health risks and optimize a mental work when operators interact with information flows with large amounts of data.

From the point of view of cognitive ergonomics, the currently developing methods of automatically generated computer graphics for the analysis of big data are of interest. In [13], a system for automatic text visualization was developed, which receives text in a natural language as input and creates an interactive three-dimensional scene on its basis. In [14], a method for obtaining information from a text using text network analysis is presented. The automatically generated structure, in combination with other indicators, can be used to determine the level of cognitive diversity of the text.

Technologies for the analysis of genetic information are also being developed. OrganellarGenomeDRAW [15] is a tool for drawing graphic maps of plastid and mitochondrial genomes with access to a local copy of the NCBI RefSeq organelle genome database, which allows visualizing large sets of organelle genomes.

2. Calculation model: genometric approach to big data analysis

Genometry is a group of DNA algorithms for multiscale information display developed by the author through the use of a system of orthogonal Walsh functions with the ability to visualize in various parametric spaces. The basic algorithm has three main steps. 1) Scaling. A sequence of characters from the set \{A, G, C, T\} or \{A, G, C, U\}, encoding nitrogenous bases, is divided into fragments of equal length N, where N is a free parameter of the algorithm - the length of "semantic units". The resulting fragments of equal length will be called N-measures or N-plets. 2) Parametrization. Taking into account the system of genetic sub-alphabets: G = C "3 hydrogen bonds" / A = T "2 hydrogen bonds"; C = T "pyrimidines" / A = G "purines"; A = C "amino" / G = T "keto" sequence of nitrogenous bases can be represented as three binary sequences, consisting of zeros and ones. The choice of encoding method (what is considered to be zero or one) affects the symmetry transformations of the final rendering. 3) Visualization. The resulting binary record of fragments is their representation in the form of three sequences of decimal or other uniquely identifying values (code frequencies, the number of certain symbols in each N-measure, and other characteristic parameters). So, converting binary N-plets to decimal numbers allows to display them in the selected coordinate system. Other numerical interpretations of N-plets for each of the sub-alphabets are possible, considered in particular at works [16-17] based on the matrix genetics of prof. Sergey Petoukhov [18-19]. Therefore, we are talking about a family of DNA-algorithms. The obtained values set the coordinates of points in visualization parametric spaces.

The study of long genetic nucleotide sequences using a genometric approach can be attributed to the problem of visual perception of big data. A conceptual diagram with options for presenting genetic information to the operator's visual analyser is shown in figure 1.

It should be noted that any information flow can be represented as a sequence of characters of some finite alphabet. An arbitrary signal can be recoded and represented in a quaternary code using the genetic alphabet of nucleotides by analogy with how information is stored in a computer's memory in
binary form. The translation of the information signal in the form of a sequence of symbols of an arbitrary finite alphabet into a quaternary code allows the use of DNA visualization algorithms. At the same time, the transformations retain the properties of displaying a characteristic cluster structure of information, if the signal is not chaotic (see figures 2, 3).

Figure 1. Comparison of the variant of visual perception of the nucleotide sequence in the form of a sequence of letter designations (DNA fragment on the bottom) and visualization of the whole chromosome in a two-dimensional space of binary-orthogonal functions of physicochemical parameters (all DNA on the top). The flow of information to the operator's visual analyzer is shown schematically. The left hemisphere is responsible for the logic and analysis of textual information, the right hemisphere is responsible for the genometric image. The DNA algorithm allows you to additionally use the right hemisphere to increase mental productivity.

Figure 2. Genometric infographics. Examples of visualization of genetic information: genome of the bacterium Ralstonia eutropha (a), genome of the bacterium Candidatus Arthromitus (b), genome of the bacterium Actinomadura madurae (c), genome of Emiliania huxleyi (d).
The result of genometric processing of arbitrary information is a visualization, which can be interpreted as a sign or symbol. Various visualization options make it possible to evaluate the symbolic composition of the information signal and track its changes in different sections with the possibility of scaling at different values of N. Also, using scaling, you can evaluate the structure of a particular section of the signal.

Thus, DNA algorithms make it possible to automatically generate cognitive computer graphics for multiscale visualization of the internal structure of signals of various nature. This can be useful for visualizing and identifying areas with different information content within a particular information signal, as well as for comparative analysis of various signals.

3. Results and discussion. Perception of information flows and big data in a text form
At figures 1-4 the results of the transformations that reveal the characteristic structure of the original information is shown. As studies have shown [16], a randomly generated symbolic sequence gives a genometric pattern, all the points of which are scattered chaotically. In this regard, the DNA algorithm can be used to visually evaluate frequency distributions in pseudo-random signals. This can be used, in particular, to evaluate the performance of noise generation algorithms. An example of visualization of the frequency distribution of the signal is shown in figure 3. It can be seen that, according to two subalphabets, the frequencies are periodic, which indicates the periodic nature of the original signal.

![Figure 3](image)

**Figure 3. Genometric infographics.** Examples of tetra [A, G, C, T] representations of various texts in Russian: the Bible phonetic composition, different projections (a), (b), "12 chairs" by Ilf and Petrov in representation of Unicode (c).

The ability to find the internal ordering of the information signal in the form of an individual fractal structure of genometric visualization is an important feature of DNA-algorithms. In this regard, the use of DNA-algorithms for the primary study of big data presented in text form can be used as a tool for the cognitive computer graphics. This makes it possible to apply this method to expand the capabilities of operators' interaction with large data of arbitrary nature, as well as to improve the ergonomic qualities of specialized graphical user interfaces (GUIs).

![Figure 4](image)

**Figure 4.** Frequency representation of the signal, decomposed into three genetic sub-alphabets. The abscissa axis is the serial number of the 12-measure of the original sequence, the ordinate is the frequency of the corresponding 12-measure. The first, second and third sub-alphabets are shown in three separate diagrams (top to bottom).
One of the promising options for using the results of research in the field of ergonomics of perception based on DNA-algorithms is the possibility of help of non-specialists to compare the genetic and phenotypic data of living organisms, as well as to search for previously unknown patterns in other subject areas. The methodology for researching visual information with the involvement of a large number of users has been shown to be effective and is known as “crowd surfing”. It should also be noted the new possibility of labeling food and seeds (by analogy with a barcode) in addition to simple genometric annotation of genomes.

Due to the fact that the features in the nucleotide composition of DNA/RNA of various organisms are clearly visible, it can be concluded that an ergonomic visualization of molecular genetic data based on the proposed approach is possible. For the analysis of textual information, DNA-algorithms are also applicable, but it seems that, unlike genetic texts, texts in natural languages do not differ from each other so much as different DNAs. At the same time, large volumes of textual data of various subjects in natural or programming languages, as well as binary data stored in the memory of computers, can have genometric representations available for visual analysis in various parametric spaces. Visualizations of sound files, biomechanical, biomedical signals, etc. are of practical interest.

4. Genometric Information System
Genometric information system is an information system used to collect, analyse, store and process information using DNA-algorithms. One of the important properties of genometric information systems is the possibility of using technologies of collective intelligence and knowledge engineering. Mind is viewed as a more general category than intelligence. Mind is inherent only in living organisms. Three directions can be distinguished based on a single technical and organizational solution:

- scientific analysis of big data incl. biological, literary and musical;
- application of DNA visualization technologies in the educational process;
- the use of the results and methods in the field of scientific art to change the way of perception of natural and creative objects. Since the method can be positioned within the framework of the direction of scientific art, it becomes possible to attract humanitarian organizations to conduct virtual exhibitions.

Thus, DNA-algorithms allow the creation of new ergonomic instruments for scientific research. Opportunities are also emerging to popularize scientific fields such as molecular genetics and Big Data through the ability to generate specific digital media content.

5. Conclusion
The considered methodology makes it possible to operate with genometric mappings at the level of semantics and semiotics of Big Data. This is possible because the result of processing any information signals using-DNA algorithms are objects of discrete geometry, the projections of which can carry the function of characteristic identification symbols. This is useful in particular for enhancing the ergonomics of graphical user interfaces. At the same time, the possibility of internal interpretation of genometric representations is remains, since DNA-algorithms allow displaying the internal structure of genetic coding.

It is shown that genetic nucleotide sequences (as sequences of symbols of the genetic alphabet) and texts in natural language (as sequences of symbols of the natural language alphabet) lend themselves to visual perception, which contributes to an increase in mental productivity of operators that are interacting with Big Data.

The question of choosing a method for coding information is actual, since this can affect the result of processing by DNA-algorithms. For example, natural language texts can be analysed at the phonetic and morphological levels. In this regard, for the further systematization of genometric mappings, a coding method is important, taking into account the peculiarities of the internal structure of the information signals themselves. It is of particular interest to take into account the cyclic fluctuations of
various biological processes, such as heart rate cycles, circadian rhythms, etc., analysis of the timing structure of musical works, speech intensity, etc. This is possible using the presented approach.

References

[1] Gauss C F 1831 Anzeige von Theoria residuorum biquadraticorum, commentatio secunda Göttingische gelehrte Anzeigen 23 pp 169-78
[2] Stepanyan I V and Denisov E I 2015 A method for reducing information loads based on semantic clustering Occupational Medicine and Industrial Ecology 9 pp 136-7
[3] Stepanyan I V, Denisov E I, Eremin A L, Bodyakin V I and Saveliev A V 2014 Algorithms for the optimization of intellectual work by methods of information visualization using cognitive semantic graphics Neurocomputers: development, application 7 pp 53-9
[4] Bodyakin V I 2012 The concept of building self-learning information management systems based on the neurosemantic paradigm Managing the Development of Large-Scale Systems (MLSD'2012): Proc. of the VI Int. conf. 2 pp 289-98
[5] Plastino A 2004 A conceptual framework for the Wheeler program Physica A: Statistical Mechanics and its Applications 340(1-3) pp 85-91
[6] Alonso-Serrano A, Dąbrowski M and Gohar H 2018 Generalized uncertainty principle impact onto the black holes information flux and the sparsity of Hawking radiation Phys. Rev. D 74(4) 044029
[7] Teo C, Giffard P, Johnston V and Treleaven J 2019 Computer vision symptoms in people with and without neck pain Applied ergonomics 80 pp 50-6
[8] Radwin R G, Lee S, Li K, Lieblich M and Daniel Park B K 2016 Discussion panel on computer vision and occupational ergonomics Proc. of the Human Factors and Ergonomics Society Annual Meeting 60(1) pp 958-60
[9] Munshi S, Varghese A and Dhar-Munshi S 2017 Computer vision syndrome — a common cause of unexplained visual symptoms in the modern era Int. J. of Clinical Practice 71(7) e12962
[10] Walker G and Strathie A 2016 Big data and ergonomics methods: a new paradigm for tackling strategic transport safety risks Applied ergonomics 53 pp 298-311
[11] Araujo C L D O, Santos T J N and Galvão H M 2018 Decision-making by data, an application of data science with cognitive ergonomics Int. J. of Advances in Agriculture Sci. 7 pp 14-9
[12] Kalakoski V, Henelius A, Oikarinen, E Ukkonen, A and Puolamäki K 2019 Cognitive ergonomics for data analysis. Experimental study of cognitive limitations in a data-based judgement task Behaviour & Information Technology 38(10) pp 1038-47
[13] Gehrmann S, Strobelt H and Rush A M 2019 GLTR: Statistical detection and visualization of generated text Preprint arXiv:1906.04043
[14] Paranyushkin D 2019 InfraNodus: Generating insight using text network analysis The World Wide Web Conf. 3584-9
[15] Greiner S, Lehwalk P and Bock R 2019 OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes Nucleic Acids Research 47(1) pp 59-64
[16] Stepanyan I V 2020 A multiscale model of nucleic acid imaging Sci. Visualization 12(3) pp 61−78
[17] Stepanyan I V and Khussein A M 2019 Scaling and visualization of nucleotide sequences EPJ Web of Conf. 224 03007
[18] Fimmel E, Gumbel M, Karpuzoglu A and Petoukhov S 2019 On comparing composition principles of long DNA sequences with those of random ones Biosystems 180 pp 101-8
[19] Petoukhov S V, Petukhova E S and Svirin V I 2019 Symmetries of DNA alphabets and quantum informational formalisms Symmetry: Culture and Sci. 30(2) pp 161-79