Scalable framework for adaptive in-silico knowledge discovery and decision-making out of genomic big data

Desislava Ivanova, and Plamenka Borovska

Citation: AIP Conference Proceedings **2048**, 060019 (2018); doi: 10.1063/1.5082134
View online: https://doi.org/10.1063/1.5082134
View Table of Contents: http://aip.scitation.org/toc/apc/2048/1
Published by the American Institute of Physics

Scalable Framework for Adaptive In-silico Knowledge Discovery and Decision-Making out of Genomic Big Data

Desislava Ivanova[1, a)], Plamenka Borovska[2, b)]

[1, 2] Bulgaria, Sofia 1000, Bul. "Kliment Ohridski" 8, Technical University of Sofia,
[1] Faculty of Applied Mathematics and Informatics, Department of Informatics, bl. 2, office 2541
[2] Faculty of Applied Mathematics and Informatics, Department of Informatics, bl. 2, office 2209

[a)] d_ivanova@tu-sofia.bg
[b)] pborovska@tu-sofia.bg

Abstract. This paper presents the concept and the modern advances of big data analytics and its influence in the area of genomics for adaptive in-silico knowledge discovery and decision-making with respect to precision and personalized medicine. The goal of the paper is to build up the scalable framework, providing a set of software tools for applying the methods in research and experimental activities for precision medicine support, establishing a modern research infrastructure that will allow for significant scientific outcomes, development of new methods and algorithms to manage big data streams, deployment of new streaming and parallel processing technologies of large sets of scientific data obtained from experiments. The scalability of the working framework reduces computational time and support optimization by involving resource reconfiguration and parallel processing. The proposed scalable framework is verified for the case studies of Multiple Sequence Alignment (MSA) based on social behavior model, enhancer-promotor interactions and early detection of breast cancer. Finally, some conclusions and future work are summarized.

INTRODUCTION

During the last years leading scientists, researchers and analysts point the Big Data as a kind of revolution in scientific studies and one of the most prospective tendencies in the field of IT [1]. This gave an intensive push in the development of methods and technologies for processing of large amount of data recently and yielded to radical changes in the scientific research paradigms. Globally, the intensive progress of ICT yielded to accumulate and store of enormous files of data, normally with low level of information density, which became a main source of knowledge. The new scientific paradigm is called "data-intensive science" [2]. The radical change in the fourth paradigm supposes a new kind of performing the experiments and knowledge detection. In spite of planning in advance an experiment and analyzing later the data, in the new paradigm the accumulated enormous data files are subject of analyze.

The new scientific paradigm "Data-Intensive Scientific Discovery (DISD) [3] yielded to a revolution in scientific research and innovations, supposing the following phases: accumulation of data, "cleaning", integration and presenting of data, analyze of data, and taking a decision (data-intensive decision making). The fourth paradigm imposes a set of challenges towards the processing technologies, summarized as follows: accumulation and storage of huge amount of data, searching, sharing, analyzing and visualization, necessity of high performance processing resources, parallel and distributed processing, parallel input/output, processing in memory. The scalability is the main obstacle when analyzing the data. The flow processing is also a big challenge in large data amount.

In nowadays the collection of data increases fast, whiles the development of methods and instruments for data management cannot follow the same pace. From this point of view it is extremely important to have effective metadata, and semantic techniques for structuring data files and content. The big challenge in this direction is the

Proceedings of the 44th International Conference on Applications of Mathematics in Engineering and Economics
AIP Conf. Proc. 2048, 060019-1–060019-7; https://doi.org/10.1063/1.5082134
Published by AIP Publishing. 978-0-7354-1774-8/$30.00

processing of combination of big data, collected in real time and data already accumulated. Finding decisions for these challenges is of great importance, especially for the fields' of scientific investigations, where enormous quantities of flow data are generated in addition to the available already big sets of real and historical data.

THE PROBLEM AREA

The fundamental scientific studies are in revolution era by the big files and flows of data. One of the fields of the fundamental science, strongly dependent from the development of big data, is the field of Genomics [4]. In the biological sciences there are very well established practices of collecting data in the public and generally accessible data bases, which are used by the scientists all over the world, working on concrete subjects. The development of the bioinformatics stimulates in high extent the methods for processing and analyses of collected data.

The next generation of methods for data analytics will have to manage huge amount of information from different kind of sources with differentiated characteristics, levels of trust, and frequency of updates. Data analytics will has to assure the information in economically effective and sustainable way. To do this, from one side it is necessary to create complex prognostic models and methods for heterogeneous and big data files. From other side, these models and methods have to be implemented in real time for big amount of flow data. This is a big challenge, because big data, besides the volume are strongly heterogeneous and dynamic, requiring high performance and scalable frameworks and platforms.

The goal of the paper is to build up scalable framework for adaptive in silico knowledge discovery and decision making based on big data streams analysis for scientific research, providing a set of software tools for applying the method in research and experimental activities for precision medicine support. The proposed scalable framework provides software tools for applying the methods for big data analytics in research and experimental activities for decision-making with respect to precision and personalized medicine, establishing a modern research infrastructure that will allow for significant scientific outcomes, development of new methods and algorithms to manage Big data streams, deployment of new streaming and parallel processing technologies of large sets of scientific data obtained from experiments.

PROPOSED SCALABLE FRAMEWORK FOR ADAPTIVE IN-SILICO KNOWLEDGE DISCOVERY AND DECISION-MAKING

The conceptual architecture of the scalable framework for adaptive in silico knowledge discovery and decision making based on big data comprises hardware and software resource reconfiguration and software will be developed utilizing streaming and parallel processing technologies, Figure 1. The architecture is compound of separate, independent components:

- ✓ Access to progressively increasing amounts of data in multiple formats, extraction, interoperability, real time integration of various types of data and information;
- ✓ Pre-processing of large data streams, including a selection of attributes, filtering, discretization;
- ✓ In-silico knowledge discovery out of Big data streams by applying methods for machine learning;
- ✓ Post-processing of data - knowledge interpretation and results visualization.

The hardware part of the scalable framework is consisted of the custom design system based on GPU accelerators developed by our team. It has the following parameters: Processor: Intel Core I7-6800K /3.4G/15MB, Accelerators: GeForce® GTX 1080 Ti Super JetStream, CUDA Cores - 3584, Memory Interface - 352bit, DRAM Type - GDDR5X, Memory Bandwidth (GB/sec) – 484, Microsoft DirectX - 12 API with feature level 12_1, OpenGL - 4.5, Bus Support - PCI-E 3.0 x 16, Maximum Digital Resolution - 7680x4320@60Hz, Height - 2.5 Slot, memory: 3T SG SATA ST3000DM008, Intel SSD 540s Series 480GB, 2.5in SATA 6Gb/s, 16nm, TLC.

For implementation of the software part of the proposed scalable framework, Linux operating system is used, various libraries for different algorithms, open source cluster computing framework, open source cross-platform management system and a wide range of data set repositories in the area of genomics are utilized.

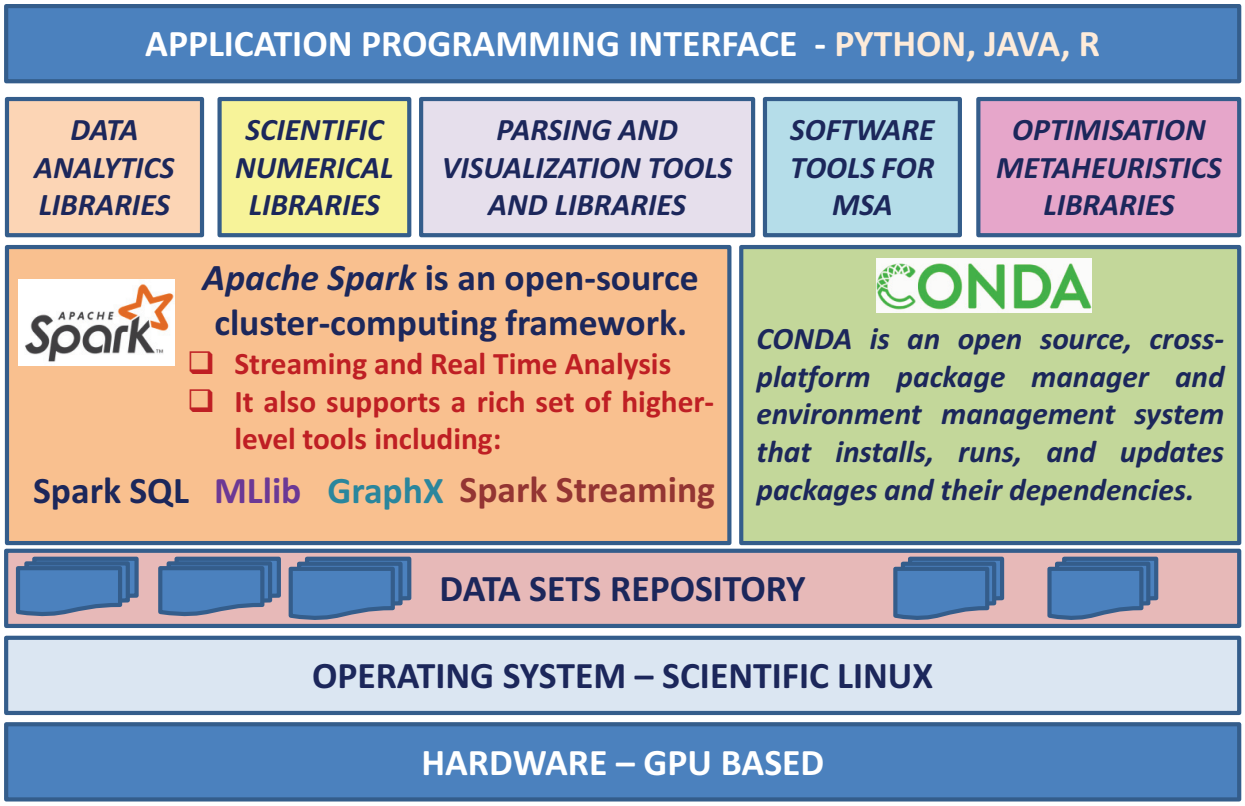

FIGURE 1. Scalable framework for adaptive in-silico knowledge discovery and decision-making

❖ Operating System

Scientific Linux (SL) is a Linux distribution that is used in the proposed scalable framework. SL is produced by Fermilab, CERN, DESY and by ETH Zurich. It is a free and open-source operating system based on Red Hat Enterprise Linux (RHEL) [5]. The choice of this operating system is based on our experience. Our team chooses RHEL because it has a lot of benefits. It is stable and supported for many year so that it is not necessary to choose between updating to get security fixes, and staying with an old release where your custom software works.

❖ Data Sets Repositories

In proposed scalable framework, it is utilized different genomic data sets repositories for big data analytics and decision-making with respect to precision and personalized medicine like breast cancer dataset repositories obtained from the University of Wisconsin Hospitals, The Cancer Genome Atlas (TCGA) and Enhancer-Promoter Interactions dataset repositories proposed by TargetFinder project.

Breast cancer dataset repositories obtained from the University of Wisconsin Hospitals has the following attributes Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. The values of the attributes are between 1 and 10. Each instance of the dataset has one of two possible classes: benign indexed with 2or malignant index with 4. The class distribution is for Benign: 458 (65.5%) and for Malignant: 241 (34.5%) [6].

The Cancer Genome Atlas (TCGA) is collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer. The TCGA dataset, comprising more than two petabytes of genomic data, has been made publically available, and this genomic information helps the cancer research community to improve the prevention, diagnosis, and treatment of cancer [7].

The enhancer-promoter interactions datasets repository provided by TargetFinder [8] project includes six cell lines (GM12878, HeLa-S3, HUVEC, IMR90, K562, and NHEK). The data for each cell line consist of enhancer-promoter pairs which are annotated as positive (interacting) or negative (non-interacting) using high-resolution genome-wide measurements of chromatin contacts in each cell line. Cell-line specific active enhancers and promoters are identified using annotations from ENCODE Project [9].

- Open Source Cluster Computing Framework

For implementation of the scalable framework for adaptive in-silico knowledge discovery and decision-making, Apache Spark is utilized. It allows streaming and real-time analyses, packet processing and export large amounts of data. Apache Spark is a fast and general cluster computing system for big data. It provides high-level APIs in Scala, Java, Python, and R, and an optimized engine that supports general computation graphs for data analysis. It also supports a rich set of higher-level tools including Spark SQL for SQL and DataFrames, MLlib for machine learning, GraphX for graph processing, and Spark Streaming for stream processing. The main abstraction of Spark is a resilient distributed dataset (RDD), which is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel. Using RDD Spark hides data partitioning and so distribution that in turn allowed them to design parallel computational framework with a higher-level programming interface (API) for four mainstream programming languages. [10].

- Open Source Cross-Platform Management System

CONDA open source environment management system is used. It easily creates, saves, loads and switches between environments. It was created for Python programs, but it can package and distribute software for any language. A popular CONDA channel for bioinformatics is BIOCONDA, which provides multiple software distributions for computational biology that is fully capable with the experimental investigations of the proposed scalable framework. Furthermore, the CONDA package and environment manager is included in all versions of Anaconda, Miniconda and Anaconda Repository [11].

- Software Libraries and Tools

Software part of the proposed scalable framework has been comprised of a set of libraries for different algorithms as Data Analytics Libraries, Scientific Numerical Libraries, Parsing and Visualization Tools and Libraries, Software Tools for Multiple Sequence Alignment and Optimization Metaheuristics Libraries.

The core of *Data Analytics Libraries* is MLlib of Apache Spark [12]. MLlib is a scalable machine learning library which includes many machine learning algorithms for classification, regression, recommendation, clustering, topic modeling and a various utilities for feature transformations, machine learning pipeline construction, model evaluation, hyper-parameter tuning, machine learning persistence, distributed linear algebra and statistics.

Scientific Numerical Libraries are used in the proposed scalable framework to perform numerical calculations. They are compound of: 1) Fundamental package for scientific computing with Python - *NumPy*; 2) Python-based open-source software for mathematics, science, and engineering – *SciPy*; 3) Software library written in Python programming language for data manipulation and analysis – PANDAS; 4) Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms – *MATPLOTLIB*.

Parsing and Visualization Tools and Libraries of the proposed scalable framework for adaptive in-silico knowledge discovery and decision-making is comprised of a free bioinformatics software *UGENE* [13] for genome sequencing data analysis and amino acid sequence visualization and a free genome browser and annotation tool *ARTEMIS [14]* that allows visualization of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation.

Software Tools for Multiple Sequence Alignment are embedded into the proposed scalable framework. This part of the framework has compound two software tools: *ClustalW-MPI* [15] and *MSA-BG* software [16]. ClustalW is a widely used multiple sequence alignment tool for DNA or proteins and implements a progressive method for multiple sequence alignment [5]. It calculates the best match for the selected sequences and lines them up so that the identities, similarities and differences can be seen. The basic algorithm behind ClustalW proceeds in three stages: pairwise alignment (PA), guide tree (GT) and multiple alignment (MA). Each of the phases produces intermediate data which is used as an input for the next one, but the calculations themselves are independent. ClustalW-MPI is a

distributed and parallel implementation on distributed computer clusters and on traditional parallel computers and uses a scheduling strategy called fixed-size chunking where batches of tasks of one fixed size are allocated to available processors. MSA-BG is parallel software for multiple sequence alignment based on social behavior metaheuristics. This software is developed by our team. It is ported on JuQueen Supercomputer at Forschungszentrum Jülich, Germany.

Optimisation Metaheuristics Libraries have included into proposed framework and worked with a set of metaheuristics algorithms like *Ant Colony Optimization (ACO), Artificial Bee Colony (ABC) and Genetic algorithms (GA).* ACO is a probabilistic technique that is used for solving computational problems which can be reduced to finding good paths through graphs. This algorithm is a member of the ant colony algorithms family, in swarm intelligence methods. It was aiming to search for an optimal path in a graph, based on the behavior of ants seeking a path between their colony and a source of food [17]. ABC is an optimization algorithm based on the intelligent foraging behavior of honey bee swarm. A set of honey bees, called swarm, can successfully accomplish tasks through social cooperation [18]. GA is a metaheuristic inspired by the process of natural selection that belongs to the larger class of evolutionary algorithms (EA). Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection [19]. All these metaheuristics algorithms are implemented in the proposed scalable framework and are used for precise and accurate experimental investigations and decision making in the area of genomic big data analytics.

The proposed framework for adaptive in-silico knowledge discovery and decision-making is scalable and supports application programming interfaces (APIs) in PYTHON, JAVA and R. It that brings together all the different components to facilitates development of software applications in the area of genomic big data analytics with respect to precision and personalized medicine.

The proposed framework provides a novel approach for designing experiments using the potential of big data ecosystem and its influence in genomics. The proposed framework is verified for the case studies of MSA based on social behavior model, enhancer-promotor interactions and early detection of breast cancer, Figure 2.

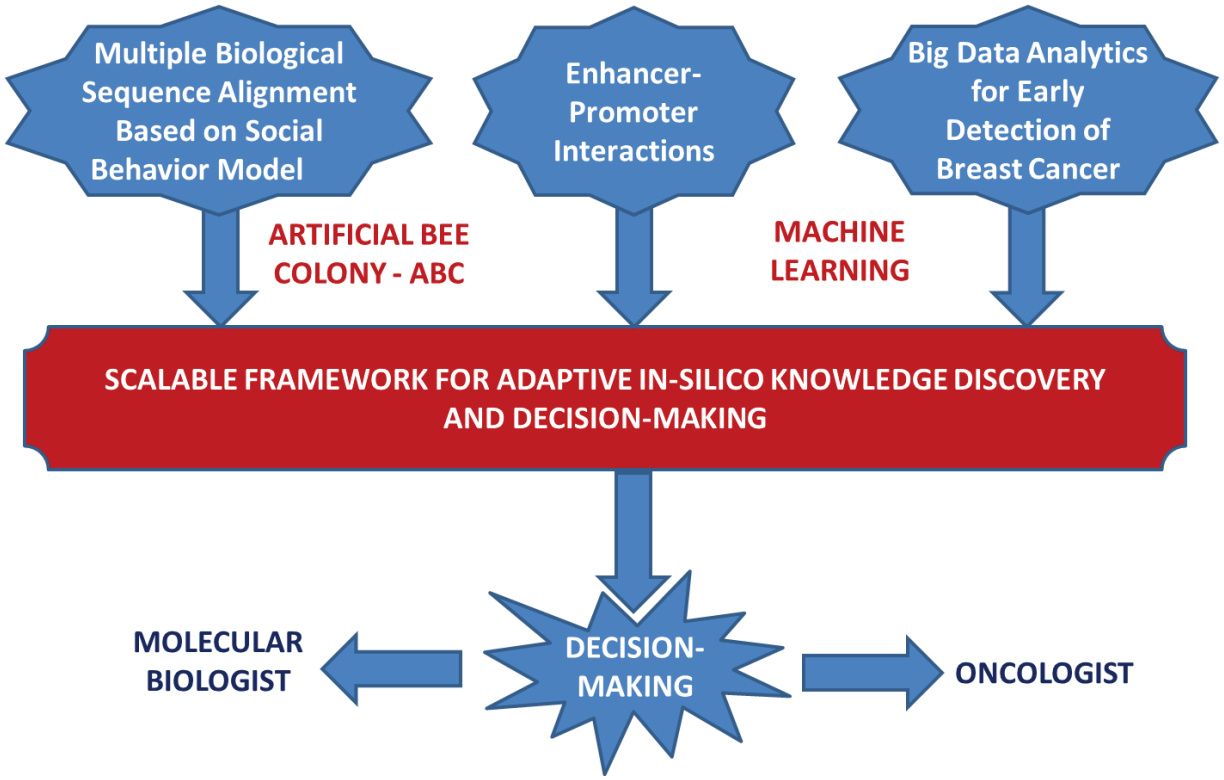

FIGURE 2. Experimental investigations based on scalable framework for adaptive in-silico knowledge discovery and decision-making

The proposed framework is used for the investigation of multiple sequence alignment with MSA-BG software for the case study of the influenza virus sequences. The goal of the experiments is to purpose a parallel multithreaded optimization including OpenMP. The experimental results show that the hybrid parallel implementation utilizing MPI and OpenMP provides considerably better performance than the original code [15].

The experimental investigations for detection of enhancer-promoter interactions from genomic big data based on Decision Tree and Support Vector Machine classifiers are implemented based on the proposed scalable framework [20]. The enhancer - promoter interactions is one of the most challenging phenomena in genomics field. The GM12878 and K562 datasets are used for experiments. The achieved performance parameters of 91 – 95 % accuracy are really competitive.

The breast cancer experiments based on Naïve Bayes classifier by using the proposed scalable framework have been carried out and the results are published in [21]. The software is written in python programming language and for the experiments the Wisconsin breast cancer datasets are used. The value of the achieved accuracy is 0.978597610089. The experimental results are really competitive. Moreover, the developed software offers a module for DNA sequence analytics. The sequences of BRCA1 and BRCA2 genes are analyzed with respect to number of nucleotides in the sequence, the adenine-thymine and guanine and cytosine content.

CONCLUSION AND FUTURE WORK

This paper has presented the scalable framework for adaptive in-silico knowledge discovery and decision-making for the area of genomics big data with respect to precision and personalized medicine. The proposed framework has been compound of hardware and software part. The hardware part is GPU-based and software part includes SL operating system, a various libraries for different algorithms like: *Data Analytics Libraries, Scientific Numerical Libraries, Parsing and Visualization Tools and Libraries, Software Tools for Multiple Sequence Alignment and Optimization Metaheuristics Libraries*, Apache Spark as a cluster computing system and CONDA cross platform management system. The proposed framework is verified for the case studies of MSA based on social behavior model, enhancer-promotor interactions and early detection of breast cancer.

The future work is to is to deploy service oriented platform to access the software tools of the framework (as a service), infrastructure (as a service) and platform (as a service), as well as open access to the constructed knowledge base and learning and testing data sets. The infrastructure will be interconnected via RRI hub to the platform of EU Responsible Research and Innovation community (http://www.rri-tools.eu/) to enable science and innovation transfer for significant scientific areas.

ACKNOWLEDGMENTS

The research is supported by the Project “Intelligent Method for Adaptive In-silico Knowledge Discovery and Decision Making Based on Analysis of Big Data Streams for Scientific Research” funded by the Bulgarian National Science Foundation, Bulgarian Ministry of Education and Science, Competition for financial support of fundamental research (2016) under the thematic priority: Technical Science, contract № ДН07/24 - 15.12.2016.

REFERENCES

[1] Gutierrez D., *Inside BIG DATA Guide to Scientific Research*, Dell, Intel: www.insidebidata.com/508-259-8570

[2] Chen P., Zhang C, *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Journal of Information Sciences,* www.elsevier.com/locate/ins

[3] *Challenges and opportunities with Big Data*, A community white paper available at: http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

[4] Akalin A., Kormaksson M., Li S., Garrett-Bakelman FE., Figueroa ME., Melnick .A, Mason CE, *Methylkit: a comprehensive R package for the analysis of genome-wide DNA methylation profile*, Genome Biol., 2012, 13:R87. 10.1186/gb-2012-13-10-r87 doi: 10.1186/gb-2012-13-10-r87.

[5] Scientific Linux: https://www.scientificlinux.org/

[6] William H. Wolberg, Breast Cancer Database, University of Wisconsin Hospitals, Madison, Wisconsin, USA.

[7] Cancer Genome Atlas, cancergenome.nih.gov/

[8] TargetFinder project: https://github.com/carringtonlab/TargetFinder
[9] ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI), website: https://www.encodeproject.org/
[10] Apache Spark RDD – Tutorialspoint: https://www.tutorialspoint.com/apache_spark/apache_spark_rdd.htm
[11] CONDA Documentation: https://CONDA.io/docs/
[12] MLlib Apache Spark: https://spark.apache.org/mllib/
[13] Unipro UGENE: http://ugene.net/
[14] ARTEMIS Free Genome Browser: https://www.sanger.ac.uk/science/tools/artemis
[15] P. Borovska, V. Gancheva, I. Georgiev, D. Ivanova, *Hybrid Parallel Multiple Sequence Alignment Based on Artificial Bee Colony on the Supercomputer JUQUEEN*, EECS 2017: European Conference on Electrical Engineering and Computer Science, Bern, Switzerland, November 17-19th, 2017.
[16] P. Borovska, V. Gancheva, *Parallelization and Optimization of Multiple Biological Sequence Alignment Software Based on Social Behavior Model*, 18th International Conference on Applied Computer and Applied Computational Science (ACACOS'18), (World Scientific and Engineering Academy and Society, Paris, France, April 13-15 2018), http://www.wseas.org/cms.action?id=16782
[17] Weifeng Gao, Sanyang Liu, Lingling Huang, *A global best artificial bee colony algorithm for global optimization*, Journal of Computational and Applied Mathematics, Volume 236, Issue 11, May 2012, Pages 2741-2753.
[18] Mustafa Servet Kıran, Oğuz Fındık, *A directed artificial bee colony algorithm*, Applied Soft Computing Volume 26, January 2015, Pages 454-462, https://doi.org/10.1016/j.asoc.2014.10.020
[19] Salvatore R. Mangano, *An Introduction to Genetic Algorithm Implementation, Theory, Application, History and Future Potential, White paper:* *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.55.1751&rep=rep1&type=pdf*
[20] D. Ivanova, P. Borovska, V. Gancheva, *Experimental Investigation of Enhancer-Promoter Interactions out of Genomic Big Data based on Machine Learning*, International Journal of Computers, Volume 3, 2018, ISSN: 2367-8895, pp. 58-62, http://www.iaras.org/iaras/journals/ijc
[21] D. Ivanova, *Big Data Analytics for Early Detection of Breast Cancer Based on Machine Learning*, Proceedings of the 43rd International Conference Applications of Mathematics in Engineering and Economics, AIP Conf. Proc. 1910, 060016-1–060016-8; https://doi.org/10.1063/1.5014010, Published by AIP Publishing. 978-0-7354-1602-4, 060016-1 - 060016-8.