Intelligent Method for Adaptive In Silico Knowledge Discovery Based on Big Genomic Data Analytics

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

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

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

Abstract. The focus of this paper is on advanced IT 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. We suggest intelligent method for adaptive in silico knowledge data discovery based on Big genomic data analytics which is adaptable to important biological, medical and computational aspects. The method is built upon the parallel phase paradigm comprising two overlapping and correlated phases – machine learning phase and operational phase. The basic functional units in both phases are scientific analytics workflows – bundles of differentiated workflows in the ML phase, and integrated workflow in the operational phase, built upon optimal differentiated workflows stored in the best models repository. The applicability of the method has been illustrated by the presented conceptual model of smart digital consultant for personalized breast cancer diagnostics and therapy recommendations deploying the suggested method for in silico KDD. The method has been verified and validated for the case studies of differentiated descriptive analytics workflows for prokaryotic and eukaryotic gene finding and mapping and differentiated diagnostics analytics workflows for breast cancer associated gene mutation detection.

THE PROBLEM AREA

Precision medicine is hot topic nowadays and initially has been defined as "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person" [1]. Actually, precision medicine is conceived as “data-driven” treatment [2]. Precision medicine is based on the intersection of “omics” and Big data technologies. The third paradigm for scientific research and the intense progress of bioinformatics and IT in last decades resulted in the accumulation of huge amounts of in silico experimentation data that has great potential to be subjected to analysis in order to obtain value. The innovative forth scientific research paradigm “Data - Intensive Scientific Discovery – DISD” [3] gives the opportunity to discover in silico knowledge that can be exceedingly helpful for medical doctors to make up personalized disease diagnostics and prescribe optimal treatment for the individual patient.

In silico oncology [4] comprises a broad spectrum of investigation areas such as in silico models of cancer, in silico modeling for tumor growth visualization, in silico models for designing and discovering novel anticancer drugs, in silico tumor models, in silico experimental modeling of cancer treatment, multiscale cancer modeling, etc.

So far as breast cancer is concerned medical data has proven that it is mostly related to mutations in two genes: *BRCA1* (BReast CAncer gene one) and *BRCA2* (BReast CAncer gene two) [5]. There are 5 types of breast cancer, and 6 types of different cancer cells [6], each type demanding specific treatment. The great variety of breast cancer treatments is due to tumor heterogeneity i.e. the differences in cancer cells. The target of some treatment plans is a particular gene or protein in the cancer cells, others treat other targets. As a result, for the case study of breast cancer, each patient needs a specific treatment. Genomic analysis gives the opportunity to estimate the activity level of breast cancer – associated genes and, consequently, to prognosticate the risk of cancer coming back

[7]. In order to be able to achieve personalized disease diagnostics it is necessary to accumulate huge amounts of patients' personal genome data. The Personal Genome Project, initiated in 2005, is a vision and coalition of projects across the world dedicated to creating public genome, health, and trait data [8]. The Global Network of Personal Genome Projects includes researchers at leading institutions around the globe such as Harvard Personal Genome Project [9], Personal Genome Project Canada [10], Personal Genome Project UK [11], Personal Genome Project Austria [12] and Personal Genome Project China [13].

Of the foregoing, we can summarize that the problem of tailoring a personal treatment for a breast cancer patient is exceedingly complicated and the medical doctor has to explore and analyze huge amounts of various data concerning the genetic specifics (mutations), patient's life style, patient's health status and environmental factors. Big data technologies can be very beneficial in supporting medical doctors in managing data exploration and analysis.

The goal of this paper is to suggest intelligent method for adaptive in silico knowledge discovery on the basis of Big genomic data analytics in order to help and assist medical doctors in disease diagnostics and personalized breast cancer therapy of the individual patient by deploying smart digital consultant.

INTELLIGENT METHOD FOR ADAPTIVE IN SILICO KNOWLEDGE DISCOVERY

We suggest a method for knowledge discovery based on in silico models involving the Big genomic data ecosystem [14, 15] and Big genomic data analytics [16]. The research methods and techniques are set out in processing pipeline for knowledge discovery and decision-making and cover the following groups: (1) data preprocessing; (2) in-silco knowledge discovery and decision making, and (3) post-processing comprising results presentations (knowledge, predictions) in understandable way, suitable for interpretation and knowledge interest evaluation.

Data preprocessing in the case of knowledge discovery encompasses: data integration (from various sources, for ex.), data filtering (in respect to accuracy), discretization, features selection (in respect to relevance). Features selection is performed by applying metaheuristic algorithms for search and optimization as well as principal component analyses - PCA, and the feature set is optimized (reduced) by iterative execution of the machine learning algorithm. Besides, post-processing encompasses verification, validation and interest (usefulness, value) evaluation of the discovered knowledge, in situ visualization.

The basic functional units for in silico knowledge data discover (KDD) actually 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). The method of workflows provides a solution, supporting the design and conducting experiments using the available data and tools. Workflow is defined as a pattern defining the consistent implementation of processes or flow of tasks which is coordinated and scheduled on the basis of a systematic plan. Scientific workflows provide a method of 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 suggested KDD method (Fig.1) comprises 2 parallel and correlated phases, learning and operational phases, which overlap and perform in parallel, exchanging data. Both phases imply model-based analytics methods. Each phase operates on 4 types of data analytics workflows – descriptive, diagnostic, predictive and prescriptive. Descriptive analytics workflow focuses on breast cancer associated gene finding and mapping, diagnostic analytics workflow aims at breast cancer associated gene mutations detection, predictive analytics workflow is targeted to determining cancer type, malignancy and expected life duration, while the issue of prescriptive analytics workflow is recommended personalized therapy taking into consideration genetic specifics, individual patient's life style and environmental factors.

The purpose of the machine learning phase of the method is to build up a repository of synthesized collection of models that will be used in the operational phase as components to build up an integrated in silico KDD workflow. ML phase performs offline on the training and validation sets and the basic computational units are bundles of differentiated workflows, each differentiated workflow performing specific type of analytics – descriptive, diagnostics, predictive or prescriptive. The operational phase is conducted online on the input streaming patient's data and the basic computational unit is an integrated computational in silico KDD workflow built up of 4 optimal differentiated workflows (descriptive, diagnostics, predictive and prescriptive).

The advantage of the new method for in silico knowledge data discovery is the automatic generation of the hypothesis and options for making decisions on the basis of the learning 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 from the target area.

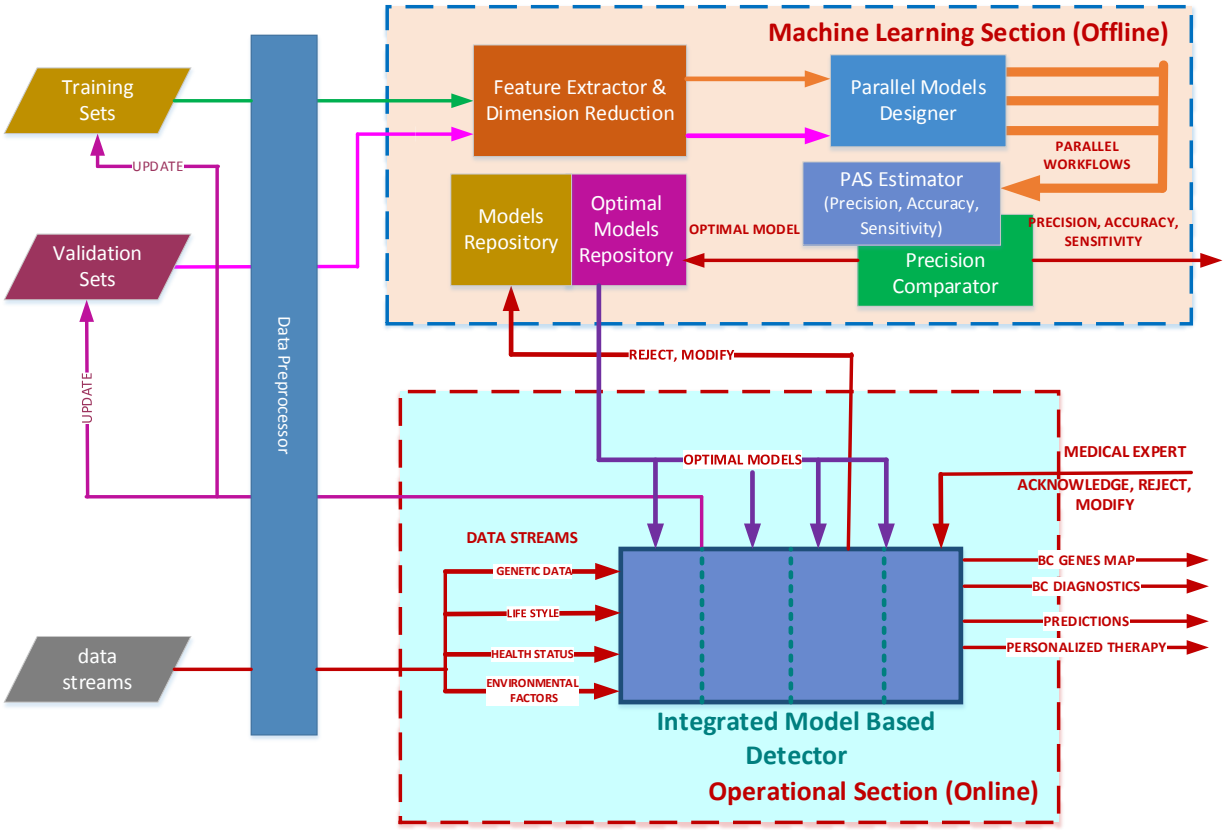

FIGURE 1. Intelligent phase-parallel model based adaptive method for in silico KDD

THE PHASE PARALLEL PARADIGM

Machine learning (ML) phase is based on parallel workflows (computational pipelines), utilizing diverse ML models, including classifying and clustering methods. Data sets are of several types: (1) patient's genetic data, (2) environmental factors, (3) patient's individual life style parameters, and (4) clinical tests results.

The learning phase operates on 4 types of differentiated data analytics computational workflows in parallel – descriptive, diagnostic, predictive and prescriptive (Fig. 2). The differentiated descriptive analytics workflow is targeted on breast cancer associated genes finding and mapping. The responsibility of the differentiated diagnostic analytics workflow is mutation detection and personalized diagnostics. The output of the differentiated predictive analytics workflow is cancer malignancy estimation and expected life duration. The task of the differentiated prescriptive analytics workflow is to yield personalized therapy recommendations.

ML phase operates offline because it requires considerable computational time. Nevertheless, differentiated analytics workflows are executed in parallel in order to accelerate the improvement process. Each analytics workflow builds up a model that is stored in models repository.

After feature extraction and dimension reduction the parallel model designer generates analytics model 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 and verified by medical or molecular

biological expert. In case validation and verification results are correct, the model joins the model repository and consequently is subjected to PAS (Precision, Accuracy, and Sensitivity) evaluation. Otherwise, training data sets is being modified and updated. The responsibility of the PAS comparator is to select the optimal differentiated workflows in respect to PAS values (one workflow per bundle) and to pass them to the operational phase to configure the integrated analytics workflow.

Within the learning phase for each type of data analytics a separate bundle of workflows is generated i.e. a bundle for descriptive analytics, a bundle for diagnostic analytics, a bundle for predictive analytics and a bundle for prescriptive analytics (differentiated workflows (models)). Each bundle of workflows is executed in parallel. The extracted knowledge of each workflow within the bundle is being evaluated in respect to precision, accuracy and sensitivity (PAS evaluation). As a result, the workflow (model) of the best PAS evaluation result is added to the optimal models repository, which stores the set of optimal differentiated workflows.

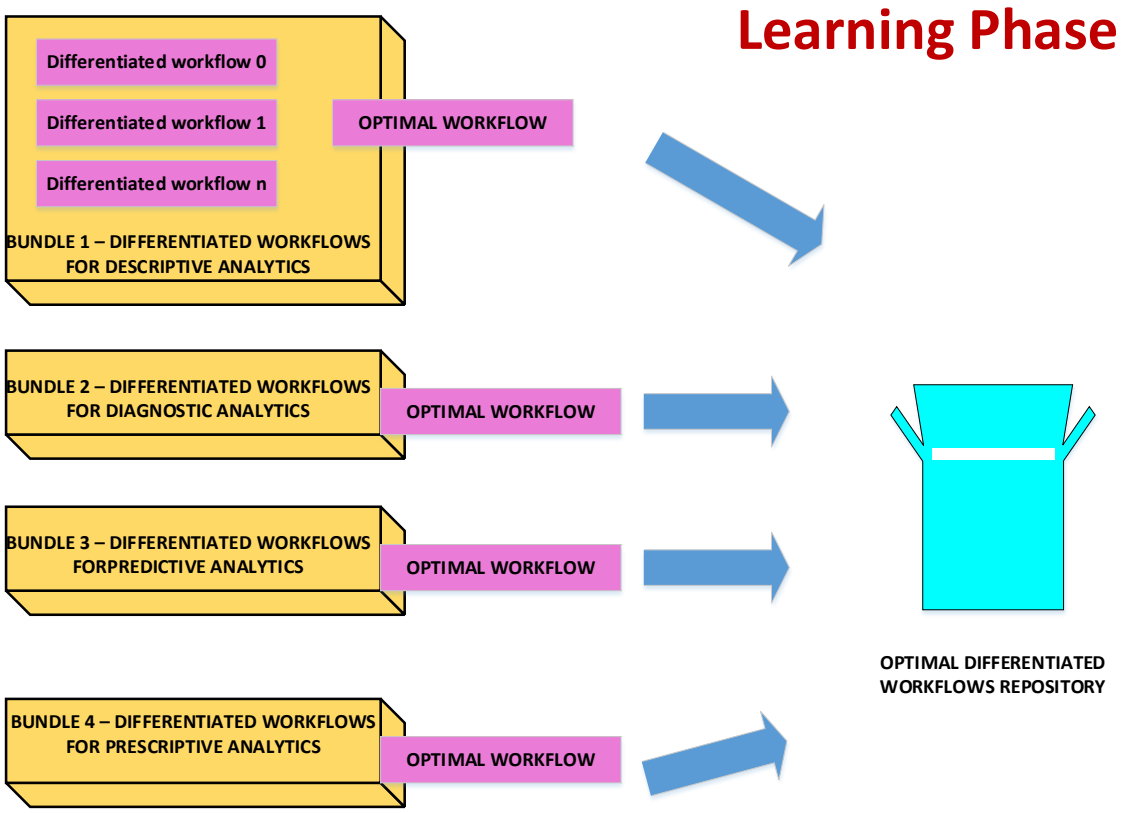

FIGURE 2 Machine learning phase of the method (offline)

Within *the operational phase* (Fig. 3) computation is organized on the bases of integrated in silico KDD workflows. An integrated workflow comprises 4 optimal differentiated KDD workflows, constructed in the ML phase and covering all 4 types of analytics i.e. the integrated in silico KDD workflow comprises 4 components – optimal descriptive differentiated workflow, optimal diagnostics differentiated workflow, optimal predictive differentiated workflow and optimal prescriptive differentiated workflow. The operational phase performs online and the input data actually are streams of patient's data – genetic data, clinical tests, individual life style parameters, and environmental factors. The outcome of the operational phase of the method is patient's breast cancer associated genes specifics, mutations detected, personalized breast cancer diagnostics, cancer malignancy estimation and expected life duration prediction, personalized therapy recommendations.

The knowledge discovered in the operational phase is subjected to medical expert's evaluation – acknowledge, reject, modify. In case of reject or modify expert evaluation results of the respective integrated workflow the building up differentiated workflows in the optimal workflows repository are marked as invalid, training and validating sets are being modified and updated and parallel workflows are initialized in the learning phase to improve the models. After clinical trials the medical doctor may initiate modification and update of data sets in order to improve quality of solution.

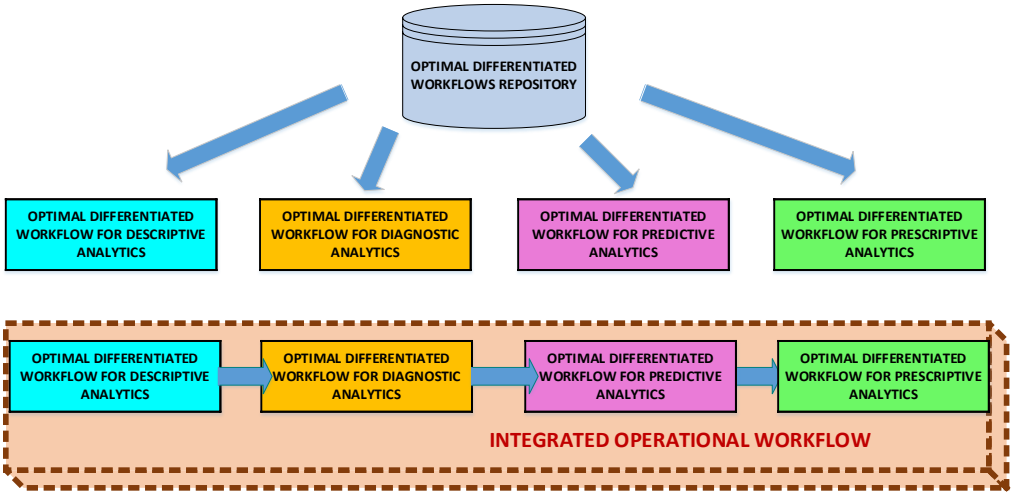

FIGURE 3 Operational phase of the method (online)

ADAPTABILITY ASPECTS AND APPROBATION

The adaptability of the method (Fig.4) comprises 3 fundamental aspects: genetic aspects, medical aspects and computational aspects. Biological aspects comprise genetic mutations and types of cancer cells. Medical aspects are related to diverse cancer therapies and the role of individual life style and environmental factors. The computational aspects have 3 major classes, which are correlated: (1) diverse models and methods; (2) scalability; and (3) polymorphic computational architecture (hardware and software resources reconfiguration).

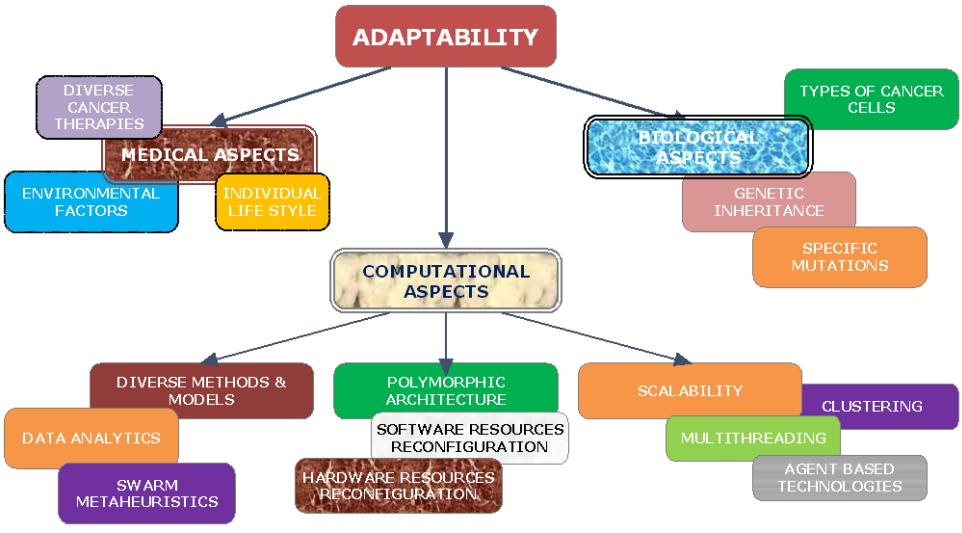

FIGURE 4 Adaptability aspects of the intelligent method for in silico knowledge discovery

The adaptability of the new method for in silico knowledge discovery and decision making is accomplished by scalable experimental framework supporting all the aspects. The hardware architecture is based on GPU-accelerated high performance workstation accessed remotely via MobaXterm local terminal. The reconfigurable software environment runs under Scientific Linux OS which does not limit the number of active local terminals.

The software architecture consists of separate, independent components: (1) access to progressively increasing amounts of data in multiple formats, extraction, interoperability, real time integration of various types of data and information; (2) pre-processing of large data streams, including a selection of attributes, filtering, discretization; (3) in-silico knowledge discovery out of Big data streams by applying methods for machine learning and (4) post-processing of data - knowledge interpretation and results visualization.

For our experimentation we have used the unified analytics engine for big data processing Apache Spark with built-in modules for streaming, SQL, machine learning and graph processing [17]. In order to verify and validate the suggested method a selection of specific up-to-date case studies have been constructed for which Big data from in silico experiments has been accumulated at the suggestion of experts in molecular biology. With the help of their expertise we have created data sets for training and testing on the basis of a representative patterns of relevant databases following value of the data criterion.

Gene finding and mapping based on promoter prediction have been subjected to intensive study, however, the techniques applied by now are not quite efficient and satisfactory, especially for genome scale analysis due to considerable rates of false positive predictions [18, 19, 20].

For the case study of identifying and mapping of genes we have built up a local database of transcriptional and translational regulatory genetic elements of referent prokaryotic genomes, in our case E. Coli. The nucleotide sequences of the selected transcriptional and translational regulatory genetic elements have been subjected to analysis to detect correlations on the basis of promoter sequences. The experimental results for E. Coli promoter prediction show precision 89-93% for applying computational tool for Artificial Neural Networks (ANN), 93-95% for Decision tree and 93-97% for Support Vector Machine (SVM).

The differential workflow for gene mapping and finding is constructed on the concept of the proposed computational pipeline for detection of enhancer-promoter interactions in [21] by applying Decision Tree and Support Vector Machine classifiers and GM12878 and K562 datasets.

Mutation detection approach for the diagnostics analytics differentiated workflow has been verified and validated on the bases of innovative method for multiple biological sequence alignment based on swarm metaheuristics for the case study of Influenza Virus AH1N1. The results are presented in [22].

The diagnostic analytics differentiated workflow concept has been verified for the case study of predicting the type of breast cancer (malignant, benign) on the basis of gene associated mutations BRCA1 and BRCA2, and the type of the tumor cells for data sets, obtained from the University of Wisconsin Hospitals where the samples arrive periodically [23].

APPLICABILITY - CONCEPTUAL MODEL OF PRECISION MEDICINE SMART DIGITAL CONSULTANT

The conceptual model of PM Smart Digital Consultant for the case study of breast cancer is shown in figure 4. The applicability of the suggested method for adaptive in silico KDD focuses on the design and implementation of Precision Medicine Smart Digital Consultant.

The PM smart digital consultant helps and assists medical doctors to process, manage and interpret the huge amounts of related data in diagnostics and personalized therapy prescription, and facilitating artificial and natural intellect interaction and cooperation for the benefits of wellbeing of people and healthcare.

The knowledge base comprises knowledge available on the web – data sets, case studies, benchmarks, etc. The intelligent system is being subjected to learning utilizing the available web knowledge and the suggested intelligent method for adaptive in silico KDD. The user (medical expert) may enrich the knowledge and/or modify the models according to his/her own experience.

CONCLUSIONS AND FUTURE WORK

The focus of this paper is on advanced IT 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. We suggest intelligent method for adaptive in silico knowledge data discovery based on Big genomic data analytics which is adaptable to important biological, medical and computational aspects. The method is built upon the parallel phase paradigm comprising two overlapping and correlated phases – machine learning phase and operational phase.

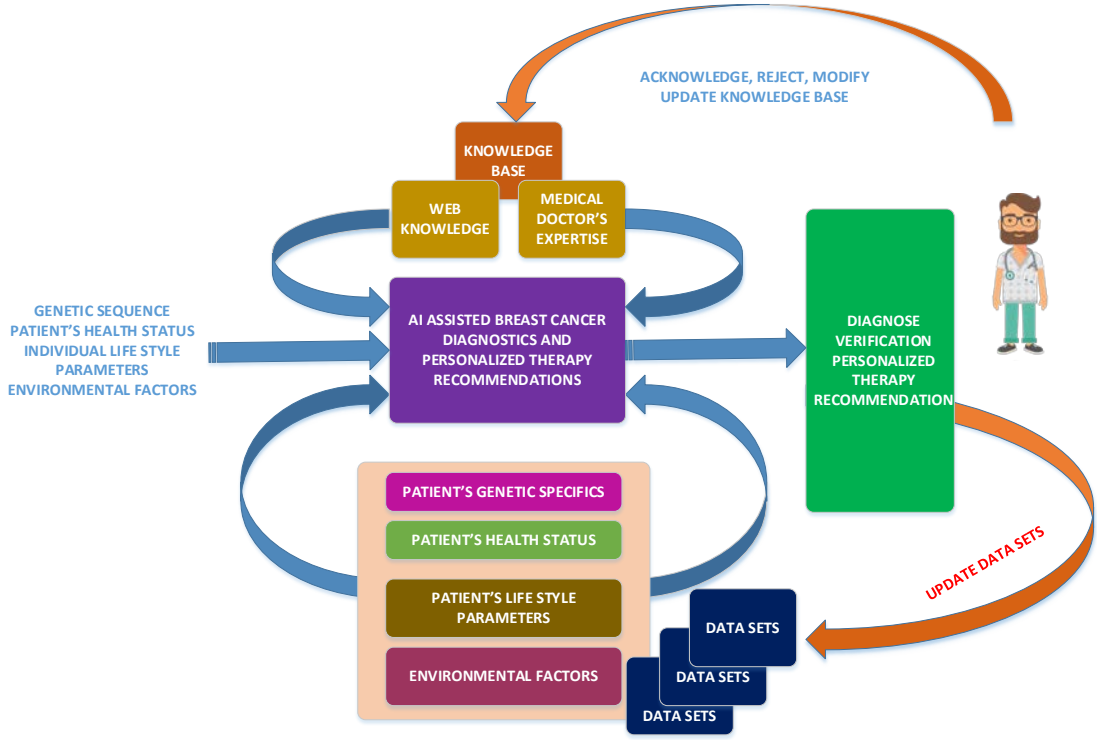

FIGURE 4 The conceptual model of PM Smart Digital Consultant for the case study of breast cancer.

The basic functional units in both phases are scientific analytics workflows – bundles of differentiated workflows in the ML phase, and integrated workflow in the operational phase, built upon optimal differentiated workflows stored in the best models repository. The applicability of the method has been illustrated by the presented conceptual model of smart digital consultant for personalized breast cancer diagnostics and therapy recommendations deploying the suggested method for in silico KDD. The method has been verified and validated for the case studies of differentiated descriptive analytics workflows for prokaryotic and eukaryotic gene finding and mapping and differentiated diagnostics analytics workflows for breast cancer associated gene mutation detection. Future work encompasses the design and implementation of digital smart consultant on breast cancer software based on the suggested intelligent method for Adaptive In-silico Knowledge Discovery (Fig. 4).

ACKNOWLEDGMENTS

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 ДН07/24, financed by the National Science Fund, Competition for Financial Support for Fundamental Research – 2016, Ministry of Education and Science, Bulgaria.

REFERENCES

1. The Precision Medicine Initiative https://obamawhitehouse.archives.gov/node/333101
2. Genetics Home Reference https://ghr.nlm.nih.gov/primer/precisionmedicine/definition
3. The Forth Paradigm "Data-Intensive Scientific Discovery" https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/
4. In silico Oncology and In Silico Medicine Group http://in-silico-oncology.iccs.ntua.gr/
5. http://www.breastcancer.org/risk/factors/genetics
6. http://www.breastcancer.org/pictures/types
7. http://www.breastcancer.org/treatment/planning/types_treatment/diff_treatments
8. https://www.personalgenomes.org/
9. https://pgp.med.harvard.edu/
10. https://personalgenomes.ca/
11. https://www.personalgenomes.org.uk/
12. http://genomaustria.at/das-projekt/
13. http://pgpchina.org/
14. P. I. Borovska, "In Silico Technologies and the Fourth Paradigm for Scientific Research", in "*In-Silico Intellect" Scientific Journal*, (Association "Innovation Center for Information and In-silico Technology and Expert Knowledge Transfer – In-silico Intellect", Sofia, 2017) vol. 1, No1, 5-12, ISSN 2534-8531 https://insilicojournal.com//wp-content/uploads/2018/02/journal-1-2017.pdf
15. P. I. Borovska, "Big Data Analytics and Genetic Research", Proceedings of International Conference "Big Data, Knowledge and Control Systems Engineering – BdKCSE'2017", (Bulgarian Academy of Science, Sofia, Bulgaria, Dec. 2017), 1-8
16. P. I. Borovska, "Big Data Analytics and Internet of medical Things Make Precision Medicine a Reality", Plenary lecture, 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. Apache Spark https://spark.apache.org/
18. S. de Avilla, Silva and Sergio Echeverrigaray, Bacterial Promoter Features Description and Their Application on E. Coli In Silico Prediction and Recognition Approaches, 2012, DOI: 10.5772/48149
19. Araceli Huerta-Moreno, Method to predict promoters recognized by the alternative sigma factors in E. coli K12, 2011, http://www.ccg.unam.mx/Computational_Genomics/PromoterTools/
20. M. Abbas, M. Mohie-Eldin, Y. EL-Manzalawy, Accessing the Effects of Data Selection and Representation on the Development of Reliable E. Coli Sigma 70 Promoter Region Predictors, PLoS One 2015; 10(3): e0119721, doi: 10.1371/journal.pone.0119721
21. D. Ivanova, P. Borovska, V. Gancheva, Experimental Investigation of Enhancer-Promoter Interactions out of Genomic Big Data based on Machine Learning, 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
22. P. Borovska, V. Gancheva, Parallelization and Optimization of Multiple Biological Sequence Alignment Software Based on Social Behavior Model, International Journal of Computers, ISSN: 2367-8895, vol. 3, 2018, pp. 69-74, http://www.iaras.org/iaras/journals/ijc
23. 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