Overview of possibilities of genometric information systems

I V Stepanyan and M Y Lednev
Mechanical Engineering Research Institute of the Russian Academy of Sciences, 4,M. Kharitonyevskiy Lane, Moscow, 101990, Russia

neurocomp.pro@gmail.com

Abstract. The actuality of the research is due to the development of methods of matrix genetics and the genometric approach that emerged from it. The article presents the results of combining these developments in a single software package in the form of a genometric information system. The developed information system is implemented in C ++ 14 using the Qt library. The software package implements the parameterization of the following variables: scale parameter, offset relative to the beginning of the sequence, overlap step. The following systems of sub-alphabets are implemented: nucleotide, dinucleotide. Visualization types are implemented: structural, frequency, integral. Visualization spaces are implemented: one-, two- and three-dimensional. In the implementations of the algorithms, methods of parallel computations were applied. The introduction of a new parameter - the overlap step - is presented, which allows obtaining clearer graphic displays. An overview of modern bio-information systems and the main directions of their development is given. Genometric bio-informatics tools play a fundamental role in understanding the modulation mechanisms of biological signals and in proposing new hypotheses.

1. Introduction. Bioinformation systems and the concept of a genometric information system
The size of biological databases is growing rapidly, that requires the development of solutions for their accelerated analysis. Bioinformation systems are the tools for analyzing biological data, including the extraction of a new knowledge. Bioinformatics tools play a fundamental role in understanding the modulation mechanisms of biological signals and in proposing of a new hypotheses.

Today bioinformatics resources are scattered across several databases and often do not have a single portal for annotating and analyzing data. The National Center for Biotechnology Information (NCBI) created the dbSNP database [1], the data of which is integrated with other information sources - GenBank, PubMed, LocusLink, and others [2]. Paper [3] discusses several publicly available bioinformatics tools and data resources that are widely used for integrated multivariate analysis and imaging for cancer immunotherapy purposes. Work [4] describes the GENEASE system with access to more than 50 different databases to facilitate the search for genes and their analysis in real time. At the same time, the totality of bioinformatics resources is rapidly expanding, which poses a challenge for scientists to select appropriate tools. To address this challenge, the European Biological Information Infrastructure has provided a portal https: /bio.tools [5], through which scientific communities and developers can describe and share their own bioinformatics resources.

Highly specialized bioinformation systems are developing intensively. Review [6] is devoted to new automated instruments and tools related to mass spectrometry. In [7], a review of about a thousand bioinformatics tools developed for the study of micro-RNAs was made. The review [8]
collected about one hundred bioinformatics tools in publications devoted to circular RNAs. In [9], it is shown that bioinformatics tools facilitate the acquisition of biological data for visualization of phosphorylation networks for efficient tracking of biological problems in a clear and concise manner. In [10], bioinformatics tools for CRISPR-Cas9 research are considered to facilitate editing of target genes in eukaryotic genomes. Clustering and classification of biological data on gene expression offers a powerful approach to cancer detection [11].

In doctoral dissertation [12], a multithreaded GPU alignment solution using CUDA based on the NVIDIA TESLA C2070 GPU card with 448 cores is proposed. Work [13] describes a Ribbon alignment visualization tool that uses an intuitive visualization technique to compare genomes.

Bioinformation frameworks are also intensively developing. BioJava [14] uses the Java language to implement bioinformatics applications by implementing independent modules. BioShell [15] is an open source package for processing biological data using a set of command line utilities for the Linux environment (functionality of this package assumes knowledge of C++ or Python).

However, it should be noted that despite the emergence of new algorithms for studying large biological data, the disadvantage of all these systems is the impossibility of a simple, understandable and convenient visual identification of genetic objects. Thus, there is a need to develop powerful tools for efficient and effective solution of specific genomics issues based on imaging techniques.

Based on the results described in [16], S.V. Petoukhov convincingly showed that the genetic code is an algebraic code, the features of which are associated with hypercomplex double (or hyperbolic) numbers and double stochastic matrices. Genometry is a group of so-called DNA-algorithms for multiscale information display with the using of a system of orthogonal Walsh functions encoding the main chemical characteristics of DNA, with the possibility of visualization in various parametric spaces. These algorithms were developed in [17-20] based on the matrix genetics of S.V. Petoukhov [16, 21-22]. Genometry contributes to gaining new knowledge about the properties of nucleotide sequences through visualization algorithms that display the cluster structure of genetic information.

In this regard, at this work, the concept of a genometric information system is introduced. This is an information system that serves to collect, analyze, store and process information using DNA algorithms. The relevance of the study is due to the development of methods of matrix genetics and the genometric approach that emerged from it. These developments must be combined into a single software package - a genometric information system for practical use.

The aim of the study is to develop an genometric information system. The task of the study is to review the main capabilities of the genometric information systems. The research objective is the accumulation of genometry methods in a single multi-module bioinformation application.

2. Materials and methods: Walsh genetic functions and Hadamard matrices

The biological meaning of the genometric approach lies in the basic DNA-algorithm, which consists of three main stages - scaling, parameterization, and visualization. Each of the stages has different ways of implementation, which affects the final result of genometric processing. Due to the variety of options at each step of the basic algorithm, we are talking about a group of DNA algorithms. Let's consider these stages.

*Figure 1.* Step of overlapping transformation window.

Scaling stage. A sequence of characters from the set {A, G, C, T / U} is divided into fragments of equal length N. These fragments are called N-measures or N-plets. The development of this stage is the introduction of a new parameter - the overlap step. This parameter provides overlapping of N-plets (figure 1). It turned out that the introduction of this parameter allows to obtain a clearer genometric
visualizations, which is associated with an increase in the number of points, the coordinates of which reflect N-plets that displays the nucleotide composition.

Parameterization stage. There is a known system of genetic sub-alphabets that implement the system of Walsh’s orthogonal genetic functions (figure 2): G = C "3 hydrogen bonds" / A = T "2 hydrogen bonds"; C = T "pyrimidines" / A = G "purines"; A = C "amino" / G = T "keto". At the stage of parameterization, the nucleotide sequence is represented in the form of three binary sequences (channels) in accordance with the system of these genetic sub-alphabets.

Visualization stage. The resulting binary record of fragments is their representation in the form of three sequences of uniquely identifying values (relative or absolute frequencies, the number of certain symbols in each N-measure, decimal representations and other parameters of N-plets) with visualization in one or another coordinate system.

![Figure 2. Purine bases: adenine, guanine, pyrimidine bases (from left to right): uracil, thymine, cytosine. Left - a variant of the Hadamard matrix, displaying the encoding of nucleotide sub-alphabets.](image)

There is a variant of the DNA algorithm for dinucleotide imaging based on the alphabet system of 16 doublets. In [21] a system of sub-alphabets are given, which were obtained by analyzing the mosaic matrix of doublets [C A; T G] as variants of the decomposition of this matrix. These sub-alphabets are shown in table 1.

| Table 1. Sub-alphabets for systems of 2-block U-complex numbers, which S V Petoukhov obtained in the analysis of the Walsh representation of the mosaic genomatrix [C A; T G]⁹. |
|---|---|
| 1 sub-alphabet UC1 | 3 sub-alphabet UC3 |
| CC = CG = TC = TG = 00e | CC = GC = CA = GA = 00s |
| TT = TA = CA = CT = 01e | AC = TC = TA = AA = 01s |
| GG = GC = AG = AC = 10e | GG = CG = GT = CT = 10s |
| AA = AT = GT = GA = 11e | TG = AG = AT = TT = 11s |
| 2 sub-alphabet UC2 | 4 sub-alphabet UC4 |
| CC = CG = AT = AA = 00q | CC = GG = AG = TC = 00p |
| CA = CT = AC = AG = 01q | AA = TT = CT = GA = 01p |
| GG = GC = TA = TT = 10q | AT = TA = CA = GT = 10p |
| GT = GA = TC = TG = 11q | CG = GC = AC = TG = 11p |
3. Results and discussion. Overview of the main capabilities of genometric information systems

The developed information system is implemented in C++ 14 using the Qt library. In the implementations of the algorithms, methods of parallel computations were applied. The main feature of the developed genometric information system is the multi-scale visualization of the physicochemical parameters of genetic nucleotide sequences in various parametric spaces, taking into account various systems of genetic sub-alphabets.

Compared to earlier versions of the software implementation in the Python language, the computational speed has significantly increased, new possibilities have appeared for constructing projections of the DNA nucleotide composition (including triangular ones, see figure 4). A convenient graphical user interface has appeared. Switching projection types is implemented in the interface of the software package.

Two additional transformations were implemented - frequency and integral. As noted, when building visualizations, it is possible not only to convert N-plexes from binary to decimal representation (structural transformation). With the frequency type of visualization, each coordinate of every point is the frequency of occurrence of the N-plet in the sequence for the corresponding sub-alphabet. In the integral type of visualization, each point corresponds to coordinates, the values of which are calculated by summing the ones (or zeros) of the corresponding N-plet of the corresponding sub-alphabet.

The software package implements the parameterization of the following variables: scale parameter N. The parameter N can be set from the range 1-1023; offset relative to the beginning of the sequence; step of overlapping the transformation window. The following systems of sub-alphabets have been implemented: nucleotide (figures 3-5); dinucleotide, based on alphabets of 16 doublets (figure 6). The following types of visualization are implemented: structural (figures 3-6), integral (figures 8-9), frequency (figures 7, 10). The following spaces are implemented: one-dimensional (figure 7), two-dimensional (figures 3-9), three-dimensional (figure 10).

Figure 3. Two-dimensional structural genomics of the nucleotide composition of the genome of coronavirus at different parameters of the overlap step. The abscissa and ordinate axes correspond to the decimal representations of the binary encoding of each N-plet.
**Figure 4.** Triangular projections of three-dimensional representation according to the system of nucleotide sub-alphabets of the genome composition of the anaerobic bacterium.

**Figure 5.** Two-dimensional structural genometrics of 3 projections of the nucleotide composition of the actinobacterium genome. The abscissa and ordinate axes correspond to the decimal representations of the binary encoding of each N-plet.

**Figure 6.** Two-dimensional structural genometrics of 3 projections of the nucleotide composition of the actinobacterium genome. The abscissa and ordinate axes correspond to the decimal representations of the binary encoding of each N-plet.
Figure 7. One-dimensional frequency genometrics of the nucleotide composition of the genome of the anaerobic bacterium in three-channel representation.

Figure 8. Two-dimensional integral genometrics of the nucleotide composition of the genome of the first chromosome of gray mouse. The coding step is 1. The abscissa and ordinate axes correspond to the number of ones in the binary coding of each N-plet.

Figure 9. Two-dimensional integral genometrics of the genome of the first human chromosome. The coding step is 1. The abscissa and ordinate axes correspond to the number of ones in the binary coding of each N-plet.
In figure 10 the effect of natural frequency clustering in the three-dimensional space of nucleotide sub-alphabets is shown. Similar clustering effects are inherent in the frequency distributions of N-mers in the DNA of living organisms. The corresponding tools are provided in the developed genometric information system.

A submodule for simultaneous display on a whole range of changes of the parameter N (with a certain step) at once, which facilitates viewing several scales in one window has been developed. An example of such of multiscale visualization of genetic structures is shown at figure 11.

Figure 10. Three-dimensional frequency genomics of fly genome chromosome X.

Figure 11. Multiscale one-dimensional structure visualization of the red bayberry.

4. Conclusions

- The developed genometric information system is a complex of computational and logical modules and means of multiscale visualization: structural, frequency and integral in various parametric spaces of various dimensions.
- The work shows some options for visualization. Fractality is a characteristic property of visualizations. This indicates the fractional dimension of genomic objects that characterize the fractal properties of the objects under study - DNA and RNA of various organisms and viruses.
- Implemented in the genometric information system, the toolkit is also applicable for symbolic sequences of various nature, after their transformation to tetra-code. This will allow studying the properties of digital signals with the using of genometric information systems.
- The transition from Python to C++ allowed us to accelerate the construction of mappings by ~10 times and achieve visualization capabilities with large N (up to 1023).
- The concept of genometric information systems contributes to the development of bioinformatics due to the instrumental means of monitoring genetic information, the ability to scale and select the type of DNA algorithms.

References

[1] Sherry S T, Ward M and Sirotkin K 1999 Genome Res. 9 pp 677-9
[2] Sherry S T, Ward M, Kholodov M, Baker J, Phan L, Smigielski E M and Sirotkin K 2001 dbSNP: the NCBI database of genetic variation Nucleic Acids Research 29(1) pp 308-11
[3] Palmisano A, Krushkal J, Li M C, Fang J, Sonkin D, Wright G and McShane L 2020 Bioinformatics tools and resources for cancer immunotherapy study Biomarkers for immunotherapy of cancer (New York: Humana) pp 649-78
[4] Ghandikota S, Hershey G K K and Mersha T B 2018 GENASE: real time bioinformatics tool for multi-omics and disease ontology exploration, analysis and visualization Bioinformatics 34(18) 3160-8

[5] Ison J, Ménager H, Brancotte B, Jaaniso E, Salumets A, Raček T and Hancock J M 2020 Community curation of bioinformatics software and data resources Briefings in bioinformatics 21(5) 1679-705

[6] Tsai P L and Chen S F 2017 A brief review of bioinformatics tools for glycosylation analysis by mass spectrometry Mass Spectrometry 6(2) 64

[7] Chen L, Heikkinen L, Wang C, Yang Y, Sun H and Wong G 2019 Trends in the development of miRNA bioinformatics tools Briefings in Bioinformatics 20(5) 1836-52

[8] Chen L, Wang C, Sun H, Wang J, Liang Y, Wang Y and Wong G 2020 The bioinformatics toolbox for circRNA discovery and analysis Briefings in Bioinformatics

[9] Shi X X, Wu F X, Mei L C, Wang Y L, Hao G F and Yang G F 2020 Bioinformatics toolbox for exploring protein phosphorylation network Briefings in Bioinformatics

[10] Alkhnbashi O S, Meier T, Mitrofanov A, Backofen R and Voß B 2020 CRISPR-Cas bioinformatics Methods 172 pp 3-11

[11] Mabu A M, Prasad R and Yadav R 2020 Mining gene expression data using data mining techniques: A critical review J. of Information and Optimization Sci. 41(3) pp 723-42

[12] Said M A M 2017 Parallel acceleration of bioinformatics algorithms Doctoral dissertation

[13] Nattestad M, Aboukhalil R, Chin C S and Schatz M C 2020 Ribbon: Intuitive visualization for complex genomic variation Bioinformatics

[14] Bhongale M P and Vyas R 2020 Role of BioJava in the Department of bioinformatics tools Advances in Bioengineering Springer Singapore pp 29-49

[15] Macnar J M, Szulc N A, Kryś J D, Badaczewska-Dawid A E and Gront D 2020 BioShell 3.0: Library for processing structural biology Data Biomolecules 10(3) 461

[16] Petoukhov S V 2020 The genetic code, algebraic codes and double numbers

[17] Stepanyan I V and Petoukhov S V 2017 The matrix method of representation, analysis and classification of long genetic sequences Information 8(1) 12

[18] Stepanyan I V and Khussein A M 2019 Scaling and visualization of nucleotide sequences EPJ Web of Conf. IV Int. Conf. Modeling of Nonlinear Processes and Systems 224 (Russia: Moscow)

[19] Stepanyan I V 2020 Scale-parametric visualization of nucleic acids Sci. visualization 12(3) pp 61-78

[20] Stepanyan I V 2020 Biomathematical system of methods for describing nucleic acids Computer research and modeling 12(2) pp 417-34

[21] Petoukhov S V 2017 Genetic coding and united-hypercomplex systems in the models of algebraic biology Biosystems 158 pp 31-46

[22] Petoukhov S V, Petukhova E S and Svirin V I 2019 Symmetry of DNA alphabets and quantum informational formalisms Symmetry: Culture and Sci. 30(2) pp 161-79