Big Data Analytics and Genetic Research

Plamenka Borovska
Technical University of Sofia
8, Kliment Ohridksy boul. , Sofia, Bulgaria
pborovska@tu-sofia.bg

Abstract: This paper focuses on the potential of the innovative Big data technologies to facilitate genetic research for the case study of gene mapping. The problem of gene mapping has been formulated on the basis of the computer model of DNA and from the point of view of computer based technologies. Conceptual model of the Big genomic data ecosystem has been suggested and the relevant most popular genomic data platforms revealed. In silico knowledge data discovery pipeline for genome mapping based on promoters has been built up and the functionality of each stage of the pipeline has been defined.

Keywords: Big Genomic Data, in silico technology, in silico knowledge data discovery, bioinformatics

1 Problem Area

The fundamental scientific studies are revolutionized by the big data files and streams. One of the areas of fundamental science, strongly dependent on big data, is molecular and computational biology [1]. In biological sciences there are well established practices of accumulating data in public open access biological databases, which facilitate scientists all over the world in conducting their research. Modern intensive research in bioinformatics stimulates creating innovative methods for processing and analyzes of biological data. In silico technologies, as well as next generation sequencing, resulted in exponential grow of experimental data, facing the new challenges of Big data technologies. Many scientific research teams predict that most analyses for the period till 2025 will encompass astronomy, molecular and computational biology, medicine and meteorology, as directions of fundamental science, strongly dependent and influenced by big data technologies [2].

Genetics is the branch of science concerned with genes, heredity, and variation in living organisms. It seeks to understand the process of trait inheritance from parents to offspring, including the molecular structure and function of genes, gene behaviour in the context of a cell or organism, gene distribution, and variation and change in populations [3]. One of the

major topics of genetic research is gene mapping. Gene mapping may be treated in two major aspects: (1) investigating the way how each gene works and its function, respectively, and (2) study the role that variations in genes play in disease. The results of such studies will lead to many advances in disease prevention and treatment and will stimulate biotechnologies (for ex. bacteria dissolving CO_2) as well as the design of new types of medicines tailored to an individual's unique genetic profile.

The goal of this paper is to investigate the relationships of Big data technologies and genetic research and the potential of big data analytics methods and software tools to help solving the problem of gene mapping in the field of genetic research.

2 The Computational Paradigm for Biology Research and the Problem of Gene Mapping

The computational paradigm for biology research is shown in Fig.1. Exploring the evolution of biology experimentation we can distinguish throughout time several stages such as in vivo, in vitro and in silico experimentation. In silico biology involves developing and implementing computer models and conducting computer based simulations in order to investigate the most important aspects of living organisms and explore biological phenomena. In silico technologies save time and finances as well as the lives of millions of animals in developing and testing new drugs and therapies. The wide spectrum of modern high-performance experimental technologies (genomic, transcriptomic, proteomic and metabolic) generate huge amounts of data facilitating research and innovations.

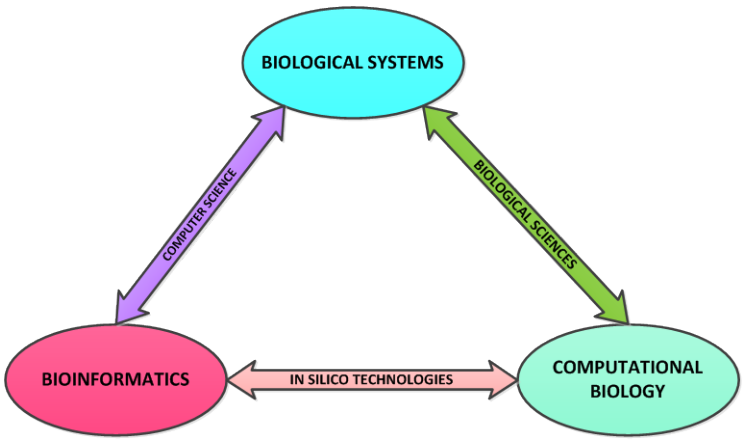

Fig.1. The computational paradigm for biology research

The first fundamental dataset in biology comprise molecular sequences. The advent of the DNA small sequence alphabet (4 nucleotides, as compared to 20 amino acids) stimulated and resulted in the comlete automation of genetic research. DNA is built up of 4 components, called nucleotides, denoted by the symbols A, C, G, T, respectively (Fig.2).

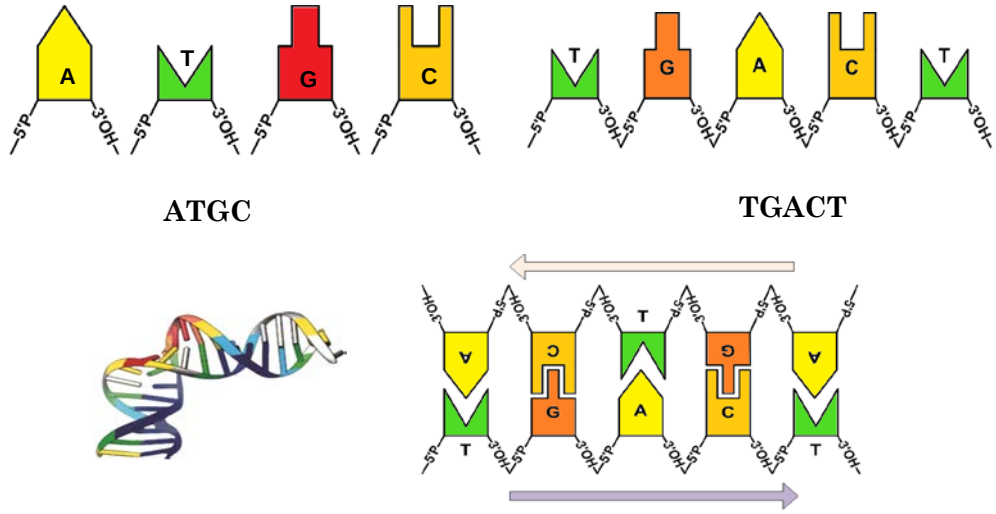

Fig.2. The computer model of DNA.

The first task of functional genomics is to identify and map the genes in new sequenced genome. The human genome was sequenced in 2000, but so far only 10-15% of the genes have been identified and the function of 99% of the human DNA remains unknown [4]. To a greater extend this is valid to thousands of fully sequenced genomes of other organisms. One reason for the backwardness of the functional genomics in comparison with the structural is the lack or the shortage of reliable and efficient methods for analysis of raw genomic information. Such methods can be developed on the basis of genetic regulatory elements. They are perfect base for identification and mapping of genes as they have characteristic structures (nucleotide composition and primary structure) and are directly related to the functions of the genes [5]. The section of DNA that controls the initiation of transcription is called a gene promoter, or a promoter in the field of genetics.

Unknown gene identification and mapping in sequenced genome is fundamental task of functional genomics. Important task of the scientific research in the area of molecular biology is identifying regulatory genetic elements out of sequenced genomes, which will be used for the purpose of unknown genes identification and mapping. The initial phase of genomics aims to map and sequence an initial set of entire genomes in order to know all the genes in a

genome, and the sequence of the proteins they encode. The fact that much DNA in large genomes is non-coding is a complicating circumstance. We must have in mind that non-coding DNA include introns in genes, regulatory elements of genes, multiple copies of genes, including pseudo- genes, inter-genic sequences and interspersed repeats.

As far as coding regions are of concern there are several kinds of exons: initial coding exons, internal exons and terminal exons.

The problem of gene mapping can be summarized as follows: considering the DNA coding regions we have to identify the start and the stop codon of each gene within a genome, having in mind that the start codon of a gene is “ATG” and the stop codon has 3 options – “TAA”, “TAG” or “TGA” (Fig. 3).

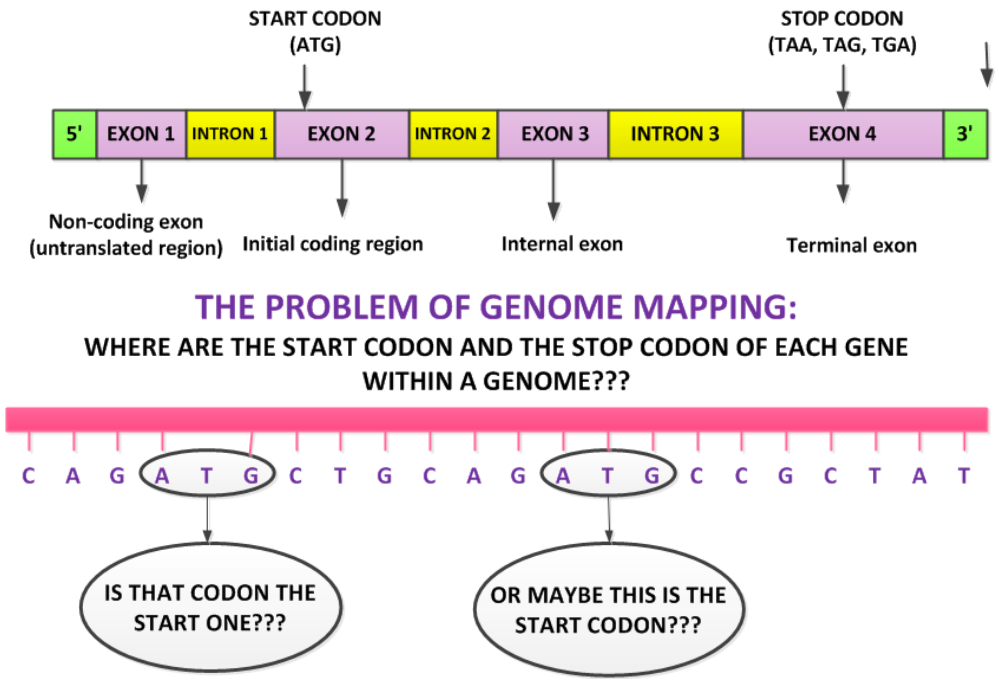

Fig.3. The problem of gene mapping

The important biological questions to answer are [6]:

(1) Which portions of DNA actually do something?

(2) Which portions of DNA actually code for protein or some other product?

(3) Which portions of DNA regulate expression?

(4) Which portions of DNA are used in replication?

3 The Big Genomic Data Ecosystem

Since 2000 medical and biological sciences entered the post-genomics era, which is associated with the emergence and intense development of the new science genomics. The prerequisites for the emergence of genomics worldwide encompass the new technology of complete genomic sequencing, i.e. determining the nucleotide sequence (primary structure) of the entire genomes. This resulted in rapid sequencing of tens of thousands of prokaryotic and thousands of eukaryotic genomes, including the human genome. DNA databases were deployed and their capacity is exponentially growing for the last decade. Genomic data has doubled up every 5 month for the last 8 years [7].

The conceptual model of big genomic data ecosystem is presented in Fig.4. The major sources of genomic data acquisition are the new DNA sequencing technologies, "omics" data generation, in silico technologies, generating huge amounts of in silico experimental data, genome databases, cancer genome databases, and the related technologies Internet of medical Things (IomT) and cloud technologies. Consequently, there is a new challenge - the capacity of genomic databases is growing up faster than the capacity of analytic tools.

Precision medicine is a hot topic nowadays. It starts with genomics and relies on the omics platforms for the analysis and interpretation of multi-scale data. Nowadays, there is a wide spectrum of Big genomic data platforms such as Google Genomics, IBM Reference Architecture for Genomics, SAP® Connected Health platform, etc. These platforms unite the efforts of developers, researchers and healthcare organizations to innovate patient-centered solutions for improving healthcare, reducing costs and providing connected healthcare services.

Google Genomics [8] offers Infrastructure as a Service (IaaS) available through their Cloud Platform, to run large-scale workloads on virtual machines on the pay-per-use principle. The Google web-based Genomics API serves as a gateway to access and use software tools from the Google portfolio of solutions. The software tool BigQuery enables very fast SQL-like queries of massive biological and medical data sets. Based on the MapReduce programming model the platform provides opportunities to discover co-relationships in genomic data through machine learning and other Knowledge Data Discoveries (KDD) methods.

Amazon's Genomics in the Cloud offers infrastructure, software tools and data sets for genomic analyses with the aim of facilitating personalized medicine [9].

IBM® end-to-end reference architecture defines the most critical capabilities for genomics computing: Data management (Datahub), workload orchestration (Orchestrator),

and enterprise access (AppCenter). It can be deployed with various infrastructure and informatics technologies obeying 3 main principles: software-defined, data-centric and application-ready [7].

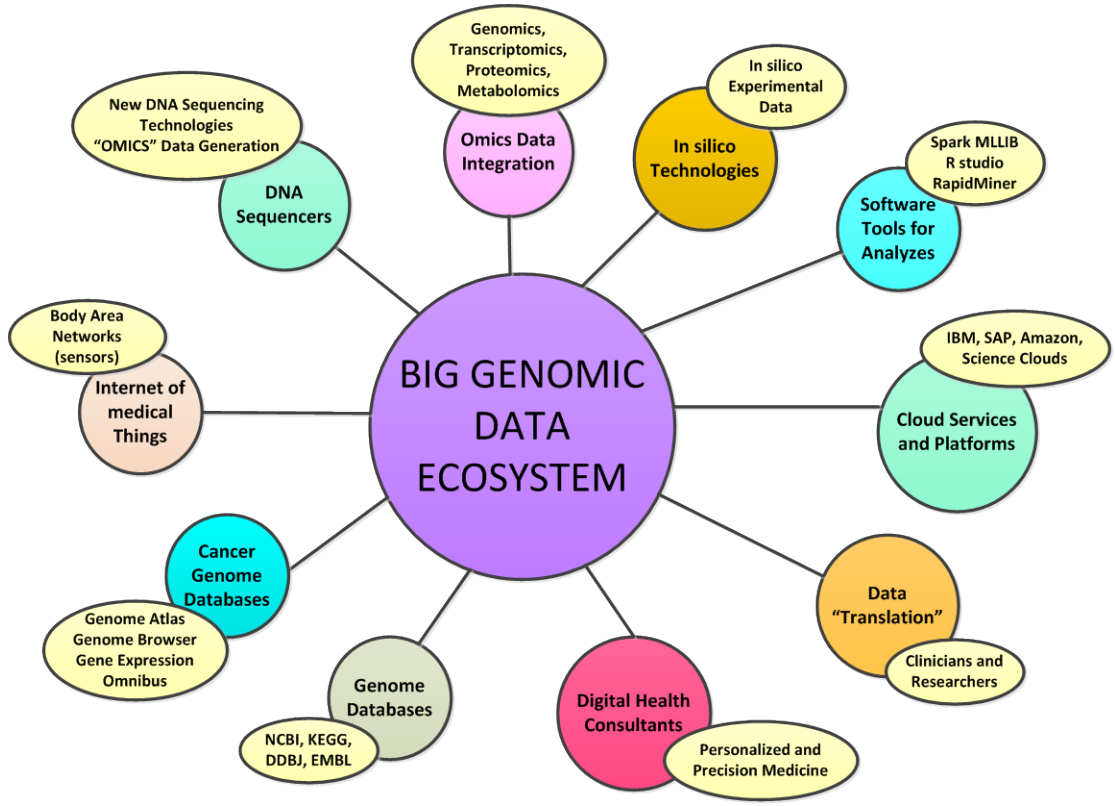

Fig.4. Conceptual model of the big genomic data ecosystem

In 2015, Intel and the Oregon Health and Science University launched a joint project - the *Collaborative Cancer Cloud* [10], actually a high-performance analytics platform accumulating and integrating private medical data targeted for cancer research. Intel intends to establish federated cloud network to other institutions, extending research to Parkinson's disease.

4 In Silico Knowledge Data Discovery for the Case Study of Gene Mapping

The major goal is to conduct scientific research in the area of molecular biology for identifying regulatory genetic elements (promoters) out of sequenced genomes, which will be used for the purpose of unknown genes identification and mapping. The idea is to extract knowledge out of genomic data of a specific organism for identifying the gene promoters and, consequently, to be able to map the genes within the genome [5].

The in silico knowledge data discovery pipeline for genome mapping based on promoters is shown in Fig.5. The first stage of the knowledge discovery process is descriptive analysis (data mining) as a result of which 2 fundamental sets are being built up: training set and validating set. After the stage of the diagnostic analysis (identifying the promoters) we come to the knowledge discovery – the actual gene mapping within the genome based on the identified promoters. Then there is the stage of predictive analysis – gene function prediction – that builds up the hypothesis followed by the stage of visualization and interpretation.

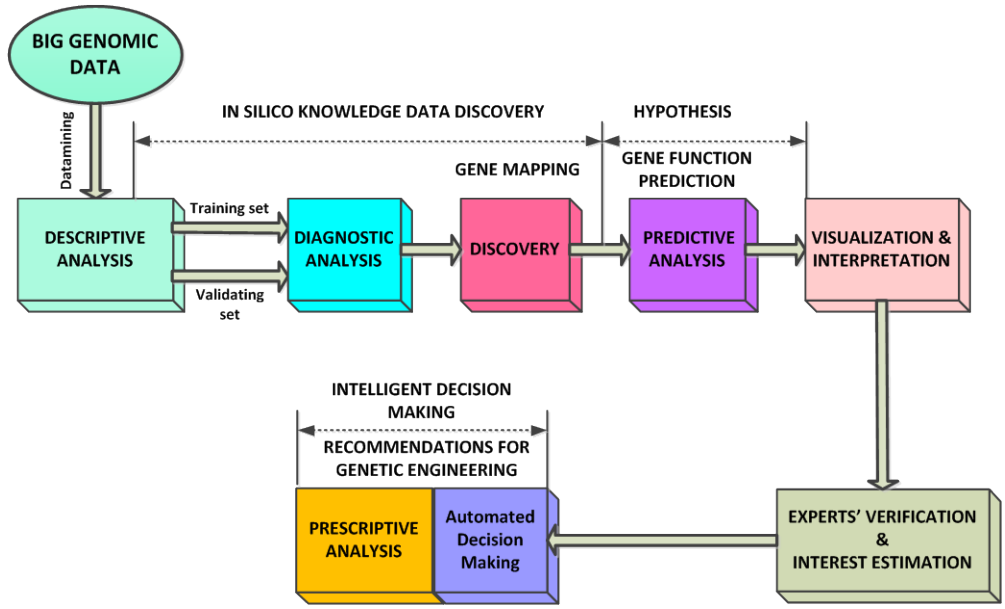

Fig.5. In silico knowledge data discovery pipeline for genome mapping based on promoters

The results obtained are verified obligatory by experts in molecular biology and genetics and the interest of the discovered knowledge has to be estimated. The final stage of pre-

scriptive analysis involves intelligent methods for automated decision making based on metaheuristics and involves recommendations for genetic engineering.

5 Conclusions and Acknowledgements

In this paper the potential of the innovative Big data technologies to facilitate genetic research for the case study of gene mapping has been revealed. The problem of gene mapping has been formulated on the basis of the computer model of DNA and from the point of view of computer based technologies. Conceptual model of the Big genomic data ecosystem has been suggested and the relevant most popular genomic data platforms revealed. In silico knowledge data discovery pipeline for genome mapping based on promoters has been built up and the functionality of each stage of the pipeline has been defined. Future work will involve implementation of the in silico knowledge data discovery pipeline for genome mapping based on promoters and deployment of the relevant workflows.

This work is part of scientific research project "Intelligent Method for Adaptive In-silico Knowledge Discovery and Decision Making Based on Analysis of Big Data Streams for Scientific Research", ДН-07/24, financed by the National Science Fund, Competition for Financial Support for Fundamental Research – 2016, Bulgaria.

References

[1] 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

[2] Xiaolong Jin, Benjamin W.Waha, XueqiCheng, YuanzhuoWang, Significance and Challenges of Big Data Research, Big DataResearch2(2015)59–64, www.elsevier.com/locate/bdr

[3] http://www.genetics.org/

[4] https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/

[5] B. Bachvarov, K. Kirilov and I. Ivanov *"Codon Usage and Gene Expression in Bacteria"*. In: *Encyclopedia of DNA Research*, Chapter 1 (Eds: Samuel J. Duncan and Patricia H. Wiley), Nova Science Publishers, Inc. NY, USA. (2012) ISBN 978-1-61324-305-3.

[6] https://www.genome.gov/10000715/genetic-mapping-fact-sheet/

[7] http://www.redbooks.ibm.com/abstracts/redp5210.html?Open

[8] https://cloud.google.com/genomics/

[9] https://aws.amazon.com/health/genomics/

[10] https://www.intel.com/content/www/us/en/cloud-computing/ohsu-precision-medical-analytics-video.html