Platform for Adaptive Knowledge Discovery and Decision Making Based on Big Genomics Data Analytics

Plamenka Borovska[1], Veska Gancheva[1](✉), and Ivailo Georgiev[2]

[1] Technical University of Sofia, Kliment Ohridski 8, 1000 Sofia, Bulgaria
{pborovska, vgan}@tu-sofia.bg
[2] The Stephan Angeloff Institute of Microbiology,
Georgi Bonchev 26, 1113 Sofia, Bulgaria
ivailo@microbio.bas.bg

Abstract. In the past years, researchers and analysts worldwide determine big data as a revolution in scientific research and one of the most promising trends that has given impetus to the intensive development of methods and technologies for their investigation and has resulted in the emergence of a new paradigm for scientific research Data-Intensive Scientific Discovery (DISD). The paper presents a platform for adaptive knowledge discovery and decision making tailored to the target of scientific research. The major advantage is the automatic generation of hypotheses and options for decisions, as well as verification and validation utilizing standard data sets and expertise of scientists. The platform is implemented on the basis of scalable framework and scientific portal to access the knowledge base and the software tools, as well as opportunities to share knowledge and technology transfer.

Keywords: Big genomic data analytics · Data integration · Knowledge discovery from data

1 Introduction

During the last years leading scientists, researchers and analysts determine big data as revolution in scientific studies and one of the most challenging trends in technological innovations. As a result of the computer simulations a huge amount of data was generated during the in silico experiments [1]. One of the fundamental scientific areas, strongly dependent on big data technologies, is molecular and computational biology [2]. The technological progress, as well as next generation sequencing yielded exponential growth of experimental genomic data, and as a result the well-known methods and technologies became not applicable to the new challenges of big data. This has stimulated the development of methods and technologies for processing of large amount of data and radical changes in the scientific research paradigms.

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 scientists all over the world, in building up solutions for specific problems. The advance in bioinformatics stimulates innovative methods for processing and analyzing the collected data.

I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11466, pp. 297–308, 2019.
https://doi.org/10.1007/978-3-030-17935-9_27

With the advances in bioinformatics, the volume of collected biological data and the number of databases in which they are stored have been steadily growing. They are supported by various organizations and institutions dealing with human genome research, virus research and their mutations, protein research, drug synthesis, and so on. A major problem for the biological data integration is the data structure and format. They are not always standardized and data access is not centralized, making it extremely difficult and time-consuming to search throughout all databases.

Biological knowledge is distributed in specialized databases data sources. Each database has its own complex data structures reflecting the scientific concept of the model [3]. Many data sources have overlapping data elements with conflicting definitions. Data integration from heterogeneous sources is very important for the effective use of biological information. It is important to interpret the different data formats, download data from different sources and convert them to integrated information. Biological data sources are characterized by an extremely high degree of heterogeneity with regard to the type of data model and the relevant data pattern, as well as the incompatible formats and nomenclatures of values [4].

Biological databases are highly decentralized, with high levels of terminology, record specificity, data representation, and request formats [5]. This in turn is associated with problems with manually executing queries from multiple databases. Therefore, there is a need to automate the integration of biological databases, with much more than simply extracting and modifying the data [6, 7]. Integration requires the use of binding formats in different databases, but large scale and redundancies make such integration impossible.

The next generation of methods for data analysis has to manage huge amounts of data from various types of sources with differentiated characteristics, levels of trust, and frequency of updates. Data analyses have to acquire knowledge in effective and sustainable way. To achieve this, it is necessary to build up complicated predictive models and methods for heterogeneous and big data analytics. On the other hand, these models and methods have to be implemented in real time for big data streams. This is a great challenge, because big data, besides its huge volume, are strongly heterogeneous and dynamic, requiring high performance and scalability.

The vast amount of accumulated data contains valuable hidden knowledge that can be useful to facilitate and improve the decision making process. That is why there is an objective need to create automated methods for extracting knowledge from data. The extracted knowledge have to meet the following requirements: be accurate, understandable and useful. In addition, knowledge should have the potential to predict. Data acquisition tasks include classification, dependency modeling, clustering, function retrieval, and associative rules detection.

Knowledge discovery and data mining is an area focused on methodologies for extracting useful knowledge from data. The ever increasing growth of data and the pervasive use of databases have demanded challenges for innovative knowledge data discovery methodologies. Data acquisition skills are based on research in statistics, databases, modeling, machine learning, data visualization, optimization, and high performance computing to deliver automated sophisticated and intelligent solutions.

The work presented in this paper is a part of a project that offers a scientific platform for adaptive in silico knowledge data discovery based on big genomic data analytics. The focus is on advanced information technologies and the fourth scientific research paradigm Data Intensive Scientific Discovery (DISD) in support of precision medicine, specifically, for the case study of fighting breast cancer.

An intelligent method for adaptive in silico knowledge discovery based on big genomic data analytics which is adaptable to important biological, medical and computational aspects has been suggested in [8]. The method is built upon the parallel phase paradigm comprising two overlapping and correlated phases – machine learning phase and operational phase.

The goal of this paper is to suggest platform for adaptive knowledge discovery and decision making based on big genomic data analytics based on the designed method for knowledge discovery stated above. The platform integrates scalable framework and scientific portal to access the knowledge base and the software tools, as well as opportunities to share knowledge and technology transfer.

The paper is structured as follows: Knowledge discovery from data pipeline is discussed in Sect. 2. Scalable adaptive method for knowledge discovery and decision making based on big data analytics is presented in Sect. 3. Conceptual architecture of data retrieval and integration system is presented in Sect. 4. Section 5 is focused on the scientific application specific platform.

2 Knowledge Discovery Based on Data Analytics

The objectives of knowledge discovery based on data analytics are predictive and descriptive as follows.

- The predictive includes the use of some variables or fields in a database to predict unknown or future values of other variables of interest.
- The description focuses on finding interpretative models describing the data.

Predictive and description objectives are achieved by using the following data retrieval tasks:

- Classification is a function that classifies a data element in one of several predefined classes.
- Regression is a function that classifies a data element of a predictive variable with a real value.
- Clustering is a general descriptive task that identifies a limited set of categories or clusters to describe the data.
- Grouping includes methods for finding a compact description for a subset of data.
- Modeling the relationship is finding a model that describes significant dependencies between variables.
- The change and detection of deviation is focused on detecting the most significant changes in data from previous measured or normative values.

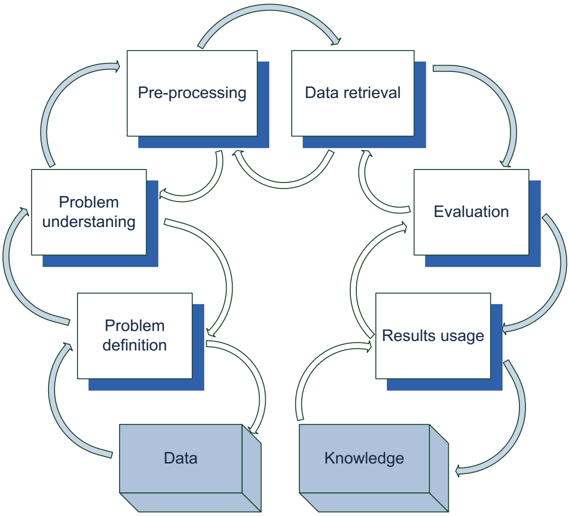

Fig. 1. The knowledge discovery based on data analytics process

Knowledge discovery and data mining process (Fig. 1) consists of six main stages:

1. Understanding the problem area - this is the initial stage that focuses on defining research goals and relevant requirements from the user's point of view. Once this stage has been completed, this knowledge has to be translated into definitions of data retrieval tasks and a preliminary plan for how these goals can be achieved.
2. Understanding the data - begins with initial data collection and continues with activities aimed at deepening the knowledge in respect to the data nature. At this stage, it is necessary to identify problems related to the data quality, to get an initial opinion on the data nature, to find the relevant subsets of data in order to form initial hypotheses about the hidden knowledge.
3. Data preparation - covers all the activities of creating raw data out of the final set of raw data. The stage of data preparation often has to be repeated many times at different stages of the computational pipeline. The tasks of data preparation include data selection, determining of attributes, exploring individual records, as well as data transformation and clearing data.
4. Modeling - this stage consists of selecting and applying various modeling techniques to derive data dependencies. The model parameters are adjusted to their optimal values. Since some models have their own specific data format requirements, it is often necessary to return to the data preparation stage.
5. Model evaluation - consists of carefully reviewing all the steps implemented in building up this model to ensure that they achieve the specific goals. At the end of this stage, a decision is made to use the results obtained during the drilling process.
6. Exploitation of the model - related to the monitoring and exploitation strategy applied. At this stage it should be determined whether and when to resume the procedure of data mining and under what conditions.

Machine learning is artificial intelligence technique for data analytics and its algorithms allow software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to design algorithms that can process input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

Classification is the most intensively studied task. Data subjected to analysis are divided in two groups: training set and validation set. The algorithm for knowledge extraction should build up rules, applying the training set. After completing the process of learning and establishing the classification set of rules, the effectiveness of the rules is evaluated on the basis of the validation set. The task of dependency modeling can be regarded as generalization of the classification task. In this case the goal is to predict the values of several attributes.

The conceptual model of knowledge discovery and decision making based on big data analytics is shown in Fig. 2.

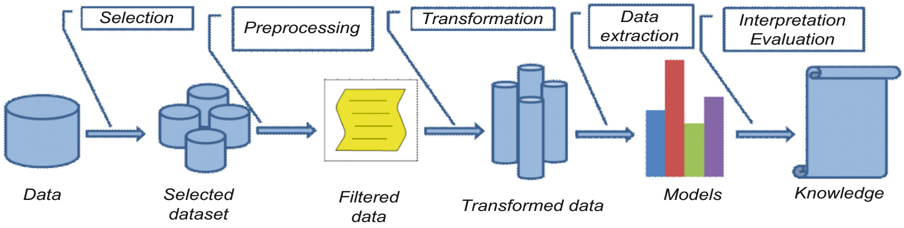

Fig. 2. Conceptual model for knowledge discovery and decision making based on big data analytics

The complete process of extracting and interpreting data models involves re-implementing the following steps:

- Determine the purpose of the knowledge discovery process - define the task and the relevant advance knowledge and its application aspects.
- Define application domain, relevant knowledge and end-user's goals.
- Create target data set: selecting a data set or focusing on a subset of variables or data samples.
- Filter and pre-process: remove noise or negative values; collect the necessary information for modeling or noise reporting; establish strategies for processing the missing data fields.
- Simplify the dataset by removing undesirable variables: find useful functions for presenting data in respect to the goal of to the task; apply dimensional or transformation methods to reduce the effective number of variables under consideration or to find invariant data representations.
- Combine the objectives of the data discovery process with data mining methods - determine whether the objective of the knowledge-based process is classification, regression, cauterization, etc.

- Select data mining algorithm. This process involves making decision which models and parameters may be appropriate for the overall process: select the method(s) to be used to search for model in data; determine which models and parameters may be appropriate; conformity of a particular data mining method with the common criteria of the knowledge discovery process.
- Data extraction - searching for models of interest in a specific presentation form or a set of such representations such as classification rules or trees, regression, clustering, etc.
- Interpret basic knowledge of extracted models.
- Utilize knowledge and its inclusion in another system for further action.

3 Scalable Adaptive Approach for Knowledge Discovery and Decision Making Based on Big Data Analytics

The aim is to suggest an integrated approach for supporting the knowledge discovery and decision making processes based on big data analytics, adaptive machine learning and adaptive procedures for generating rules according to the specific goal of the scientific research. The main advantage is the automated generation of hypotheses and solutions for the specific case under study. The interest of the knowledge discovered is estimated by the expertise of scientists in the relevant field. The adaptability of the approach is achieved by means of a synthesized collection of modules based on techniques such as data analysis, machine learning, and metaheuristics. Depending on the specific purpose of the study, the relevant modules are applied for pre-processing of data flows, for knowledge extraction and post-processing. Regarding the knowledge extraction, a hybrid approach of machine learning methods and procedures for rules generation is applied.

A scalable framework for adaptive knowledge discovery is based on big data streams analytics, providing a set of software tools for applying the method in research and experimental activities for wide spectrum of scientific areas. Streaming technology eliminates the need for significant resources of disc memory as raw data derived from databases are subjected directly to online processing. The scalability of the working framework reduces computational time by involving additional resources and parallel processing.

The advantage of the proposed framework is the automatic generation of the hypothesis and options for decisions making on the basis of the learning data set analysis, while the verification and validation is conducted via benchmark testing set and the expertise of the researchers of the relevant area. The interest of the discovered knowledge is estimated by hybrid approach – a combination of objective criteria and the expertise of scientists.

The research techniques follow the processing pipeline for discovering knowledge of interest out of a collection of data and imply the following:

- data preparation, cleansing and selection;
- knowledge discovery and decision making, and
- results visualization and interpretation.

Pre-processing of data in knowledge discovery covers: integration of data from different sources, data clearing in terms of accuracy, sampling, and selection of functions in terms of relevance. The selection of functions is done by applying algorithms for searching and optimization, and the feature set is optimized by iterative execution of machine learning algorithm. The combination of machine learning, data mining, knowledge discovery and decision making methods is applied to achieve high accuracy and precision of the extracted knowledge. Post-processing involves verification, validation, visualization and evaluation of the discovered knowledge.

The basic functional units for knowledge data discover are scientific analytics workflows being exceedingly useful for facilitating e-science. Actually, a workflow represents the computational pipeline for KDD based on specific analytics method (model and/or rules). The workflow provides a solution, supporting the design and conducting experiments using the available data and tools and high-level definition of the objectives of the experiment, modeled by workflow of scientific tasks. Different types of tasks can be performed within a workflow, especially when outputs from one task are used as input for the next task. The workflow process for knowledge discovery from big genomics data is shown in Fig. 3.

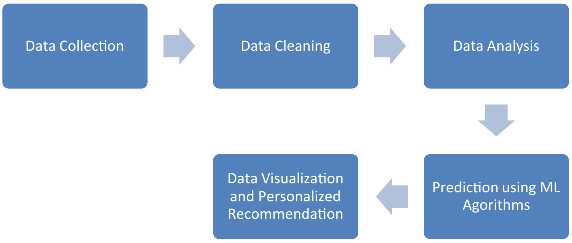
Data Collection
Data Cleaning
Data Analysis
Prediction using ML Agorithms
Data Visualization and Personalized Recommendation

Fig. 3. Workflow process for knowledge discovery from big genomics data

The purpose of the machine learning phase is to build up repositories of synthesized collection of models and rules that will be used as components to build up an integrated KDD workflow. Machine learning (ML) phase is based on parallel workflows (computational pipelines), utilizing diverse ML models, including classifying and clustering methods. ML phase performs on the training and validation sets and the basic computational units are bundles of differentiated workflows, each differentiated workflow performing specific type of analytics. Differentiated analytics workflows are executed in parallel in order to accelerate the KDD process. Each analytics workflow builds up a model that is stored in models repository or a set of rules stored in rules repository. Once feature extraction and dimension reduction have been done, analytics model is generated utilizing the training set by configuring the respective differentiated workflow. Afterwards, the outcome of the differentiated workflow is subjected to validation by utilizing the validation data set.

The components of the big data analytics system architecture are shown in Fig. 4. The conceptual architecture of the scalable framework for adaptive knowledge

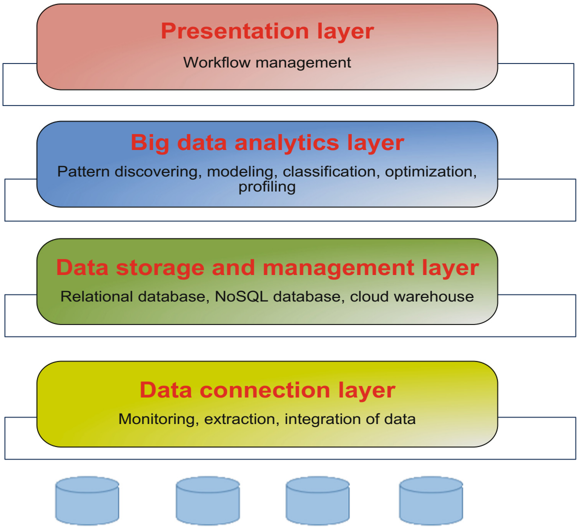
Presentation layer
Workflow management
Big data analytics layer
Pattern discovering, modeling, classification, optimization, profiling
Data storage and management layer
Relational database, NoSQL database, cloud warehouse
Data connection layer
Monitoring, extraction, integration of data

Fig. 4. Big data analytics system architecture

discovery comprises hardware and software resource reconfiguration. Software is developed utilizing streaming and parallel processing technologies (multithreading and clustering). The architecture consists 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 selection of attributes, filtering, discretization;
- knowledge discovery out of big data streams by applying methods for generating rules and machine learning and
- post-processing of data - knowledge interpretation and results visualization.

4 Data Integration

The requirements for scalable data integration systems for modern biology are indisputable due to the existence of very large, heterogeneous and complex datasets in the public database. Managing and merging these big data with local databases is a great challenge as it is the basis of computational analyzes and models that are then experimentally generated and validated through portal access to distributed modern high-performance infrastructure and software tools for big genomics data processing and visualization.

A conceptual architecture for an integrated and effective access to the exponentially growing volume of data in multiple formats is proposed (Fig. 5). The architecture allows the rapid management of large volumes of diverse data sets represented in different formats - relational, NoSQL, flat files. The integration system consists of services for transforming the common request into a specific language request for each local database, depending on its type. Additionally, the possibility of making a permanent access to the state of research in order to compare the results with the available information (access to a constantly updated representation of all the accumulated knowledge in the relevant field) is further explored.

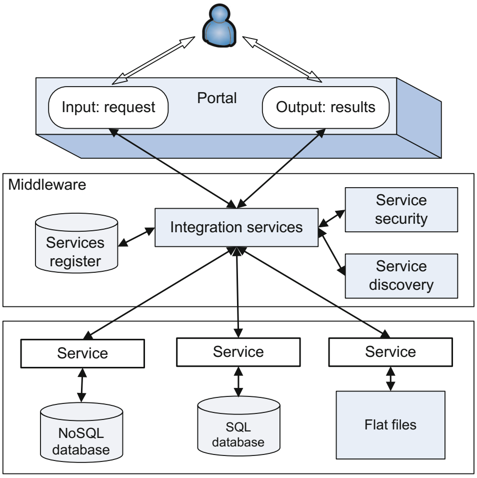

Fig. 5. Conceptual architecture of data retrieval and integration system

5 Scientific Specific Application Platform

The objective of an academic scientific portal gateway is to allow a large number of users to have transparent access to available distributed advanced computing infrastructure, software tools, visualization tools and resources.

The high-performance platform offers services for carrying out remote simulations and consists of HPC resources, resource database, knowledge discovery resources as services, software tools and web portal is presented in Fig. 6.

The portal provides user-centric view of all available services and distributed resources. The web-based environment is organized to provide a common user interfaces and services that securely access available heterogeneous computing resources, data and applications. It also allows many academic and scientific users to interact with these resources.

The experimental infrastructure is achieved through a customized application specific gateway, based on the platform and can be summarized as follows:

- identifying and clarifying the requirements of specific user communities, user scenarios and needs of the target groups of scientists;
- defining the specific user communities views in accordance with their requirements and develop custom scripts for specific applications;
- identification of methods and tools for user communities and their testing;

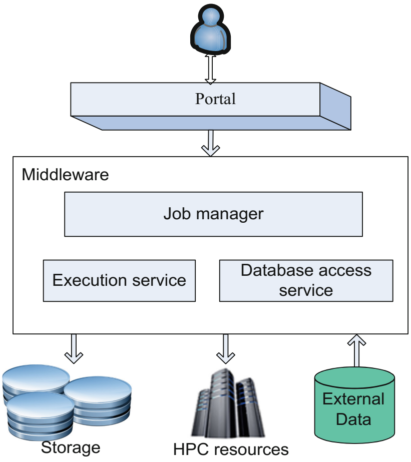

Fig. 6. High performance infrastructure

- development of workflow templates representing different user scenarios of low level and adapted to the specific requirements of customers;
- creating a knowledge base and data sets for training and testing.

The scientific platform will facilitate access and use of existing infrastructure in research and will provide:

- a central access point to the status of research in order to compare the results with the available information (access to constantly updated representation of the accumulated knowledge in the relevant area),
- interactive cooperation via various channels
- opportunities for knowledge and technology transfer, partnership and providing information and services.

The web-based environment is organized so as to provide a common user interfaces and services that provide secure access to currently available heterogeneous computing resources, data and applications.

The web portal provides also: user profile management, e.g., different views for different user; personalized access to information, software tools and processes; getting information from local or remote data sources, e.g., from databases, transaction systems, or remote web sites; aggregated the information into composite pages to provide information to users in a compact and easily consumable form. In addition, the portal also includes applications, software tools, etc.

The proposed platform is verified for the case studies of multiple sequence alignment based on social behavioral model, enhancer-promoter detection and early detection of breast cancer. An investigation for detection of enhancer-promoter interactions out of genomic big data based on machine learning is presented in [9]. A pipeline for detection of enhancer-promoter interactions using Decision Tree and Support Vector Machine classifiers is proposed. The experimental framework is based on Apache Spark environment that allows streaming and real time analysis of big data. The experimental results for detection of enhancer-promoter interactions have been performed with GM12878 and K562.

An innovative parallel method MSA_BG for multiple biological sequences alignment that is highly scalable and locality aware is designed [10]. The algorithm is iterative and is based on the concept of Artificial Bee Colony metaheuristics and the concept of algorithmic and architectural spaces correlation. The metaphor of the ABC metaheuristics has been constructed and the functionalities of the agents have been defined. The conceptual parallel model of computation has been designed. Parallelization and optimization of the multiple sequence alignment software MSA_BG in order to improve the performance, for the case study of the influenza virus sequences is proposed [11]. For this purpose a parallel MPI + OpenMP - based code optimization has been implemented and verified. The experimental results show that the hybrid parallel implementation provides considerably better performance.

6 Conclusion and Future Work

In this paper a platform for adaptive knowledge discovery and decision making based on big data analytics is proposed. The major advantage is the automatic generation of hypotheses and options for decisions, as verification and validation are performed using standard data sets and expertise of scientists. The tools for utilizing the platform are scalable framework and scientific portal to access the knowledge base and the software tools, as well as opportunities to share knowledge, and technology transfer. Web portal provides services to access and extract knowledge out of biological data and execute parallel software applications for big genomics data analysis.

An integrated approach to support knowledge discovery and decision making based on big data analytics, adaptive machine learning and adaptive procedures for generating rules according to the goal of scientific research is presented. A conceptual architecture for an integrated and effective access to the exponentially growing volume of data in multiple formats is proposed. The architecture allows the rapid management of large volumes of diverse data sets represented in different formats - relational, NoSQL, flat files. The integration system consists of services for transforming the common request into a specific language request for each local database, depending on its type.

The future work is to make in silico experiments on the platform based on big genomic data analytics for scientific research in the area of molecular biology. The spectrum of case studies under investigation comprises identifying regulatory elements in sequenced genomes, and prediction of the type and malignance of breast cancer. This will enable fast processing of clinical observations and laboratory analyzes data and comparison with the available data accumulated so far in support of precision medicine.

Acknowledgment. This paper presents the outcomes of research project “Intelligent Method for Adaptive In-silico Knowledge Discovery and Decision Making Based on Analysis of Big Data Streams for Scientific Research”, contract DN07/24, financed by the National Science Fund, Competition for Financial Support for Fundamental Research, Ministry of Education and Science, Bulgaria.

References

1. Chen, P., Zhang, C.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. J. Inf. Sci. **275**, 314–347 (2014). https://doi.org/10.1016/j.ins.2014.01.015
2. Roy, A.K.: Trends in computational biology and bioinformatics in the era of big data analytics. In: Conference International Workshop on Bioinformatics in Fisheries and Aquaculture' held at ICAR- CIFRI (2017). https://doi.org/10.13140/rg.2.2.21016.39680
3. Thiam Yui, C., Liang, L.J., Jik Soon, W., Husain, W.: A survey on data integration in bioinformatics. In: Abd Manaf, A., Sahibuddin, S., Ahmad, R., Mohd Daud, S., El-Qawasmeh, E. (eds.) ICIEIS 2011. CCIS, vol. 254, pp. 16–28. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25483-3_2
4. Nguyen, H., Michel, L., Thompson, J.D., Poch, O.: Heterogeneous biological data integration with declarative query language. IBM J. Res. Dev. **58**(2/3), 1–12 (2014)
5. Rao, C.S., Somayajulu, D.V.L.N., Banka, H., Ro, S.: Feature binding technique for integration of biological databases with optimized search and retrieve. In: 2nd International Conference on Communication, Computing & Security [ICCCS-2012], pp. 622–629 (2012)
6. Paton, N.W., Missier, P., Hedeler, C. (eds.): DILS 2009. LNCS, vol. 5647. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02879-3
7. Zhang, Z., Bajic, V.B., Yu, J., Cheung, K.-H., Townsend, J.P.: Data integration in bioinformatics: current efforts and challenges. In: Bioinformatics - Trends and Methodologies. pp. 41–56 (2011). ISBN 978-953-307-282-1. http://doi.org/10.5772/21654
8. Borovska, P., Ivanova, D.: Intelligent method for adaptive in silico knowledge discovery based on big genomic data analytics. In: AIP Conference Proceedings, vol. 2048, p. 060001 (2018). https://doi.org/10.1063/1.5082116
9. Ivanova, D., Borovska, P., Gancheva, V.: Experimental investigation of enhancer-promoter interactions out of genomic big data based on machine learning. Int. J. Comput. **3**, 58–62 (2018). ISSN: 2367-8895
10. Borovska, P., Gancheva, V., Landzhev, N.: Massively parallel algorithm for multiple biological sequences alignment. In: Proceedings of the IEEE International Conference on Telecommunications and Signal Processing (TSP), Rome, Italy, pp. 638–642. ISBN 978-1-4799-0402-0
11. Borovska, P., Gancheva, V.: Parallelization and optimization of multiple biological sequence alignment software based on social behavior model. Int. J. Comput. **3**, 69–74 (2018). ISSN: 2367-8895