Big Data Analytics and Internet of medical Things Make Precision Medicine a Reality

PLAMENKA BOROVSKA
Department of Informatics
Technical University of Sofia
Sofia 1000, 8 Kliment Ohridsky boul.
BULGARIA
pborovska@tu-sofia.bg https://www.researchgate.net/profile/Plamenka_Borovska

Abstract: - In this paper the role of advanced IT technologies such as Big data analytics and Internet of medical Things (IomT) in support and promotion of precision medicine has been revealed. The concept of precision medicine has been presented and analyzed from the point of view of computational science and the new paradigm for scientific research. The focus of the paper is on the intersection of Big data analytics and precision medicine. The computational flow of in silico knowledge data discovery has been presented and analyzed and the beneficial outcomes for the case study of genome mapping based on computer model of RNA revealed. Finally, the beneficial role of Internet of medical Things and related technologies has been discussed.

Key-Words: - Big data analytics, in silico knowledge discovery, Internet of medical Things, bioinformatics

1 The Problem Area

In Digital Era, modern society faces the challenges of advanced knowledge that technology helps us create. Advanced information and communication technologies facilitate the efficiency of scientific research in all areas – life sciences, technology, and humanities. The computational paradigm in scientific research involves computer-based models and simulations (in silico experimentation) that offer greater potential and facilities to investigate then theoretical analysis does. Globally, this resulted in the accumulation of huge amounts of in silico experimentation data that can be subjected to analysis in order to extract value. The fourth scientific research paradigm – Data Intensive Science Discovery - revolutionized fundamental and applied science research [1]. Innovative modern technologies, such as Big data analytics, Internet of Things, cloud computing, give researchers powerful opportunities for Knowledge Data Discovery and intelligent decision making [2, 3].

Precision medicine [4] has been one of the hottest topics nowadays and involves disease treatment that takes into account individual genetic profile, environmental specifics, and lifestyle of the individual. Knowledge Data Discovery (KDD) solutions are crucial for the detection of complex DNA anomalies implicated in genetic diseases and cancer. The capacity of DNA databases [5] is exponentially growing for the last decade. Precision medicine starts with genomics and relies on the "omics" platforms for the analysis and interpretation of multi-scale data.

The major sources of genomic data acquisition within the Big genomic data ecosystem are the new DNA sequencing technologies, "omics" data generation, in silico technologies, genome databases, cancer genome databases, and the related technologies Internet of medical Things (IomT) and cloud technologies. Wireless Body Area Networks (IomT) provide continuous health monitoring of patients for applications, such as remote monitoring, biofeedback and assisted living. Leader IT companies such as Amazon's Genomics in the Cloud [6], Google Genomics [7], IBM Watson Health [8], and SAP Health offer healthcare innovation portfolio on their big data platforms and cloud services for infrastructure, software tools and data sets for genomic analyses with the aim of facilitating personalized and precision medicine.

2 The Concept of Precision Medicine

In 2015 president Obama (White House), USA, launched the Precision Medicine Initiative with the aim to improve health and disease therapy by means of tailoring the treatment and prevention strategy to fit the specifics of the individual patient instead of "one-size-fits-all" approach (average patient). "Doctors have always recognized that every patient is unique, and doctors have always tried to tailor their treatments as best they can to individuals",

president Obama stated in his State of the Union address [9].

In the context of cancer treatment, getting the genetic profile is extremely important, because many severe diseases, such as breast cancer, are considered a genetic disorder due to mutations in specific genes.

Personalising patient treatment is a difficult and expensive task and instead, in the case of precision medicine, patients are grouped according to their specifics and are assigned optimal treatment for the target group.

The main objective of the Precision Medicine Initiative is the individualized healthcare. The deployment of precision medicine into patient care involves considering 3 crucial factors: genetic specifics, environmental factors and individual life style.

In order to capture the genetic profile of a wide range of individuals special instruments are utilized - automated DNA sequencers. Next generation sequencing (NGS) technology, also named massively parallel or deep sequencing, gives the opportunity to sequence an entire human genome for a day, however, the prize for this is being still quite expensive. Each nucleotide of the genome (the symbol representing it) is read several times in order to reduce the probability of error. The size of a personal sequenced genome file is about 200 Gbytes. Whole genome sequencing approach gains popularity because more than 90% of DNA regions of clinical importance are actually non-coding regions.

The first human genome was sequenced more than 15 years ago at the price of $3 billion. Nowadays, the prise of sequencing a personal human genome is less than $1000. However, still this prize is considered too high and healthcare community relies on technological innovations that will reduce this price under $100.

According to Nebula Genomics, besides whole genomic sequencing, two other factors are crucial for stimulating precision medicine and developing successful genomic market: (1) easy sharing of acquired personal genomic data and (2) “enhanced data protection”. In order to ensure the necessary personal data protection Nebula Genomics intends to deploy blockchain technology [10].

The Global Network of Personal Genome Projects, (started in 2005), is a coalition of projects worldwide aiming at establishing public genome, health, and trait data. The scope of participating partners comprises: Harvard Medical School, USA (the pilot site of the network), Canada, UK, Astria and China.

Personalized medicine and precision medicine emerge as a result of the intersection of omics technologies and Big Data.

3 How Big Data Analytics and Precision Medicine Intersect

3.1 The computer model of DNA and RNA

Genetic research relies a lot on in silico experimentation. Studies and in silico experiments are based on computer models of the gene. At present, these patterns represent extremely long string combinations (strings of symbols) within which it is difficult to identify individual genes. The problem is compounded by the fact that genes contain both encoding and non-coding information. Non-coding information may contain junk DNA, regulatory elements of genes, multiple copies of genes, including pseudogens, intergenic spaces, and scattered repeats.

The computer model of DNA practically is a string of symbols out of a 4 – letter alphabet (A, C, G and T), each letter representing a separate nucleotide (a nucleotide is the main building component of DNA). The DNA helix is presented as 2 complimentary sequences, however, in most cases, only one of the DNA sequences is subjected to analysis. A triplet of nucleotides is called a codon and each codon is encoding for a single protein (Fig.1.). For the case of RNA the alphabet comprises letter U instead of T.

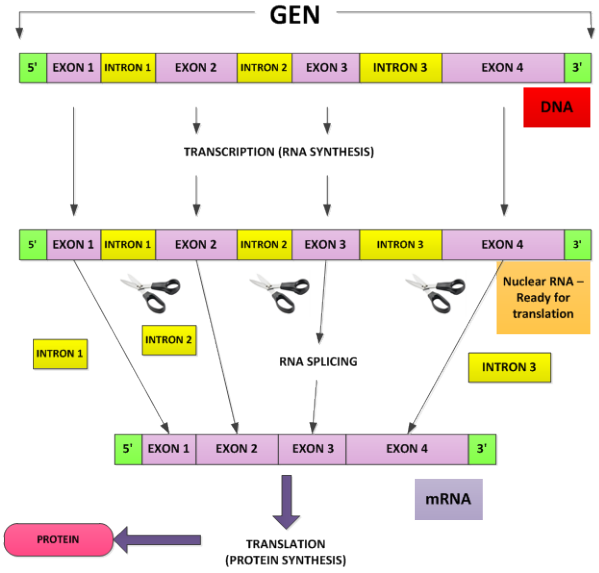

Fig.1. Eukaryotic gene structure.

3.Prokaryotic Gene Mapping Based on Machine Learning

Key problems of genetic research are genome mapping and genome analysis for diseases of genetic disorders. The purpose of gene mapping is to identify separate genes within the genome and consequently, to determine their functionality.

Each gene starts with a fixed combination of 3 nucleotides AUG, named a start (initializing) codon and terminates with stop codon (UAA, UAG or UGA). It is extremely difficult, even impossible to find out which of the numerous combinations of ATG nucleotides is the start codon. The contemporary approach is to conduct gene mapping on the basis of the analysis of intergenic spaces, the focus being on regulatory elements. The difficulty of the problem is the non-consistent varying structure of the regulatory elements.

For the purpose of clarity and simplicity, let's consider the model of RNA for the case study of bacteria E. Coli (Fig.2). For the case study under investigation intergenic spaces include regions rich in A and U. For the case of the consensus structure UAUAAU of the PB box in E. Coli, the probability of each nucleotide to be in the position stated, varies from 50% to 90%. So, the most promising approach is to focus on regulatory elements. The promoters are regulatory elements that activate or deactivate a gene.

The algorithm to find out the start codon of a gene or gene sequence comprises the following steps with the assumption that the start position of the gene is denoted as position 0, the position of each symbol to the left of position zero is marked by (-) and is considered upstream, and vice versa, the position of each symbol to the right of position zero is marked by (+) and is considered upstream:

Step 1: Analysis starts at the first symbol of the computer model of RNA;

Step 2: Starting at position -35 identify two consensus areas of 6 ntds UUGACA downstream, possibly separated by 1 or 2 nucleotides (spacers);

Step 3: Starting from the middle of the second consensus area search for PB area (usually at position -10);

Step 4: After identifying PB area, from the middle of PB shift 10 positions right;

Step 5: From current position identify SD area;

Step 6: From the terminal symbol of the SD area shift right 7 to 9 nucleotides;

Step 7: Start searching for the initial codon AUG;

Step 8: After identifying the start codon AUG, search to identify the terminal codon UAA.

In order to implement this algorithm for genome mapping by machine learning methods, software developers need to build up enormous training sets holding all the possible variations of the promoter region structures within the prokaryotic genome.

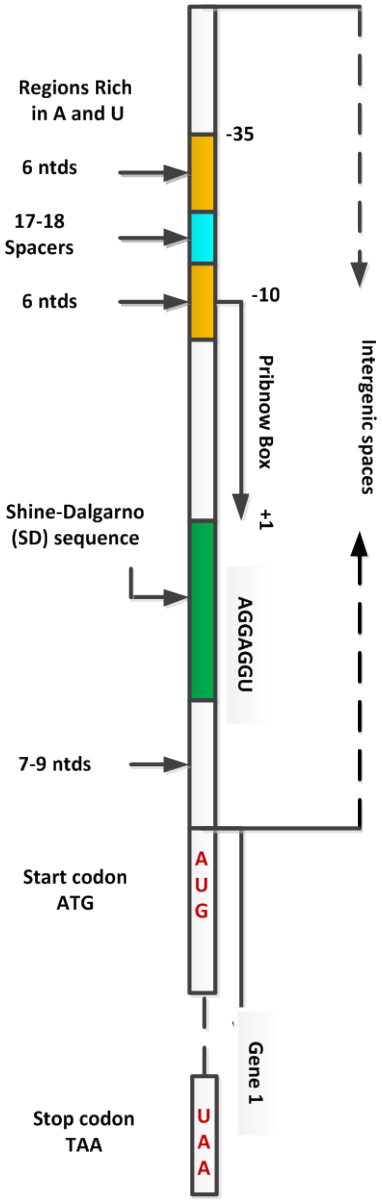

Fig.2. The computer model of prokaryotic RNA for the case study of bacteria E. Coli

3.2 Big Genomic Data Analytics and in silico Knowledge Data Discovery

The intersection of Big data analytics and precision medicine lies in "in silico" knowledge data discovery (in silico KDD) that makes possible the design and deployment of smart digital consultants helpings clinicians and researchers to make accurate disease diagnostics and prescribing the optimal therapy for an individual patient. The intersection of Big data analytics and precision medicine is shown

in Fig.3. It starts with Big genomic data ecosystem acquiring and accumulating molecular omics data, medical wearables data and in silico technology experimentation data.

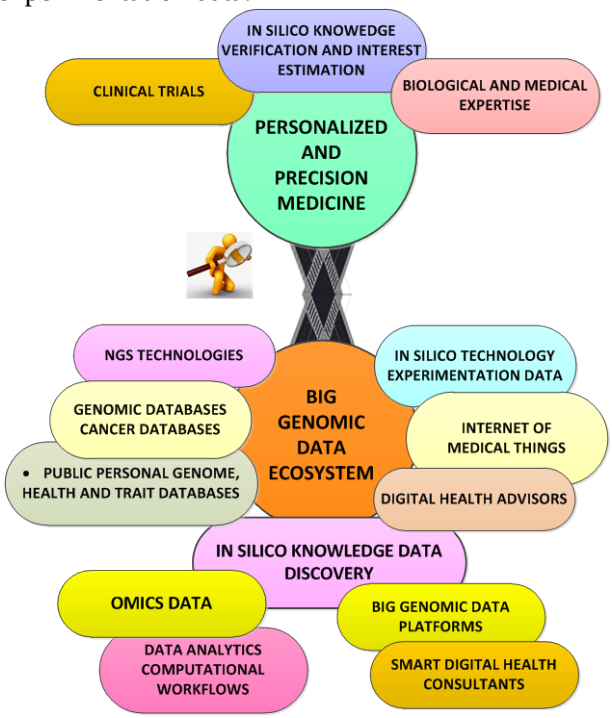

Fig.3. Intersection of Big genomic data analytics and precision medicine.

The computational workflow encompasses the stages of a standard computational pipeline for data analytics and knowledge data discovery: data acquisition and modeling, preprocessing stage (filtering, problem dimensionality reduction, multiple sequence alignment, signal processing), data analytics stage and postprocessing stage including result visualization and interpretation and expert's estimation of interest (Fig.4).

Multiple sequence alignment (MSA) is a popular technology involving aligning thousands of nucleotide or protein sequences with the aim of studying homology evolutionary relationships among the sequences under investigation. It is often used in the stage of pre-processing raw DNA or RNA data within the workflow of in silico KDD.

For the case study of prokaryotic gene mapping, it is recommendable to apply multiple sequence alignment methods to find out the consensus areas of the promoters and even of all the regulatory elements.

For the purpose of in silico knowledge data discovery quantitative as well as qualitative types of data analytics have to be conducted.

Descriptive analysis is conducted by means of data mining tools and in this stage of the computational pipeline the training and validating data sets are being built up. Then the diagnostic analysis is performed, the outcome of which is, for example, finding out the specific genetic disorder. The predictive analysis is following, that prognosticates the progress of the disease. For the case study of cancer, for example, cancer malignancy and life duration are being predicted.

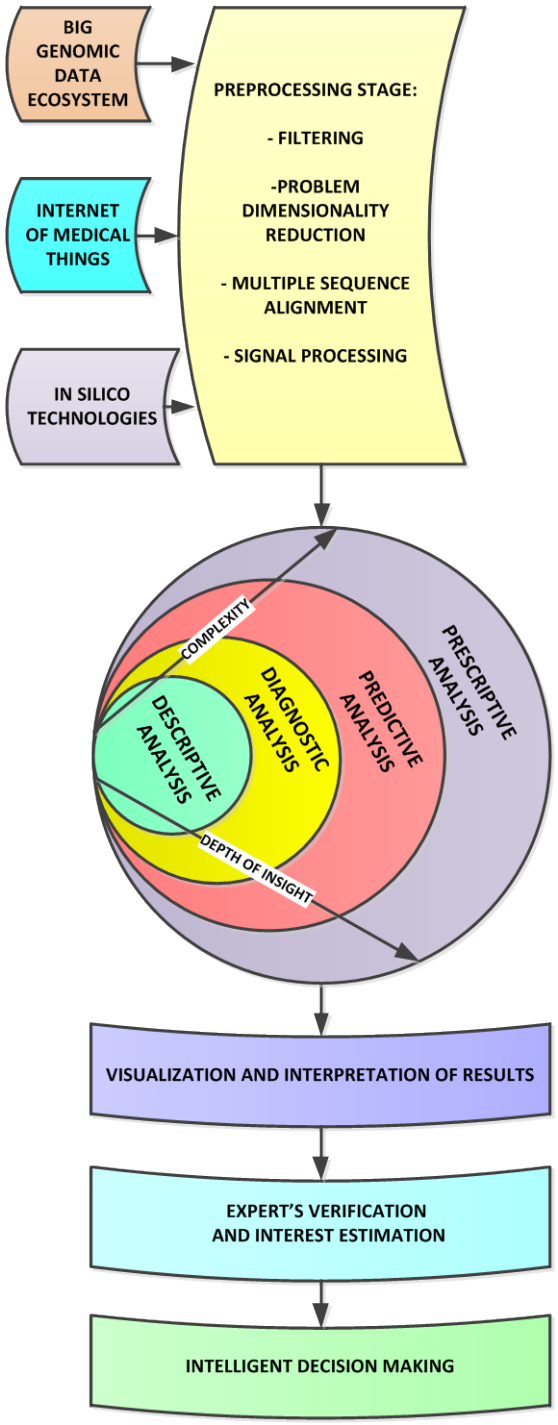

Fig.4. Computational workflow for in silico knowledge data discovery

The final and most important analysis is the prescriptive analysis, based on optimization methods, the issue of which is the individual therapy for the patient. The final stage comprises visualization and interpretation of the results obtained and the discovered knowledge is subjected to expert's verification and interest estimation to give answer to the question if the discovered knowledge is significant or of minor interest. Afterwards follows the stage of intelligent decision making i.e. decision taking has to be fully automated thus overpassing the potential of Decision Support Systems. In the context of intelligent decision making, lately, IBM Watson for Oncology launched a cognitive computing tool (2017) that offers meaningful decision support capabilities.

4 The Role of Internet of medical Things for Precision Medicine

Internet of medical things is an interconnected infrastructure of mobile devices "embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data" [12]. One of the most successful solutions for healthcare industry in the IomT-enabled infrastructure is the mobile applications for remote health monitoring system i.e. the Digital Health Advisors that facilitate communications between patients and doctors over a secured connection.

Wireless body area network (WBAN), named also body sensor network (BSN), actually make possible to acquire and accumulate personal medical data out of wide spectrum wearable devices (physiological biosensors), embedded in or on the surface of the body, or suitable for wearing in clothes, bags, etc. (Fig.5). The data acquired in-situ about the physiological status of a person is transmitted via Internet and thus is made accessible in reasonable time and in a secure way to doctors no matter of the patient's allocation.

The concept of interconnecting and remote monitoring medical imaging equipment over the Internet dates back to approximately 20 years ago. The Big Three medical imaging companies Siemens Healthcare, Philips Healthcare and GE Healthcare together built up a strategy for establishing "All-in-one Health Cloud" in 2015 [11]. The major aim is to move computer-intensive image processing to the Health Cloud ecosystem.

According to HealthIT Analytics (Intelligent Network Media) imaging analytics is the "first step to personalized medicine [13]. Medical imagery produces and accumulates huge amounts of X-rays, cardiographics, ultrasounds and other images helping in diagnostics of individual patient's disease. Crucial aspect is complete automation of medical imagery and analytics process to support high precision of diagnosis [14].

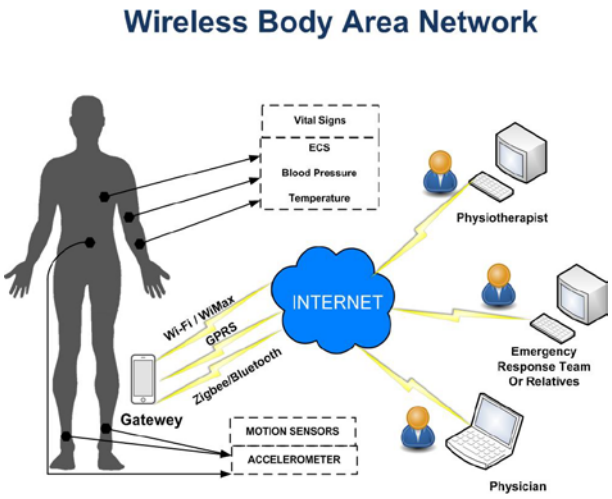

Fig. 5. Body Area Network.

Signal processing is playing an increasingly important role in modern times, mostly due to the ever-increasing popularity of IomT devices. Analysis of IomT data in medicine helps doctors to make up a reliable and accurate disease diagnosis and prognosis, for predicting disease progression and the effect of treatment as well as for drug target identification.

4 Conclusion

In this paper the role of advanced IT technologies such as Big data analytics and Internet of medical Things (IomT) in support and promotion of precision medicine has been revealed. The concept of precision medicine has been presented and analyzed from the point of view of computational science and the new paradigm for scientific research. The focus of the paper is on the intersection of Big data analytics and precision medicine. The computational flow of in silico knowledge data discovery has been presented and analyzed and the beneficial outcomes for the case study of genome mapping based on computer model of RNA revealed. Finally, the beneficial role of Internet of medical Things and related technologies has been discussed.

Acknowledgements
The 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" 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] Xiaolong Jin, B. Waha, X. Cheng, Y. Wang, Significance and Challenges of Big Data Research, Big DataResearch 2 (2015)59–64, www.elsevier.com/locate/bdr
[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung, Big data: the next frontier for innovation, competition, and productivity, Tech. rep., McKinsey Global Institute, 2011, available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
[3] Edmon Begoli, A Short Survey on the State of the Art in Architectures and Platforms for Large Scale Data Analysis and Knowledge Discovery from Data, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.468.8676&rep=rep1&type=pdf
[4] What is precision medicine? http://learn.genetics.utah.edu/content/precision/intro
[5] GenBank https://www.ncbi.nlm.nih.gov/genbank/
[6] Amazon's Genomic in the Cloud https://aws.amazon.com/health/genomics/
[7] Google Genomics Platform https://cloud.google.com/genomics/
[8] IBM Watson Health – Value Based Care https://www.ibm.com/watson/health/value-based-care/
[9] https://obamawhitehouse.archives.gov/node/333101
[10] Grishin D., K. Obbad, P. Estep, M. Cifric, Y. Zhao, G. Church, Nebula Ganomics: Blockchain-enabled genomic data sharing and analysis platform, 2018, www.nebulagenomics.io/assets/documents/NEBULA_whitepaper_v4.52.pdf
[11] The Global Network of Personal Genome Projects http://www.personalgenomes.org/
[12] Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M. Internet of things for smart cities. IEEE Internet Things J. 2014;1(1):22–32
[13] https://healthitanalytics.com/news/imaging-analytics-the-first-step-to-personalized-medicine
[14] http://www.diagnosticimaging.com/pacs-and-informatics/internet-medical-imaging-things-here