Mathematical Biology and Bioinformatics
2018. V. 13. № S. P. t1-t16. doi: 10.17537/2018.13.t1

Translation of the original article
Nazipova N.N., Isaev E.A., Kornilov V.V. et al. Mathematical Biology and Bioinformatics. 2017; 12(1):102–119. doi: 10.17537/2017.12.102

========== TRANSLATIONS OF PUBLISHED ARTICLES ==========

UDC: 004.9:004.9:004.8:577.21

Big Data in Bioinformatics

Nazipova N.N.[1], Isaev E.A.[2], Kornilov V.V.[2], Pervukhin D.V.[2], Morozova A.A.[3], Gorbunov A.A.[2], Ustinin M.N.[1]

[1]*Institute of Mathematical Problems of Biology RAS - the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences*
[2]*National Research University "Higher School of Economics"*
[3]*The Union of Enterprises the Central Scientific and Production Association "CASCADE"*

Abstract. Sequencing of the human genome began in 1994. Revealing of a human DNA draft took 10 years of collaborative work of many research groups from different countries. Modern technologies allow for sequencing a whole genome in a few days. We discuss here the advances in modern bioinformatics related to the emergence of high-performance sequencing platforms, which not only contributed to the expansion of capabilities of biology and related sciences, but also gave rise to the phenomenon of Big Data in biology. The necessity for development of new technologies and methods for organization of storage, management, analysis and visualization of big data is substantiated. Modern bioinformatics is facing not only the problem of processing enormous volumes of heterogeneous data, but also a variety of methods of interpretation and presentation of the results, the simultaneous existence of various software tools and data formats. The ways of solving the arising challenges are discussed, in particular by using experiences from other areas of modern life, such as web and business intelligence. The former is the area of scientific research and development that explores the impact and makes use of artificial intelligence and information technology (IT) for new products, services and frameworks that are empowered by the World Wide Web; the latter is the domain of IT, which addresses the issues of decision-making. New database management systems, other than relational ones, will help to solve the problem of storing huge data and providing an acceptable timescale for performing search queries. New programming technologies, such as generic programming and visual programming, are designed to solve the problem of the diversity of genomic data formats and to provide the ability to quickly create one's own scripts for data processing.

Key words: *Big Data, NGS, genome sequencing, IT technologies, bioinformatics, generic programming, visual programming, nonrelational databases, NoSQL systems, Hadoop, MapReduce.*

INTRODUCTION

Currently, the concept of Big Data has become common. Although there is still difference of opinion regarding the strict definition of the term [1, 2], under big data, we understand information of a huge volume and of a diverse composition, which is often updated and located in different sources, as well as special technologies for storage, transfer, processing and analysis of this information. This understanding has not only firmly entered into the lexicon of information technology specialists, but also evolved from a fashionable

technological trend into a concept that includes approaches, technologies and techniques that are actively used in the most diverse areas of our society. Moreover, the notion of big data, as a separate technology of collection and treatment for huge data sets, has disappeared from the analytical report “Hype Cycle for Emerging Technologies, 2015” of the Gartner Inc, where the analytics of market excitement, maturity and of benefit of more than 2000 new technological solutions is given in a graphical form [3]. The company explained its decision by pointing out that the concept of "big data" includes a large number of actively used technologies, which are parts of other popular areas and trends and have become day-to-day working tools. The main task of working with such data to date is to extract valuable knowledge from them. The greatest successes have been achieved by the business sectors, by closely interacting with the consumer and, accordingly, by being able to get the most benefit from the correct analysis and prediction of the behaviour of potential customers. This, above all, applies to banks, telecommunications, retail, energy and utilities. Now we are talking about the competent use of large amounts of data by companies in their business processes of storage and processing, and their ability to make them useful to the business. Large data tools enable organizations to manage resources more efficiently, to anticipate events that can affect their business, and make informed decisions faster.

Informatics responded to the revolutionary changes in the social life through the emergence of new scientific disciplines, the most actively developing among which are Web analytics and analysis of business data. Advanced solutions are available in these research areas already.

Web Intelligence (WI) is an area of research and development that studies the role and practical consequences of applying of artificial intelligence (knowledge representation, planning and organization of knowledge discovery, data mining, use of intelligent agents) and advanced information technologies (wireless networks, e-globalization devices, social networks, the World Wide Wisdom Web (W4), and data and knowledge networks (grids)) to the next generation of products, systems, services and activities for the World Wide Web.

Business intelligence (BI) includes technological tools for the collection, processing and analysis of business information designed to help corporate executives, business managers and other end users in business rule based decision making. Business analysis includes a wide range of tools, applications and techniques that allow organizations to collect data from internal systems and external sources, prepare them for analysis, design and execution of data requests, and to create reports, dashboards and other ways to visualize data to make the analytical results available to decision makers, as well as to decision executives.

However, new trends based on the concept of large data are coming onto the peak of popularity. One of such trends is the Internet of Things (IoT) [4]. This term implies the revolutionary transformation of the modern Internet, when a lot of devices become "smart", able to collect, analyze information and exchange it over telecommunications networks, both with people and with each other. A high value is placed here on machine learning, the means of searching for rules and links in very large volumes of information; data mining; advanced means of visualization and self-analysis of data; decision support systems and artificial intelligence; system of recognition of natural languages, etc.

However, all of the foregoing refers mainly to the industrial sphere of our society. At the same time, keeping in mind the onset of the "big data" term in 2008, we can remember that initially it was primarily concerned with the scientific sphere and, to a large extent, with bioinformatics.

The term "bioinformatics" was first used in 1970 by B. Hesper and P. Hogeweg in an article published in Dutch that is not generally accessible [5]. There it was defined as "the study of information processes in biotic systems". The authors considered the management of information in various forms, for example, the accumulation of information in the process of evolution, information transmission from DNA to intra- and intercellular processes, the interpretation of information at various levels of life as the defining property of life. The

authors claim that they came up with this term in order to separate bioinformatics as a research field in addition to biophysics and biochemistry.

Modern bioinformatics is a science that develops the use of computer methods for the analysis of a variety of genomic data. A huge role in the development of bioinformatics was played by the rapid development of computer technology and computational methods of data processing, and the emergence of modern telecommunications technologies. Bioinformatics is one of the science areas that are more dependent on the Internet and can only be successfully developed through the Internet. The very important for biology and medicine political decision, about the open accessibility of the most complex biological text – the human genome – has made this valuable source of knowledge accessible to scientists around the world and has enabled the formation of bioinformatics as a collective science, in which the achievements of separated teams are immediately made available to the entire scientific community, and where it is customary to freely distribute developed software and data.

When botanist H. Winkler proposed, in 1920, the term "genome" for the designation of a set of chromosomes, he obviously did not suspect that he was setting a fashion trend for the generation of more and more scientific words ending in "-ome" [6]. At that time, the concepts of biome (the set of living beings) and rhizome (the root system) already existed, but now scientists have thousands of different "-omes" [7]. Many of these terms are based on the Greek suffix "-ome", which means roughly "having the nature of". Simultaneous development of computer capacities and new technologies for obtaining data in various disciplines of biology related to the study of genomes led to the emergence of various disciplines called "-omics" in bioinformatics; these disciplines analyse all the organism entities (DNA, RNA, proteins, metabolites, etc.) in their structural-activity relations. Genomics, metagenomics, transcriptomics, proteomics, metabolomics, interactomics and other areas of bioinformatics are engaged in the study of genomes, metagenomes, transcriptomes, proteomes, metabolomes, interactomes and other collection of objects [8].

Each of the bioinformatics disciplines has its own objects for studying and own technologies for obtaining data. But they all generate huge amounts of data in different formats and at different levels that need to be stored, systematized, comprehended and visualized in order to deepen existing knowledge and stimulate discoveries.

MODERN BIOINFORMATICS: PROBLEMS AND SOLUTIONS

Historically, genomics is one of the first -omics of bioinformatics, which is engaged in the study of the genome. Genome is defined as the genetic substance of all the chromosomes of a living organism. Genomics deals with the structure, functioning, evolution, mapping and annotation of genomes. Genome annotation is a description of the functional and structural characteristics of the genome: determination of location and ascribing function to functional (genes, coding regions, promoter and regulatory sites, transposable elements, etc.) and structural elements (repeats, homopolymer tracts, etc.), the features of the functioning of the genome, the relationships between genes and other functional properties of the genome. Unlike the genetics, that deals with the role of genes in heredity, genomics studies gene structure and mechanisms of gene functioning. The main method of obtaining data is the sequencing of genomic DNA. Until 2001, when, as a result of international scientific cooperation, the human genome was first "read" and published [9, 10], the sequencing was carried out by the Sanger method, which primarily gave birth to genomics. New sequencing technologies (Next-Generation Sequencing, NGS, or high-throughput sequencing) that appeared at the end of the first decade of the 21st century dramatically reduced the cost of one genome processing from $100,000,000 in 2001 to $10,000 in 2011. One may say that they gave birth to metagenomics. Metagenomics studies the genetic material of a historically formed set of living species, united by a common area of distribution, called the biota.

The result of the "Human Genome" project was one sequence of more than three billion nucleotides contained in the haploid human genome. However, genomic sequences slightly

differ from one person to another, so art of sequence deciphering could become really profitable only when the methods allowing a researcher to read several copies of a genome simultaneously were developed. Several dozen techniques are now available; each of them has its advantages, limitations and disadvantages. However, all of them are cheap and fast enough for use not only in specialized laboratories, but also in medical clinics. With their introduction, a quick determination of the extended genomic sequences of an individual became available, for example, in order to identify gene mutations that could lead to the development of various diseases. New high-performance sequencing technologies allow for the *de novo* sequencing and resequencing of a genome (DNA-Seq), studying of the whole body transcriptome (RNA-Seq) and single-cell transcriptome (scRNA-Seq), of the nature of the DNA-protein interactions (ChIP-Seq), epigenomes, metagenomes, etc. [8]. A single sequencer can provide a great variety of data for various –omics, while the procedures of their generation differ only in the methods (protocols) for preparing samples for sequencing libraries.

Sequencing machines (sequencers) allow for processing of millions of fragments of nucleic acids in parallel, repeating operations of the same type for a hundred times, and each of these operations provides large data sets for analysis. For their processing, each sequencer is equipped with powerful server equipment, which helps to solve problems of proper reading of up to hundreds billions of nucleotides per hour, as a result of which biologists obtain fragments of a given length (reads). After this, researchers are faced with the task of assembling and annotating the genomic sequence. The different platforms and different manufacturers of sequencing machines are available [11]. The main platforms of the second generation NGS are the products of Illumina Inc., Thermo Fisher Scientific, Roche and Pacific Biosciences. They are all formally related to the discovery of polymerase chain reaction and automation of the main stages of DNA reading and are based on parallelizing the process of DNA reading. Several fragments of genomes can be determined with one run of the sequencer. Each of these technologies has its limitations in length and number of reads, in its price, in the availability of software and other parameters, but all of them are being actively developed in order to provide fast and cheap sequencing of any genomic data. The volumes of these data are enormous; they must first be stored, analyzed and presented in a form useful to biologists. A great variety of sequencing machines, each of which has its own application domain, forces biologists to become experts in a large number of methods, individual computer programs created in-house and through collaborative development by community members, as well as in a variety of data formats and databases. The revolutionary development of sequencing technologies poses new challenges for software developers and IT-specialists.

Software

A great many programs for work with NGS big data are available now [12]. Most of them are commercial products, the most famous and widely used among them are BaseSpace (Illumina, Inc.) [13], CLCBio [14], Lasergene (DNASTAR, Inc.) [15] and Geneious Basic [16]. Galaxy [17, 18], Globus Genomics [19], PATRIC [20], and UGENE [21, 22] are popular among free and hybrid (partially commercial, partially free) software tools. They have the broadest capabilities, most of these software packages can be installed on a local computer or server, and they work under different operating systems. However, despite the huge assortment of NGS software released in recent years, there is still no balance between the requirements of users and the capabilities that the tools offer.

Modern software solutions for analysis of the NGS data deluge should be provided with a convenient development environment. The time of biologists not concerned with IT-technologies is a thing of the past. When a researcher cannot quickly create an own innovative application can slow down and complicate the analysis of data. Modern tasks of genomic data analysis require an above average level of computer training. Researchers involved in large

projects, performed by several laboratories, need completely new interactive tools actively using high-performance computing infrastructure for the analysis of the NGS data sets. In the situation, when amount of more and more sophisticated data grows continuously, the future will belong to tools of online analysis and storage facilities that ensure the collaborative work of researchers, to methods that provide a high degree of interactive data analysis and visualization. Examples of software for managing scientific workflows designed specifically for work with large scientific data and more or less used for bioinformatics problems are KNIME [23], Pipeline Pilot [24], and TAVERNA [25].

Workflow management system is software used to accumulate performance parameters and to monitor a certain chain of tasks arranged in form of a working process. The system is a part of the infrastructure for launching, executing and monitoring scientific workflows. Such products have replaced software libraries, equipped with command line option, which were suitable only for software developers and trained users, and serve now as the basis for development of high-level software packages with an advanced graphical user interface (GUI) including menus and premade workflows or a pipelined data analysis.

Availability of an easily learned and understandable interface is a key aspect of the future development in NGS research, as it can ensure successful work of users without the knowledge of the basics of programming. The programs in large high-level software packages are carefully designed so that users do not make mistakes caused by the lack of special knowledge. In addition to the GUI, these packages provide a possibility to use single functions and premade workflows, as well as visual workflow designers. Visual designers are different from GUIs because they have their own graphical interface, which allow a user to combine existing program functions into new data processing pipelines, not available in pull-down menus or icons of a GUI.

When a user is forced to use a high-performance computing infrastructures to carry out projects related to, for example, a large number of human genomes, he must be able to use the software tool already familiar to him, with which he already dealt, but not learn a new one. An example of the correct design of application software is a software system that provides unchanging program syntax for all possible configurations (desktop, server, or distributed environment).

Despite the abundance of programs for NGS, there is a lack of opportunities for low-level programming to work with specific data structures, for example, with de Bruijn graphs used in genome assembly tasks or for performing specific functions, for example, using the Burrows-Wheeler transform for data compression. These features are not available in C++ or Java. BAMTools [26], htsliv (SAMtools / bcftools) [27], NGS++ [28], Bioclojure [29], and libStatGen [30] use standard data formats and provide little opportunity for using specific data structures and developing new algorithms needed for the analysis of NGS data. Despite the availability of developments in the technology of generic programming [31], their application to NGS is problematic due to a giant leap in the data volumes. This is also true for extensions of the open source programming languages Bioperl [32], BioRuby [33], BioJava [34], and Biopython [35], which were created due to the efforts of the Open Bioinformatics Foundation [36] for large data processing packages such as Bioconductor [37, 38].

Generic programming is a programming paradigm that consists of such a description of data and algorithms that can be applied to different types of data without changing the description itself. In one form or another, it is supported by different programming languages. The notions of generic programming first appeared in the 1970s in the form of generalized functions in the languages CLU and Ada, then in the form of parametric polymorphism in ML and its descendants, and then in many object-oriented languages such as C++, Java, Object Pascal, D, Eiffel, languages for the .NET platform and others.

Generic programming is considered as a programming methodology based on the separation of data structures and algorithms through the use of abstract descriptions of requirements. Abstract descriptions of requirements are an extension of the concept of an

abstract data type. Instead of describing a single type in generalized programming, a family of types sharing a common interface and *semantic behavior* is used. A set of requirements that describes the interface and semantic behavior is called a *concept*. The algorithm written in the generalized style can be applied to any type that satisfies its concepts, which is called polymorphism. In C++, object-oriented programming is implemented through virtual functions and inheritance, in generic programming – using class templates and functions.

To provide for all the needs of researchers capable of low-level software development and creating high-level programs for the enhancement of capabilities for analyzing NGS data, a visual programming can be applied. Visual programming is not a new concept [39]; since the early 1960s, it has been the subject of a philosophical discussion. This is the way to create a computer program by manipulating graphic objects instead of writing text. Visual programming is often presented as the next stage in the development of text-based programming languages. Recently, greater focus has been placed on visual programming due to development of the mobile sensor devices. Visual programming is mainly used to create programs with a graphical user interface. The visual programming environment allows for creation of web applications and console applications.

Data Formats

Heterogeneity of biological data is a big problem. Each manufacturer of a new device develops its own data format, making the task of data unification more and more difficult. This proves that people working with various devices should have good programming skills to have a possibility to modify existing scripts or create new ones to parse the data and convert one format to another.

In the era of big data, the usual conventional formats of data storage are changing. For data exchange, basic data formats such as PDB for spatial structures of proteins, FASTA for nucleotide and amino acid sequences have been developed and for a long time considered to be classical. Now there are new data formats. The PDB database format can serve as an example [40]. This format, appearing first in the 1970s, has long been used to store and exchange data on the structures of small proteins. However, the PDB format cannot be used for large complexes that consist of thousands amino acids, so now the PDBx/mmCIF format was introduced, which combined the PDB format and the mmCIF crystallographic data storage format [41] and officially replaced the PDB format in 2014. Another format for large scale structures that has been recently proposed is the MMTF (MacroMolecular Transmission Format) is a compact binary format for storage and transmission of large structural data for faster visualization and analysis [42]. The binary format enables extraordinary compaction of the data, allowing the entire PDB archive to be stored in less than 7GB.

Bioinformatics has always been associated with a large number of databases, which store a variety of genomic data. The most comprehensive NAR online Molecular Biology Database Collection currently contains about 1900 resource descriptions published in the annual Database issue by Nucleic Acids Research (NAR). The collection is divided into 15 subject categories subdivided further into 41 subcategories [43]. This systematization is largely nominal, because it is considered a good practice to make information resources polythematic, to organize cross-references between all other databases where relevant information is contained. The collection has been going for almost a quarter of a century, it has a "golden" core of 105 resources that have existed for a long time and are constantly being updated. The largest databases, which contain voluminous experimental data, such as nucleotide and protein sequences, data on the structures of biological macromolecules, crystallographic data, etc., are replenished mainly by the experimenters using the electronic submission system. This is due to the requirements of the journals that publish results of this kind. The authors, before submitting an article reporting a sequencing of a biomolecule structure, should deposit their data to a public resource, making them available to the entire scientific community.

Storage and exchange of data

The problem of databases completeness is solved by creating database consortia established to store sequences of DNA and protein. Nucleotide sequences are uploaded into one of the three databases GenBank [44], ENA [45] or DDBJ [46] either by the authors or by the sequencing centres. Between these three databases, daily data exchange is carried out, so that daily updates on the NCBI servers, where GenBank is stored, include the latest available sequence data from all three sources.

However, in the new era of big data, conventional database management systems based on the relational principle do not correspond anymore to large amounts of data, a variety of formats, the need to share data originated in various spots all over the world and to a variety of search queries [**Ошибка! Источник ссылки не найден.**]. Relational organization of data storage assumes the availability of pre-defined search fields and logical structure of requests. To store data in a relational schema, tables are constructed for each search field, in which the field values are written. When resolving queries, temporary tables are built, which, with huge amounts of data, makes the work inefficient, and in some cases, impossible. Therefore, key Internet players such as Amazon, Inc. and Google, Inc. in the early 2000th began the development of new database management systems.

One such solution is NoSQL ("Not Only SQL"), a class of non-relational database management systems (DBMS) designed to work with big data [47]. NoSQL systems provide a fast response time to search queries with a high throughput of processing the flow of requests. These systems can be divided into two groups, according to the type of storage organization.

The first group uses the logical principle of "key-value". It is, in fact, an associative table; each key has a unique value. The second group of systems is document-oriented. This is not a fully systematized storage; the tables are not used here.

Other non-relational data management systems are graph databases that use the structure of graphs with nodes, edges, and node properties for semantic querying to represent and store data. The key concept of the system is the graph (or edge, or relationship) that is directly related to the data elements. Links allow for connecting data to each other directly, and in many cases, for retrieving with a single operation.

This distinguishes graph systems from traditional relational databases, where data interconnections are implemented using tables, and complex search queries are resolved by combining tables that satisfy elementary queries. Graph databases provide a simple and quick extraction of complex hierarchical structures that are difficult to find in relational systems [48]. The principle of storing graph databases is constantly changing. Some systems use elements of a relational organization, that is, they store graphs in the form of tables, while others use the key-value principle or document-oriented concept for data storage, which makes them essentially NoSQL structures.

Extracting data from a graph database requires a special query language other than SQL, which was designed for relational databases and is not able to traverse the graph elegantly. None of the query languages became universal and conventional, like SQL was for relational databases; there is a wide variety of systems that are tightly bound to a particular product. However, some standardization efforts were implemented, and it led to the appearance of query languages such as Cypher, which can become standard [49, 50]. In addition to the existence of query languages, some graph databases are available through API.

Another example of a nonrelational storage model is HBase [51]. It is designed for work with the file system of the distributed operating system Hadoop (HDFS, Hadoop Distributed File System). The Hadoop system was specifically designed for work with large data. A typical file system consists of a table of file descriptors and a data area. In HDFS, the name server (NameNode) is used instead of the table, and the data is distributed to data servers (DataNodes). The information about the machines on which the data blocks are located allows for running the same computing processes on them and for performing most of the

calculations locally, i.e. without data transmission over the network. Just this idea underlies the paradigm of MapReduce and its specific implementation in Hadoop. The classic Hadoop cluster configuration consists of one name server, one MapReduce wizard (the so-called JobTracker), and a set of working machines, each running a DataNode and a TaskTracker. Each MapReduce work consists of two stages, separated by data transfer between nodes:

– The map phase is executed in parallel and (if possible) locally over each data block. Instead of delivering terabytes of data to the program, a small, user-defined program is copied to the data server and does to them all the operation except for those requiring shuffling or moving.

– The reduce phase completes the map phase with aggregating operations.

Hbase is a distributed, column-oriented, multi-version "key-value" database, modelled after BigTable [52] developed by Google. The data are organized into lines indexed by the primary key, referred as RowKey in HBase. For each RowKey key, an unlimited set of attributes (or columns) can be stored. Columns are organized into groups of columns, called Column Family. When columns share the same patterns of use and are stored together, they are united into one Column Family, when they. For each attribute, several different versions can be stored. Different versions differ in Timestamp. The records are physically stored in the order of sorted RowKey values. In this case, the data corresponding to different elements of the Column Family type are stored separately, which allows for reading data only from the desired family of columns if necessary. Attributes that belong to the same column group and correspond to the same key are physically stored as a sorted list. Any attribute of any key is not obligatory, the absence of an attribute does not cause overhead to store empty values. The four-dimensional model of HBase data can be formulated as a key-value relationship of the following kind:

<table, RowKey, Column Family, Column, Timestamp> $\Rightarrow$ Value.

Based on the results of studies on the HBase applicability in bioinformatics for NGS data [50], it was recognized that the scalability and reliability of the data-oriented HBase is large enough. It has also been shown that this architecture allows for the rapid integration and analysis of large and heterogeneous data, using for their storage a small number of tables.

Visualization

New NGS technologies of data obtaining in biology open new horizons in the formulation of novel ideas and concepts for researchers, but big data are hard for analysis and visualization. Visualization plays a key role in the detection of new patterns and trends; the lack of specialized representation tools is the limiting factor for data interpretation.

Since data visualization forms the basis in the interpretation of sequencing data, many software developers are engaged in creating software tools for visual analysis of data. These developments are more specialized than other NGS data analysis packages; each of them has its own visualization object and its range of applicability [47]. Some of them (ngs.plot [53] and Integrative Genomics Viewer [54]) allow for the integration of heterogeneous data, such as gene annotations, clinical information and phenotypic data, from different sources. Girafe [55] can be used to visualize the process of reads aligning with genomic fragments; it is user-friendly because it works together with the R/Bioconductor package [38]. Bioconductor is a large-scale project with open source software that provides many separate packages for bioinformatics research. It uses the programming language R, which is cross-platform (supports Linux, most UNIX-like systems, Mac OS X and Windows). Despite its ability to work with dynamic graphs [56] on web pages, these solutions are not sufficient to manage dynamic graphs for big data and sufficient interactivity.

The main feature of huge data visualization is the need to show many millions of points at the screens of monitors with a limited number of pixels. An important technical problem is need to interact dynamically with the graphs, for example, to change the type of graph, to

zoom in and zoom out, to view and change parameters and instantly get a new picture. In addition, the revolution in the genomics technologies and large amount of available data increases importance of the teamwork. For NGS analysis, it is usually necessary to have close on line cooperation of scientists from different teams and different locations. This collaboration is provided by the creation of web applications, which should support the increasing potential of the data research intensity. This calls for the use of web technologies for data exchange, analysis tools and results through web applications that are accessible from the Internet. However, web applications have some technical limitations related to the capabilities of modern web browsers. Numerous problems arise when creating web applications that include interactive visualization tools for Big Data analysis. For example, web browsers cannot support huge interactive graphics and tables with thousands data fragments. Thus, developers of web applications for the analysis and visualization of NGS data should create solutions supporting large amounts of data and increasing demands of scientists in interactivity of their data combined with the limitations and diversity of web browsers.

IT companies have ready-made innovative solutions for visualizing large data developed for Business Intelligence applications. In last decades, these products have become significantly more powerful and less expensive. Examples include Tibco Spotfire [57] or SAS [58] packages, both of which are successfully used now in life sciences and can greatly help to visualize investigation of NGS data. The main advantage of these solutions is that they offer powerful visualization with numerous types of graphs for data representation and provide a high level of interactivity for changing image parameters or scaling.

METHODS OF ORGANIZING CALCULATIONS WITH LARGE VOLUMES OF DATA

In modern bioinformatics, a large scale analysis of the genomes of systems of different complexity, from microorganisms to humans, has become possible. Accumulated data contain extremely important new information about yet unknown mechanisms of functioning of these codes. However, the amount of generated data is so huge (for example, storing of one genomic sequence on a computer storage medium will require hundreds of gigabytes), that not only the storage or the transmission of these data becomes a problem, but even greater difficulties arise at data processing. The very assembly of the genome from billions of sequenced reads is a difficult task, but a solution of applied problems of bioinformatics using genomic sequences requires the application of the latest developments of distributed and parallel programming.

Thus, as far as gathering of huge volumes of genomic information is concerned, more effective, accurate and specific methods for performing analysis of this heterogeneous information, based on modern methods of analyzing large amounts of data, become critical. To create them, one can borrow IT solutions that established for big data management and analysis in such areas as networked artificial intelligence and business information and analytics.

First of all, these are mathematical and statistical methods of analysis and data processing and information retrieving algorithms applicable to huge data sets. These methods include advanced Data Mining tools (for example, cluster and regression analysis), natural language processing techniques (including tonal analysis), predictive analytics, statistical analysis algorithms (such as A/B-testing and time series analysis), machine learning algorithms, and others.

Secondly, these are instrumental and software technical means of information technologies, which allow for storing and processing of extremely large amounts of data. The main way to solve these problems is to organize distributed computing using a large number of computational nodes, most often combined into a parallel computing system [59]. The most well-known implementations of the distributed computing model in large parallel

clusters of computing nodes include the MapReduce programming platform [60], proposed by Google, and the freely distributed Hadoop framework [61], created and maintained by the Apache Software Foundation and designed to develop and support the execution of distributed computing programs. Hadoop is based on the implementation of the MapReduce model for the distributed file system (HDFS), which is designed to store large files distributed between the nodes of the computational cluster. Hadoop infrastructure also includes big data processing software applications [62], such as Apache Pig, Apache Hive and Apach Spark.

Apache Pig is a platform for analyzing big data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets. The Apache Hive data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. Apache Spark is a software framework for large-scale distributed processing of unstructured and weakly structured data.

In addition, there are a large number of commercial implementations of software systems based on the technology of different classes, capabilities and destinations offered by large vendors, such as IBM or Microsoft, and relatively small companies such as Cloudera, Hortonworks, MapR, and others.

The third component of the technological toolware for processing of big data is represented by specialized high-performance hardware/software complexes. These products use the paradigm of in-memory analytics, this is a technology that maximize the use of RAM for running processing systems, such as IBM's recently announced mainframe z13s [63]. This supercomputer, focused on solving problems in the area of big data, has up to 20 eight-core, 5.0 GHz dedicated processors, super-fast memory with a capacity of up to 10 TB, supports up to 8000 virtual servers and has a design maximum performance of more than 111,000 MIPS (millions of operations per second). Such specialized supercomputers are also offered by other suppliers of solutions such as Teradata, Oracle, SAP, SAS and others.

Thus, today we have a fairly advanced set of methods for big data analytics, allowing for extraction of the necessary knowledge. However, as it was mentioned before, the overwhelming number of technologies and methods of work with big data is implemented and successfully used almost exclusively in business sectors.

CONCLUSION

The revolutionary changes that high-throughput sequencing (NGS) caused in the biological sciences contribute to the high-speed emergence of new data. Obviously, big data have a variety of formats and are of great interest for different groups of researchers. There is no general consensus yet as to what data can be considered big, it is generally believed that these are collections of data that are too large to be managed and analyzed using traditional approaches. According to this point of view, the scale and method of big data obtaining are specific to each research area. Data that are suitable for this definition in biology and medicine are generated from numerous sources, including laboratory experiments, and are accessible through online databases. Medico-biological big data are the result of fusion of small data sources.

For example, most scientists make the laboratory experiments to study gene expression public by depositing them in the ArrayExpress database [64], which is one of three collections of results of costly experiments, along with Omnibus in NCBI (USA) and DDBJ Omics Archive (Japan). These databases collectively store processed data and metadata describing properties of the samples and the technical details of experiments, including raw data and protocols of experiments. The raw DNA sequences are placed in the European Nucleotide Database (ENA) and immediately access the data collections associated with this resource in GenBank and DDBJ.

These common big data are profitable, since the cost of obtaining data is shared by many laboratories, and could be used by specially developed computing methods. Medico-biological big data make it possible to develop predictions based on evidence that complement hypotheses based on previous knowledge. Since these data are supplied to the collection by various research teams from various institutions, and the systems under examination are diverse, the discoveries will be more likely to be generalized.

Big data not only open up new opportunities, but also set new tasks. It becomes necessary to adapt training programs on bioinformatics to new realities [65]. It is necessary to develop training programs that provide skills for the effective and confident use of big data and to critically evaluate the results.

Problems raised by big data require that training programs prepare students for solving problems, including combining data, overcoming computational difficulties and storage constraints, the skills of multiple testing of hypotheses, and working with biased and mixed data. Data consolidation encompasses the problem of obtaining the necessary data in the appropriate format and their normalization in order to make them comparable by sources. Computing constraints relate to the difficulties and costs associated with storing, moving, and analyzing data. Multiple hypothesis testing refers to the problem of detecting the statistical probability of parasitic associations in large sets of data. Discrepancies and inconsistency of the data correspond to problems related to what experiments were performed or what processes were most often analyzed.

This area of knowledge is developing rapidly, and the problems formulated here are not static, they are also changing rapidly. Scientists working in the field of bioinformatics in the era of big data should be able to understand the computing environment and ways of analyzing and obtaining of analytical conclusions from large-scale data most effectively in this environment. In addition, they must be well versed in the algorithms for assembling genomes, in order to choose the most suitable from the vast majority of existing.

Significant resources are allocated all over the world to prepare scientists for the analysis of large-scale data. The US government has allocated $200 million to finance Big Data and the NIH Big Data to Knowledge (BD2K) initiative [66]. The Big Data program in particular, aims to significantly improve the tools and methods for accessing, organizing and collecting data on discoveries associated with huge amounts of digital data.

The reality is that the use of familiar relational database management systems can no longer satisfy the increased data volume loads. The relational database cannot adapt to a large number of queries; the volumes of tables needed to implement this model grow too fast for a large amount of stored data. The relational model no longer matches the performance criterion, since this data model operates large amount of temporary tables in which intermediate results are stored. Other large database management systems are necessary to meet the increased demands of modern life.

A big problem is the heterogeneity of data formats and software tools for their processing. Each producer of sequencing machines develops, as a rule, its own data format, which makes unification of data an urgent task. This situation requires that biologists have a serious level of knowledge in programming languages, so that they can use existing scripts or create new ones to analyze data and extract useful knowledge. A lot of tools for converting and analyzing data are posted on the Internet. All of them are written in different programming languages and are designed for various computer platforms. The difficulty lies in understanding the level of coordination between different tools and the organization of their workflow, as well as in updating and maintaining software.

New distributed computing technologies are necessary. In the areas of business analytics, a number of solutions have already been created that could be applied to bioinformatics. These problems could be solved by a new generation of bioinformatics experts.

New methods of visualization are necessary. These methods will help a human mind to comprehend the data of various omics. Therefore, in the era of technology of quick and cheap

obtaining of any data, may be working methods should be changed. For example, a concept "plurality should not be posited" should become a rule, that is, a researcher should not accumulate raw data, but carefully plan research and visualize the results. Visualizing designers should be involved at a work process at the stage of its planning but not after the experimental data have already been obtained. In such a case, the data structure will be more thoughtful and optimal.

Reflections on the methods of visualization can lead to the development of alternative representations for the same data. This may entail the development of other approaches to the collection, organization and retrieval of data that contribute to the maximum meaningfulness of experimental data, which in turn stimulates intuition, leads to information interactions and valuable discoveries.

Our world has changed, our society has become information-driven, and fully dataflow-controlled; knowledge and skills are becoming the core values. Big data tools allow us to manage resources more efficiently, to anticipate future events, to make informed decisions faster.

In bioinformatics and in computational biology also, amount of data became too large to analyze them "in the old fashion," and the speed of their emergence is increasingly growing, and the complexity of their analysis is very high because of their specific structure and organization. At the same time, the application of big data technologies in bioinformatics, biomedicine and health care [67] can only improve, but radically and in a revolutionary manner change the situation in this area. However, despite some successes in the development of methods of analysis and in the practical application of new technologies for work with big data, bioinformatics and biomedicine have a huge untapped potential for their development.

Acknowledgements

We would like to thank Professor Ansuman Lahiri from University of Calcutta (Department of Biophysics, Molecular Biology & Bioinformatics) for his valuable comments and thorough work on translation of the manuscript into English.

The study was partially supported by RFBR grants 15-07-05783 (N.N.N), 16-07-00937 and 16-07-01000 (U.M.N) and the Program of Fundamental Scientific Research of the Presidium of the Russian Academy of Sciences I.33P. (U.M.N).

REFERENCES

1. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Byers A.H. *The Next Frontier for Innovation, Competition, and Productivity*. San Francisco: McKinsey Global Institute, 2011. URL: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (accessed 17.02.2017).
2. Jacobs A. The Pathologies of Big Data. *Communications of the ACM*. 2009. V. 52. No. 8. doi: 10.1145/1536616.1536632
3. What's New in Gartner's Hype Cycle for Emerging Technologies, 2015. *Gartner*. URL: http://www.gartner.com/smarterwithgartner/whats-new-in-gartners-hype-cycle-for-emerging-technologies-2015/ (accessed 17.02.2017).
4. Chui M., Löffler M., Roberts R. The Internet of Things. *McKinsey Quarterly*. 2010. URL: http://www.mckinsey.com/industries/high-tech/our-insights/the-internet-of-things (accessed 17.02.2017).
5. Hogeweg P. The Roots of Bioinformatics in Theoretical Biology. *PLOS Computational Biology*. 2011. V. 7. No. 3. Article No. e1002021.
6. Winkler H. *Verbreitung und Ursache der Parthenogenesis im Pflanzen - und Tierreiche*. Jena: Verlag Fischer, 1920.
7. Baker M. The 'Oms Puzzle. *Nature*. 2013. V. 494. P. 416–419.

8. Ohashi H., Hesegawa M., Wakimoto K., Miyamoto-Sato E. Next-generation technologies for multiomics approaches including interactome sequencing. *BioMed Research International*. 2015. V. 2015. Article No. 104209.
9. International Human Genome Sequencing Consortium. Human genome. *Nature*. 2001. V. 409. P. 860–921.
10. Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., et al. The sequence of the human genome. *Science*. 2001. V. 291. No. 5507. P. 1304–1351.
11. Buermans H.P.J., den Dunnen J.T. Next generation sequencing technology. Advances and applications. *BBA – Molecular Basis of Disease*. 2014. V. 1842. No. 10. P. 1932–1941.
12. Bioinforx Inc. *Next Generation Sequencing Software*. URL: http://bioinfo.wisc.edu/knowledge_base/next-gen-seq_software.php (accessed 17.02.2017).
13. *BaseSpace Sequence Hub*. URL: https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_basespace.pdf (accessed 17.02.2017).
14. *CLCBio*. URL: http://www.clcbio.com (accessed 17.02.2017).
15. *DNASTAR Lasergene*. URL: https://www.dnastar.com/t-allproducts.aspx (accessed 17.02.2017).
16. Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012. V. 28. No. 12. P. 1647–1649.
17. Giardine B., Riemer C., Hardison R.C., Burhans R., Elnitski L., Shah P., Zhang Y., Blankenberg D., Albert I., Taylor J., et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*. 2005. V. 15. No. 10. P. 1451–1455.
18. Goecks J., Nekrutenko A., Taylor J., Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010. V. 11. No. 8. Article No. R86.
19. Madduri R.K., Sulakhe D., Lacinski L., Liu B., Rodriguez A., Chard K., Dave U.J., Foster I.T. Experiences Building Globus Genomics: A Next-Generation Sequencing Analysis Service using Galaxy, Globus, and Amazon Web Services. *Concurr. Comput*. 2014. V. 26. No. 13. P. 2266–2279.
20. Wattam A.R., Abraham D., Dalay O., Disz T.L., Driscoll T., Gabbard J.L., Gillespie J.J., Gough R., Hix D., Kenyon R., et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*. 2014. V. 42. P. D581–D591.
21. Golosova O., Henderson R., Vaskin Y., Gabrielian A., Grekhov G., Nagarajan V., Oler A.J., Quinones M., Hurt D., Fursov M., Huyen Y. Unipro UGENE NGS pipelines and components for variant calling, RNA-seq and ChIP-seq data analyses. *PeerJ*. 2014. V. 2. Article No. e644.
22. Okonechnikov K., Golosova O., Fursov M. UGENE Team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*. 2012. V. 28. No. 8. P. 1166–1167.
23. Jagla B., Wiswedel B., Coppree J.-Y. Extending KNIME for next-generation sequencing data analysis. *Bioinformatics*. 2011. V. 27. No. 20. P. 2907–2909.
24. Warr W.A. Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of Computer-Aided Molecular Design*. 2012. V. 26. No. 7. P. 801–804.
25. Oinn T., Addis M., Ferris J., Marvin D., Senger M., Greenwood M., Carver T., Glover K., Pocock M.R., Wipat A., Li P. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*. 2004. V. 20. No. 17. P. 3045–3054.

26. Barnett D.W., Garrison E.K., Quinlan A.R., Stromberg M.P., Marth G.T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*. 2011. V. 27. No. 12. P. 1691–1692.
27. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009. V. 25. No. 16. P. 2078–2079.
28. Nordell Markovits A., Joly Beauparlant C., Toupin D., Wang S., Droit A., Gevry N. NGS++: a library for rapid prototyping of epigenomics software tools. *Bioinformatics*. 2013. V. 29. No. 15. P. 1893–1894.
29. Plieskatt J., Rinaldi G., Brindley P.J., Jia X., Potriquet J., Bethony J., Mulvenna J. Bioclojure: a functional library for the manipulation of biological sequences. *Bioinformatics*. 2014. V. 30. No. 17. P. 2537–2539.
30. *libStatGen*. URL: https://github.com/statgen/libStatGen/ (accessed 17.02.2017).
31. Pitt W.R., Williams M.A., Steven M., Sweeney B., Bleasby A.J., Moss D.S. The Bioinformatics Template Library – generic components for biocomputing. *Bioinformatics*. 2001. V. 17. No. 8. P. 729–737.
32. Stajich J.E., Block D., Boulez K., Brenner S.E., Chervitz S.A., Dagdigian C., Fuellen G., Gilbert J.G., Korf I., Lapp H., et al. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*. 2002. V. 12. No. 10. P. 1611–1618.
33. Goto N., Prins P., Nakao M., Bonnal R., Aerts J., Katayama T. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics*. 2010. V. 26. No. 20. P. 2617–269.
34. Holland R.C., Down T.A., Pocock M., Prlic A., Huen D., James K., Foisy S., Drager A., Yates A., Heuer M., et al. BioJava: an open-source framework for bioinformatics. *Bioinformatics*. 2008. V. 24. No. 18. P. 2096–2097.
35. Cock P.J., Antao T., Chang J.T., Chapman B.A., Cox C.J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009. V. 25. No. 11. P. 1422–1423.
36. *Open Bioinformatics Foundation*. URL: https://www.open-bio.org/wiki/Main_Page (accessed 17.02.2017).
37. Huber W., Carey V.J., Gentleman R., Anders S., Carlson M., Carvalho B.S., Bravo H.C., Davis S., Gatto L., Girke T., et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*. 2015. V. 12. No. 2. P. 115–121.
38. Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004. V. 5. No. 10. Article No. R80.
39. Milicchio F., Rose R., Bian J., Min J., Prosperi M. Visual programming for next-generation data analytics. *BioData Mining*. 2016. V. 9. Article No. 16.
40. Bernstein F.C., Koetzle T.F., Williams G.J., Meyer E.F.Jr., Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T., Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*. 1977. V. 112. No. 3. P. 535–542.
41. Bourne P. E., Berman H.M., McMahon B., Watenpaugh K.D., Westbrook J.D., Fitzgerald P.M.D. Macromolecular crystallographic information file. *Methods in Enzymology*. 1997. V. 277. P. 571–590.
42. Bradley A.R., Rose A.S., Pavelka A., Valasatava Y., Duarte J.M., Prlić A., Rose P.W. MMTF – an efficient file format for the transmission, visualization, and analysis of macromolecular structures. *PLOS Computational Biology*. 2017. V. 13. No. 6. Article No. e1005575. doi: 10.1371/journal.pcbi.1005575

43. Galperin M.Y., Fernández-Suárez X.M., Rigden D.J. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. *Nucleic Acids Res*. 2017. V. 45. P. D1–D11.
44. Benson D., Lipman D.J., Ostell J. GenBank. *Nucleic Acids Res.* 1994. V. 22. P. 3441–3444.
45. Rice C.M., Fuchs R., Higgins D.G., Stoehr P.J., Cameron G.N. The EMBL Data Library. *Nucleic Acids Res.* 1993. V. 21. P. 2967–2971.
46. Tateno Y., Gojobori T. DNA Data Bank of Japan in the age of information biology. *Nucleic Acids Res*. 1997. V. 25. No. 1. P. 14–17.
47. de Brevern A.G., Meyniel J.-P., Fairhead C., Neuvéglise C., Malpertuy A. Trends in IT Innovation to Build a Next Generation Bioinformatics Solution to Manage and Analyse Biological Big Data Produced by NGS Technologies. *BioMed Research International*. 2015. V. 2015. Article No. 904541.
48. Lith A., Mattsson J. *Investigating Storage Solutions for Large Data. A comparison of well performing and scalable data storage solutions for real time extraction and batch insertion of data*: Master of Science Thesis. 2010. URL: http://publications.lib.chalmers.se/records/fulltext/123839.pdf (accessed 17.02.2017).
49. Svensson J. Relational vs. graph databases: Which to use and when? *SD Times*. 2016. URL: http://sdtimes.com/guest-view-relational-vs-graph-databases-use/#sthash.yHI6aoDv.dpuf (accessed 17.02.2017).
50. Have C.T., Jensen L.J. Are graph databases ready for bioinformatics? *Bioinformatics*. 2013. V. 29. No. 24. P. 3107–3108.
51. Taylor R.C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics*. 2010. V. 11. Article No. S1.
52. Chang F., Dean J., Ghemawat S., Hsieh W.C., Wallach D.A., Burrows M., Chandra T., Fikes A., Gruber R.E. Bigtable: A distributed storage system for structured data. In: *The 7th Symposium on Operating System Design and Implementation Seattle*. WA: Usenix Association, 2006. 14 p. URL: https://static.googleusercontent.com/media/research.google.com/ru//archive/bigtable-osdi06.pdf (accessed 17.02.2017).
53. Shen L., Shao N., Liu X., Nestler E. Ngs.plot: quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*. 2014. V. 15. No. 1. Article No. 284.
54. Robinson J.T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E.S., Getz G., Mesirov J.P. Integrative genomics viewer. *Nature Biotechnology*. 2011. V. 29. No. 1. P. 24–26.
55. Toedling J., Ciaudo C., Voinnet O., Heard E., Barillot E. Girafe – an R/Bioconductor package for functional exploration of aligned next-generation sequencing reads. *Bioinformatics*. 2010. V. 26. No. 22. P. 2902–2903.
56. Nolan D., Lang D.T. Interactive and animated scalable vector graphics and R data displays. *Journal of Statistical Software*. 2012. V. 46. No. 1. P. 1–88.
57. *TIBCO Spotfire Homepage*. URL: http://spotfire.tibco.com/ (accessed 17.02.2017).
58. Wexler J., Thompson W., Aponte K. Time Is Precious, So Are Your Models. SAS provides solutions to streamline deployment. In: *SAS Global Forum 2013*. Paper No. 086-2013. URL: https://support.sas.com/resources/papers/proceedings13/086-2013.pdf (accessed 17.02.2017).
59. Tanenbaum A.S., van Steen M. *Distributed Systems. Principles and Paradigms*. Prentice-Hall Inc., 2002.
60. Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun. ACM*. 2008. V. 51. No. 1. P. 107–113.
61. White T. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 2015. 756 p.

62. *The Apache Software Foundation Home page*. URL: http://www.apache.org/ (accessed 17.02.2017).
63. *IBM z Systems – z13s*. URL: http://www-03.ibm.com/systems/z/hardware/z13s.html/ (accessed 17.02.2017).
64. Rustici G., Kolesnikov N., Brandizi M., Burdett T., Dylag M., Emam I., Farne A., Hastings E., Ison J., Keays M., et al. ArrayExpress update – trends in database growth and links to data analysis tools. *Nucleic Acids Res*. 2013. V. 41. P. D987–D990.
65. Greene A.C., Giffin K.A., Greene C.S., Moore J.H. Adapting bioinformatics curricula for big data. *Briefings in Bioinformatics*. 2016. V. 17. No. 1. P. 43–50.
66. Margolis R., Derr L., Dunn M., Huerta M., Larkin J., Sheehan J., Guyer M., Green E.D. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J. Am. Med. Inform. Assoc*. 2014. V. 21. P. 957–958.
67. Luo J., Wu M., Gopukumar D., Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomed. Inform. Insights*. 2016. V. 8. P. 1–10.

Received 16 March 2018.
Published 03 April 2018.