3D Visualization to Analyze Multidimensional Biological and Medical Data

V. L. Averbukh, I. O. Mikhailov, M. A. Forghani, P. A. Vasev

Abstract Biological and medical databases continue to grow in size, volume, and dimension that lead to facing big data issues. The data obtained as a result of complex computer modeling, as well as in analyzing various sources of big data are complex and poorly structured. Visualization of such data is an important task for their interpretation that affects a final obtained decision from data. Since traditional approaches such as projection, the use of pictograms, colors, shapes, etc., are not enough to demonstrate the multidimensional relationship, it is necessary to develop a visualization system that is flexible to represent the desired visualization for an expert, doctor or researcher. The aim of the current paper is the development of visualization systems for multidimensional medical and biological data with additional reality. The main idea is the set of projections from multidimensional space to three- dimensional cube and representation of patients' data in the form of points cloud. The remarkable advantage is that the proposed system is user-friendly and flexible to define visualization axes. Moreover, additional reality provides better visualization of the information content. In case of clustering of proteins by genomic signal processing techniques, physico-chemical properties of amino acids can be

V. L. Averbukh
N.N. Krasovskii Institute of Mathematics and Mechanics, 620990, 16 S. Kovalevskaya str., Ekaterinburg, Russia,
Ural Federal University, 620002, 19 Mira street, Ekaterinburg, Russia,
e-mail: averbukh@imm.uran.ru

I. O. Mikhailov
Ural Federal University, 620002, 19 Mira street, Ekaterinburg, Russia,
e-mail: igormich88@gmail.com

M. A. Forghani
Ural Federal University, 620002, 19 Mira street, Ekaterinburg, Russia,
e-mail: majid.forqani@gmail.com

P. A. Vasev
N.N. Krasovskii Institute of Mathematics and Mechanics, 620990, 16 S. Kovalevskaya str., Ekaterinburg, Russia, e-mail: vasev@imm.uran.ru

used to convert an alphabetical sequence to numerical. Since there are many possible conversions using AAindex database, we suggest using dimensional reduction methods before genomic signal processing. This decreases the time of computation, provides the overall picture of physico-chemical changes and increases the quality of visualization. A wavelet-based algorithm can represent the relationship between proteins in different scales. Using this idea, a user is able to define the visualization scale to see small or large differences between protein sequences.

1 Introduction

Biological databases are growing fast due to the progress and development of genetics technologies such as high-throughput sequencing. The major objectives of biological databases are to store, organize and share data in a structured and searchable manner, with the aim to facilitate data retrieval and visualization for humans, and also to provide web application programming interfaces (APIs) for computers to exchange and integrate data from various database resources in an automated manner [20]. Besides biological databases, medical are growing in quantity and quality due to the growth of better measuring in the medical system in recent decades. This leads the researcher or doctor to face with interpretation task of a multidimensional database.

Approaches for visualization of the multidimensional databases are considered in the scientific literature for several decades (see, e.g., the overview in [19]). Considered as general approaches to the visualization of multidimensional data [5], and specialized implementations of systems that provide a representation of large amounts of data obtained as a result of mathematical and computer modeling of complex phenomena and processes [17]. As an example, a visual analytics framework presented in [15], that is used for effective treatment decisions from complex genomics data. Visual data mining techniques play an important role in exploratory data analysis. Data mining aims to search and analyze data to find useful information. An idea for such visualization is to represent as many data items as possible by mapping each data value to a pixel and arranging the pixel adequately [10]. One of the most common methods for representing multidimensional data is their projection onto a two-dimensional or three-dimensional space. For example, back in 1991, the idea of Hyperbox was considered. A hyperbox is a two-dimensional depiction of an N-dimensional box (rectangular parallelepiped). The authors [1] defined the visual syntax of hyperboxes, state some properties, and sketch two applications. Hyperboxes can be evocative visual names for tensors or multidimensional arrays in visual programming languages. They can also be used to display all pairwise relationships in an N-dimensional data set simultaneously. This can be helpful in choosing a sequence of dimension-reducing transformations that preserve interesting properties of the dataset.

To represent, in practice, the multidimensional data arising from the analysis of dynamic networks, the idea of Matrix Cube is suggested in [2]. Matrix Cube is a

novel visual representation and navigation model for dynamic networks; inspired by the way people comprehend and manipulate physical cubes. Users can change their perspective on the data by rotating or decomposing the 3D cube. These manipulations can produce a range of different 2D visualizations that emphasize specific aspects of the dynamic network suited to particular analysis tasks. The closed ideas can be found in [14], which describes a system designed for visual analysis of multidimensional data. The developed system can display a multidimensional cloud of data and allows the user to analyze it in a lower-dimensional space (2D and 3D), propose and test various hypotheses about the original data, with the possibility of making assumptions for using calculating techniques, using geometric constructions in interactive mode.

An important factor in visualization is the user's interaction. A human can analyze complex events within a short time, to find important information to make a decision. Comparing with a computer, human handles with vague descriptions and inaccurate knowledge, using general knowledge, easily makes complex conclusions [10]. The performance of visualization can be improved considering better user's interaction with a visualization system. A tool called Interaction+ [12] was developed that enhances the interactive capability. It takes existing visualizations as input, analyzes the visual objects, and provides users with a suite of interactions to facilitate visual exploration. Another idea of an interactive system had been presented in [9], in which a set of low-dimensional parallel coordinates plots are interactively constructed by sampling user-selected subsets of the high-dimensional data space. This allows a user to specify the most relevant lowe-dimensional data and provides the visualization of the most meaningful dimensions. The interactive visual analytics tool, Winnows [3] had been designed to enable users to easily filter and compare patient subgroups based on data visualization of multiple outcome measures. It also provides the investigation of inter-relationships across outcome measures in various domains or relationships between multiple disease features and their changes over time.

Recently, two visualization systems have been developed by us, one for medical and another for biological data. The first system is an interactive visual analytic system for medical data. It assumes the use of a projection of multidimensional space into a three-dimensional cube. It provides an ability for a user to choose a set of measurements to be mapped on cube axes. Furthermore, it allows mapping other data dimensions onto visual attributes like color, marker shape and, size, etc. [13]. In the second system for biological data, a new dimension was defined and added to the phylogenetic tree to track the physico-chemical changes in proteins. Moreover using multidimensional scaling, the physico-chemical properties space dimension is decreased that presents general changes in the protein. The wavelet-based algorithm considers the neighbor effect of amino acids in a new dimension. Also, virtual reality was added to improve the quality of 3D visualization of the phylogenetic tree [7].

2 Visualization of Medical Data

In major biological systems, we only speculate on the process that reveals the relationship between different variables and the visual exploration helps to understand relationships, processes or forming a hypothesis. Novel multidimensional visualization techniques enable us to display large, high dimensional data set in a meaningful, more descriptive manner [4]. Users' understanding of the visualization and interpretation defines the way that the system can interact with the user. Since exploration plays an important role in diagnostic from medical data and to enhance the interactivity, our idea is the visualization based on user-organizing projections.

As a result of the examination of a large number of patients, significant amounts of multidimensional data have been obtained. Visualization and analysis of multidimensional data is an important area for many scientific fields. It should be noted that there are no general approaches to the visualization of multidimensional sets, although important results have been obtained in special cases and there are many publications on this subject. In the presented work, methods development for visualization of multidimensional medical data collected by the medical system "qMS" and provided by the company SPARM was considered. The work was carried out as part of the project to analyze the Big Data of the Academic Partnership Dell EMC (project healthcare Optimization, Dell EMC External Research and Academic Alliances - ERAA Dell EMC). Our goal is to support the data analysis for Medical Information System (MIS). The records of MIS qMS [11] collected during the period from 2013 to 2017, from Russian hospitals. Patients have Diabetes Mellitus Type 1 and Arterial Hypertension. Patients' data includes ICD-10 clinical diagnoses, records about implemented investigation procedures, operations, pharmacological treatment.

In our case, the data represents the results of patients monitoring collected by one of the clinics. This study aims to analyze the efficiency of treatment. That is determined by a set of parameters, which can be considered as measurements of the obtained data space. It is suggested a set of projections of a multi-dimensional discrete space into a three-dimensional cube and representation of patients' data in the form of points cloud (see figure 1). The possibilities of this visualization system include the ability to simultaneously display up to 5 axes with the ability to interactively highlight clusters and automatically find the correlation. It is also possible to split the data into groups according to several characteristics and compare them (see figure 2). Thus, the user, a specialist in the field, has the opportunity to independently select the visual mapping, necessary for the analysis and interpretation of real data. The system is developed by free software products and is cross-platform.

Briefly, an interactive environment for the 3D-visualization of MIS data was developed. The method of analysis was applied to a sample of patients with type 1 diabetes. The multidimensional data space is considered, where the characteristics of patients and the results of their examination and treatment (columns of the metadata table) can be used as measurements. The developed prototype of the system allows combining several types of data in a single three-dimensional field. There is the possibility of scaling and hyperactive detailing information about each specific

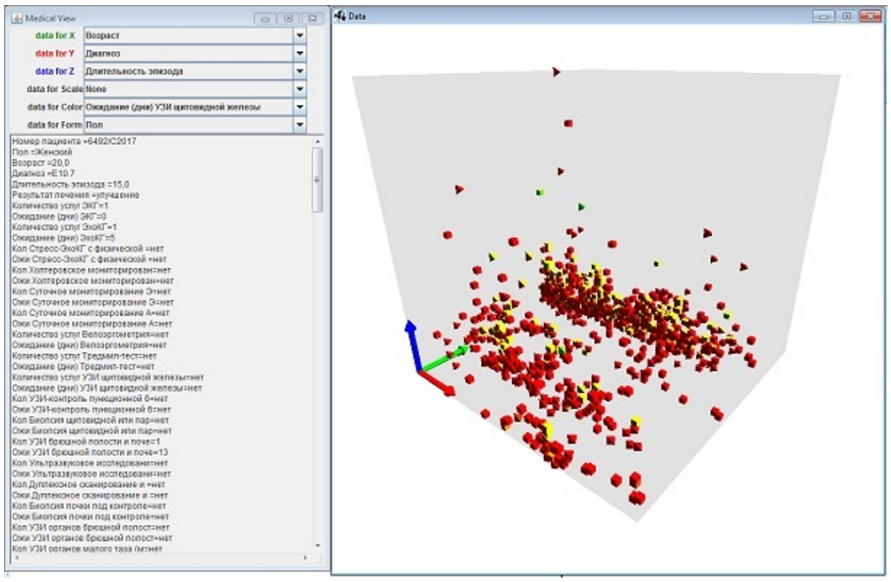

Fig. 1 The interface window of the developed system for 3D-visualization of MIS data (Patients with type 1 diabetes). The possibility to visualize patients' data in a three-dimensional virtual space is shown virtual space can be rotated, and data for a particular patient can be seen separately, choosing from a set of figures. The vertical axis shows the duration of the hospitalization episode. The color shows the waiting time for the thyroid ultrasound: yellow color corresponds to the average expectation of ultrasound, green to the long waiting for ultrasound; red color corresponds to the absence of ultrasound. This configuration of options allows the doctor to immediately take a look at the picture of the distribution of patients for the duration of hospitalization and wait for an ultrasound of the thyroid gland, in conjunction with age, sex and codes/subsections for ICD-10 clinical diagnoses.

patient. It is possible to change the set of measurements during the analysis of data, and visual space can simply be rotated (see figures 3, 4 and 5). In the future, it will be investigated the possibility of virtual and extended reality (or additional reality) usage within the system.

3 3D Visualization of Phylogenetic Tree

Evolutionary tree diagrams can be found in even the earliest descriptions of evolution, and their visualization still plays a key role in modern phylogenetics. However, although trees visualize an organism's evolutionary history, tree's construction is based on biological data which in turn contains the information that distinguishes each organism. Sequence alignments are the most common data used in phylogenetic analysis, and their visualization assists in understanding the molecular mech-

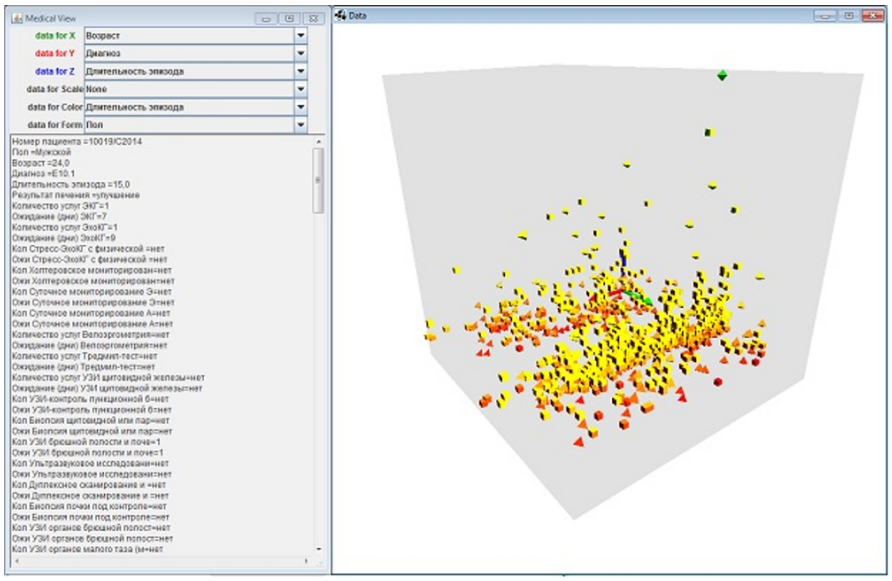

Fig. 2 In this case, the length of the hospitalization episode is mapped both to the vertical axis and to the color.

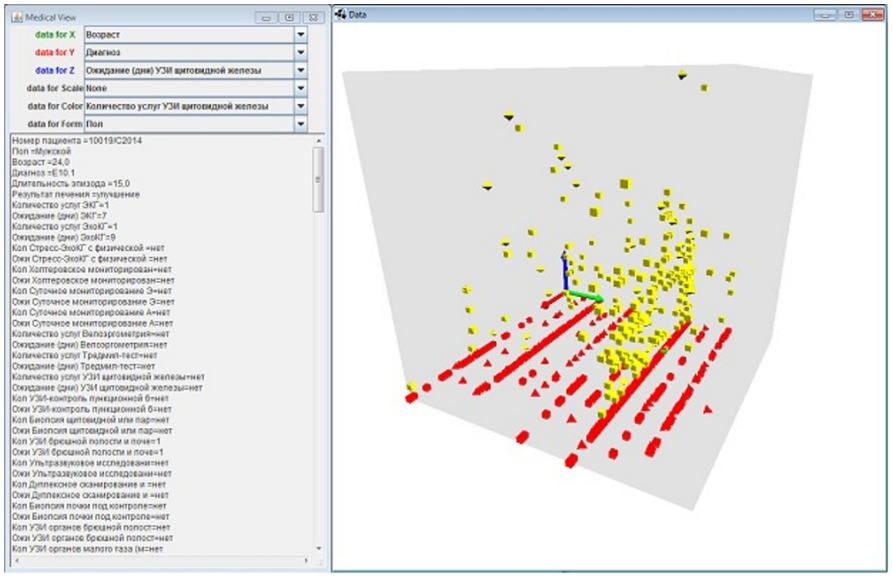

Fig. 3 The vertical axis is the waiting time of the thyroid ultrasound. The red color of the marker represents no ultrasound and yellow color represents one examination.

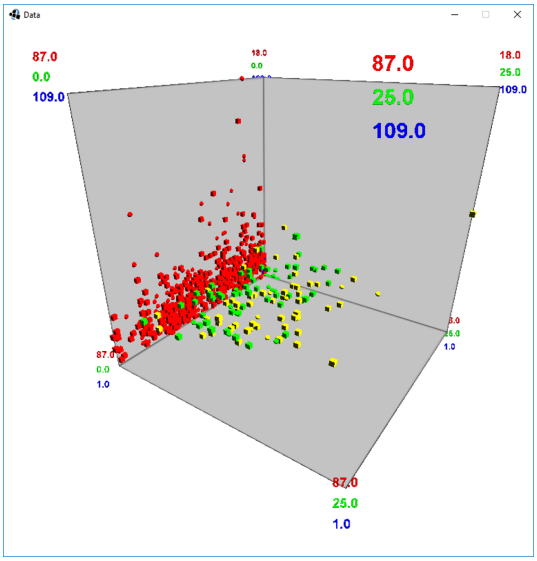

Fig. 4 Quantity and waiting time for monitoring of the electrical activity of the cardiovascular system with Holter monitor. (Holter monitor is a type a portable device for cardiac monitoring). The vertical axis is a duration of hospitalization. Front-looking axis (increases from back to front) indicate the age of patients. The horizontal axis (increases from left to right) indicate waiting time for the medical procedure. The form of elements indicates the sex of patients (spheres for men, cubes for women). The color of elements indicates the count of medical procedures: red means no procedures at all, yellow means one and green means two. Many red elements indicate patients that don't have that procedure and their waiting time is zero. These elements can be filtered to improve visualization.

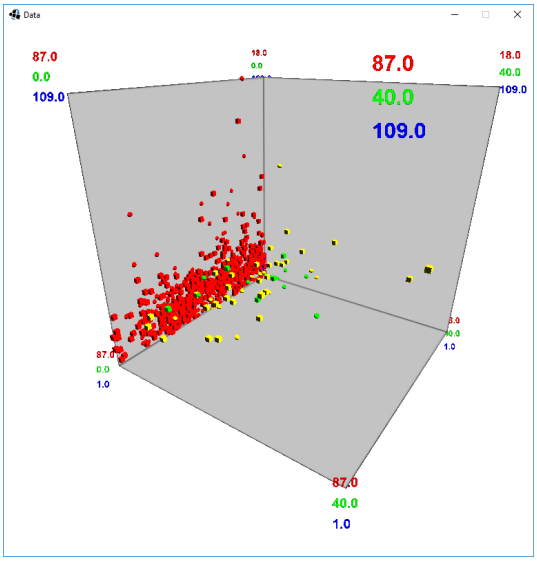

Fig. 5 Radiography of joints/bones of hands and feet.

anisms that differentiate each species, down to the level of the individual nucleotide bases and amino acids [16].

To better visualization of the tree with a mass of leaves, it was suggested to use 3D visualization. The idea of visualizing phylogenetic trees in three-dimensional hyperbolic spaces with the Walrus graph visualization tool was introduced in [8]. This system can visualize and navigate phylogenetic trees with more than 100,000 nodes. Recently a 3D-visualization of a phylogenetic tree has been developed [7] by adding significant information from physico-chemical changes in amino acids as a new dimension (see figure 7).

To apply genomic signal processing methods on protein data, a primary alphabetical sequence is converted to numerical one. The numerical representation should reflect biological properties in the numerical domain. A way to define such conversion is by using an amino acid index that includes 20 numbers of an amino acid property. A rich collection of indices can be found in AAindex database (www.genome.jp/aaindex/).

Previous researchers had indicated there is a correlation between amino acid substitutions and their physico-chemical properties. Each of these properties provides a viewpoint in the study of biological functions. Taking into account all of them leads to a multi-viewpoint representation and provides more options to observe and study the target biological phenomena. In other words, the combination of all amino acid physico-chemical properties would result in a complex high-dimensional feature space, possibly including redundant features [6] and causes a sophisticated visualization.

To handle this issue, before the conversion of protein sequences to numerical, we suggest considering a dimensional reduction on AAindex using clustering and principle component analysis (PCA) (see figure 6). Beside of AAindex data, new indices obtained by AAindex clustering gives the user to choose an individual specific property from AAindex or a general picture of property changes from new extracted indices (see figure 9).

The system clusters tree leaves according to the physico-chemical property of amino acid while each leaf has its protein sequence. Information that is used for clustering includes physico-chemical changes as well as neighbor effect, the effect of adjacent amino acids on a target amino acid in protein primary sequence. This information is achieved by a new algorithm developed by us based on the wavelet packet transform. An improvement of visualization can be done by considering the demonstration of leaves relationships in different scales of the wavelet. This allows observing the overall picture of changes while it is possible to see small changes between protein sequences during the evolution using phylogenetic tree. In addition to usual 3D-visualization, virtual reality is provided(see figure 8) [18]. Due to the limitation of monitor view, it is difficult to visualize a complex tree. The virtual reality can dramatically increase the information content of visualization and provides a wide range of view to see the general picture of a tree with details.

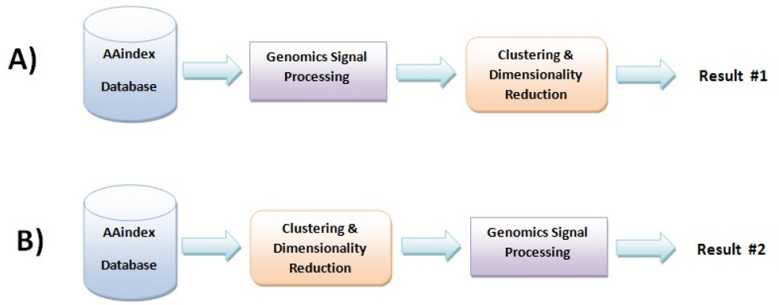

Fig. 6 Two methods for clustering of protein sequences. Method A uses AAindex database for numerical representation of protein and then applies genomic signal processing techniques, each index can provide a different result, while method B uses a few indices and gives a overall picture of physico-chemical changes. Method B is considered in this paper.

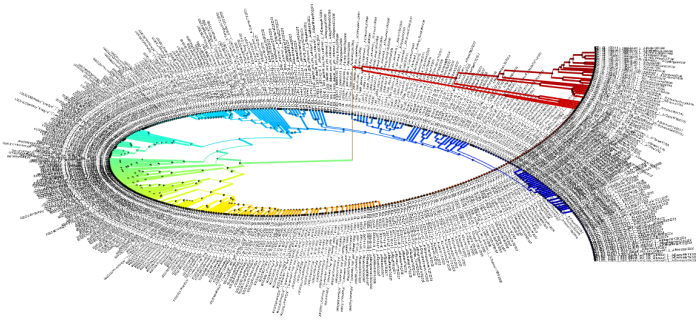

Fig. 7 A general presentation of the 3D phylogenetic tree of influenza virus (without virtual reality).

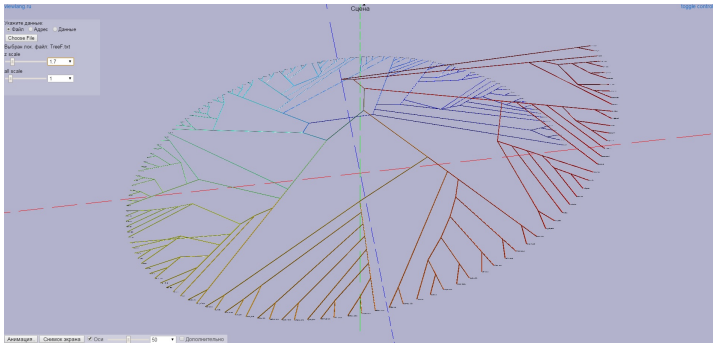

Fig. 8 A general presentation of the 3D phylogenetic tree in virtual reality environment (note that this tree is different from the tree in figure 7).

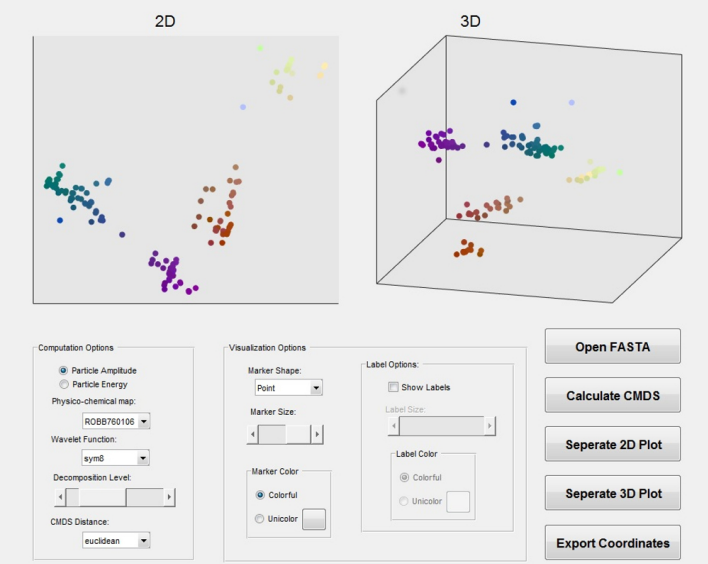

Fig. 9 2D & 3D representation of clustering of hemagglutinin protein sequences using the wavelet-based algorithm and physico-chemical properties.

4 Conclusion and Future Work

Both of the presented systems define visualizations that are flexible to interact with. Some medical parameters have priority over others for decision making. Considering this priority, a doctor or specialist can arrange the visualization to see a specific relationship between different parameters in data and increase the speed of decision making.

Taking the influenza virus as a case study, besides the genetic relationship visualization (in phylogenetic tree) and antigenic relationship visualization (in antigenic cartography), the physico-chemical changes in amino acids provide additional information to better describe the evolutionary processes. Since this information also includes the neighbor effect extracted by the wavelet-based algorithm, they can directly be used in mathematical modeling of biological functions. The clustering of phylogenetic tree leaves and adding virtual reality representation affords an interactive environment for the researcher to explore and find a simple interpretation of complex data.

At the next stage of our research and development, it is supposed to try out other methods of multidimensional visualization to improve the results of mathematical modeling. For the visualization of medical data, the platform will be translated into web-based visualizations by using Viewlang (viewlang.ru) system and to provide an interactive virtual reality presentation. In the case of the phylogenetic tree, we plan to improve the accuracy of the algorithm by applying the principal component analysis in a different level of wavelet decomposition. Depend on the wavelet fam-

ily, the obtained components of PCA in the level of decomposition can be varied. Accordinlgy, the choice of proper wavelet family for the visualization is the subject for future work.

References

1. Alpern, B. and Carter, L.: The hyperbox. In Visualization, 1991. Visualization91, Proceedings., IEEE Conference on (1991), IEEE, pp. 133139.
2. Bach, B., Pietriga, E., Fekete, J.-D.: Visualizing dynamic networks with matrix cubes. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems(CHI 2014), Apr 2014, Toronto, Canada (2014), ACM, pp. 877886.
3. Cheng, H.C., von Coelln, R., Gruber-Baldini, A.L., Shulman, L.M. and Varshney, A.: Winnow: Interactive Visualization of Temporal Changes in Multidimensional Clinical Data. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (pp. 124-133). ACM.
4. Cvek, U., Trutschl, M., Stone, R., Syed, Z., Clifford, J.L. and Sabichi, A.L.: Multidimensional visualization tools for analysis of expression data. World Acad Sci Eng Technol 54, 54 (2009), 281289.
5. Dzemyda, G., Kurasova, O. and Zilinskas, J.: Multidimensional data visualization: Methods and applications. Springer Optimization and Its Applications, ISSN 1931-6828 72 (2012).
6. Eng, C. L., Tong, J. C., Tan, T. W.: Predicting host tropism of influenza a virus proteins using random forest. BMC medical genomics 7, 3 (2014), S1.
7. Forghani, M., Vasev, P., Averbukh, V.: Three dimensional visualization for phylogenetic tree. Scientific Visualization 9, 4 (2017), 5966.
8. Hughes, T., Hyun, Y. and Liberles, D.A.: Visualising very large phylogenetic trees in three dimensional hyperbolic space. BMC bioinformatics 5, 1 (2004), 48.
9. Itoh, T., Kumar, A., Klein, K. and Kim, J.: High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots. Journal of Visual Languages and Computing, 43, pp.1-13.
10. Keim, D.A. and Kriegel, H.P.: Visualization techniques for mining large databases: A comparison. IEEE Transactions on knowledge and data engineering 8, 6 (1996), 923938.
11. Kolesnichenko, O., Kolesnichenko, Y., Minushkina, L., Mazelis, L., Mazelis, A., Nikolaev, A., Shahgeldyan, C., Averbukh, V., Mikhailov, I., Martynov, A., Pulit, V., Dolzhenkov, A., Grigorevsky, I., Smorodin, G.: Big data analytics of medical information system (mis) records. In Proceedings of National Supercomputer Forum (NSCF). The Program Systems Institute of RAS, Pereslavl-Zalesskiy, Russia (2017), NSCF.
12. Lu, M., Liang, J., Zhang, Y., Li, G., Chen, S., Li, Z., Yuan, X.: Interaction+: Interaction enhancement for web-based visualizations. pp. 6170.
13. Mikhailov, I., Averbukh, V.: Design and development methods for visualization multidimensional discrete data. In Proceedings of National Supercomputer Forum (NSCF). The Program Systems Institute of RAS, Pereslavl-Zalesskiy, Russia (2017), NSCF.
14. Maslenikov, O. P., Milman, I. E., Safiulin, A. E., Bondarev, A. E., Nizametdinov, S. U., Pilyugin, V. V.: Development of a system for analyzing of multidimensional data. Scientific Visualization 6, 4 (2014), 3049.
15. Nguyen, Q. V., Khalifa, N. H., Alzamora, P., Gleeson, A., Catchpoole, D., Kennedy, P. J., Simoff, S.: Visual analytics of complex genomics data to guide effective treatment decisions. Journal of Imaging 2, 4 (2016), 29.
16. Procter, J. B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F., Barton, G. J.: Visualization of multiple alignments, phylogenies and gene family evolution. Nature methods 7 (2010), S16S25.

17. Perevalov, D. S., Vasev, P. A.: On development of methods of multidimensional visualization. In Proceedings of GraphiCon 2002 (2002), pp. 431437.
18. Vasev, P.: Three-dimensional visualization in a web based environment based on qml declarative description. In Proceedings of International (47th all-Russian) youth school-conference, Yekaterinburg, January 31 - February 6, 2016 (2016).
19. Wong, P. C., Bergeron, R. D.: 30 years of multidimensional multivariate visualization. In Scientific Visualization (1994), pp. 333.
20. Zou, D., Ma, L., Yu, J., Zhang, Z.: Biological databases for human research. Genomics, proteomics and bioinformatics 13, 1 (2015), 5563.