Data-Driven Genomic Computing: Making Sense of Signals from the Genome

(Extended Abstract)

© Stefano Ceri, © Arif Canakoglu, © Abdulrahman Kaitoua, © Marco Masseroli, © Pietro Pinoli

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB)
Politecnico di Milano – P.za L. Da Vinci 32,
Milano, Italy

Abstract. Genomic computing is facing a technological revolution. In this paper, we argue that the most important problem of genomic computing is tertiary data analysis, concerned with the integration of heterogeneous regions of the genome – because regions carry important signals, and the creation of new biological or clinical knowledge requires the integration of these signals into meaningful messages.

Keywords: genomic computing, high-level data models and languages, data integration.

1 Introduction

Genomics is a relatively recent science. Historically, the double helix model of DNA, due to Nobel prices James Watson and Francis Crick, was published on Nature on April 1953; and the first draft of the human genome, produced as result of the Human Genome Project, was published on Nature in February 2001, with the full sequence completed and published in April 2003. The Human Genome Project, primarily funded by the National Institutes of Health (NIH), was the result of a collective effort involving twenty universities and research centers in the United States, the United Kingdom, Japan, France, Germany, Canada, and China.

In the last 15 years, the technology for DNA sequencing has made gigantic steps. Figure 1 shows the cost of DNA sequencing in the last fifteen years; by inspecting the curve, one can note a huge drop around 2008, with the introduction of Next Generation Sequencing, a high-throughput, massively parallel technology based upon the use of image capturing. The cost of producing a complete human sequence dropped to 1000 US$ in 2015 and is expected to drop below 100 US$ in 2018.

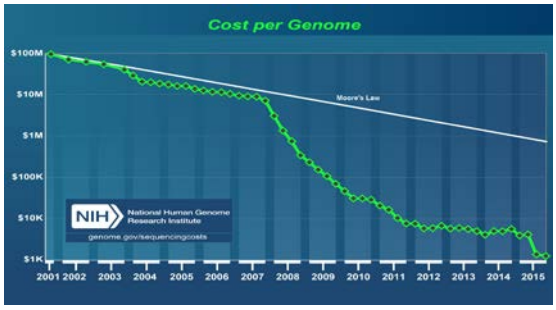

Figure 1 Cost of DNA Sequencing, Source: NIH

Each sequencing produces a mass of information (raw data) in the form of "short reads" of genome strings. Once stored, the raw data of a single genome reach a typical size of 200Gbyte; it is expected [1] that between 100 million and 2 billion human genomes will be sequenced by 2025, thereby generating the biggest "big data" problem for the mankind.

2 From strings to signals

Technological development also marked the generation of new methods for extracting signals from the genome, and this in turn is helping us in better understanding the information that the genome is bringing to us. Our concept of genome has evolved, from considering it as a long string of 32 billions of base pairs, encoding adenine (A), cytosine (C), guanine (G), and thymine (T), to that of a living system producing signals, to be integrated and interpreted.

The most interesting signals can be classified as:

(a) **mutations,** telling us specific positions or region of the genome where the code of an individual differs from the expected code of the "reference" human being. Mutations are associated with genetic diseases – which are inherited; mutations occur on specific positions of the genes – and other diseases such as cancer – which are also produced during the human life, and correlate with factors such as nutrition and pollution.

(b) **gene expression,** telling us in which specific conditions genes are active (i.e. they transcribe a protein) or inactive. It turns out that the same gene may have a big activity in given conditions and no activity in other.

(c) **peaks of expression,** indicating specific position of the genome where there is an increase of short reads due to a specific treatment of DNA; these in turn indicate specific biological events, such as the binding of a protein to the DNA.

These signals can be observed by using a genome browser, i.e. a viewer of the genome. All signals are aligned to a reference genome (the standard sequence characterizing human beings; such sequence is constantly improved and republished by the scientific

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10–13, 2017

community). The browser is open on a window of a given length (from few bases to millions of bases), and the signals are presented as tracks on the browser; each track, in turn, show the signal – either by showing their intensity or just by showing their position. Figure 2 presents a window; the red, blue, and yellow tracks describe gene expression, peaks of expressions, and mutations. The black line indicates the position of (four) genes – these are known information, or "annotations", that can be included in the window. An interesting biological question could be: "find genes which are expressed, where there are three peaks (i.e., peaks representing three experiments are confirmed by all experiments) and with at least one mutation. Such question would, in this specific example, extract the second gene.

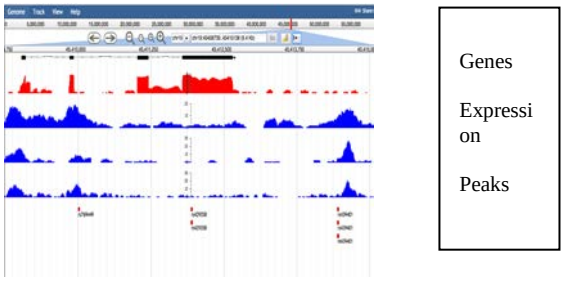

Figure 2 Signals corresponding to mutations, gene expression and peaks as shown on a genome browser

3 Tertiary Data Analysis and GeCo

Signals can be loaded on the browser only after being produced as result of long and complex bioinformatics pipelines. In particular, analysis of NGS data is classified as primary, secondary and tertiary (see Figure 3): primary analysis is essentially responsible of producing raw data; secondary analysis is responsible of extracting ("calling") the signal from raw data and align the signals to the reference genome; and tertiary analysis is responsible of a number of tasks all concerned with data integration.

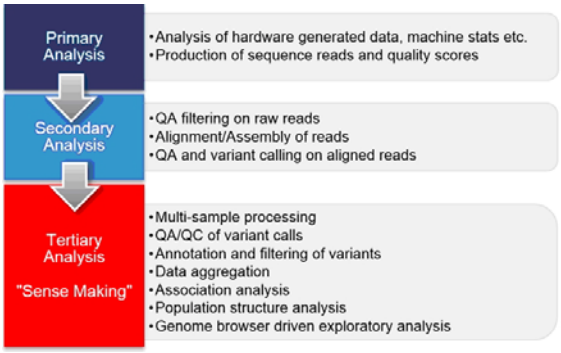

Figure 3 Classification of data analysis for genomics, http://blog.goldenhelix.com/grudy/a-hitchhiker%E2%80%99s-guide-to-next-generation-sequencing-part-2

The bioinformatics community has produced a huge number of tools for secondary analysis. So far, it has not been equally engaged in tertiary data analysis, which is clearly the most important aspect of future research.

Figure 4 shows that the number and variety of tools for secondary analysis. Instead, only four few systems are focused on tertiary analysis.

GeCo is developed by our group at Politecnico di Milano as outcome of an Advanced ERC Grant. GeCo is based on GMQL, a high-level language for genomic data management, and has a system architecture based on cloud computing, implemented on engines such as Spark and Flink [3]. **SciDB**, a scientific database produced by the spinoff company Paradigm4, supports a genomic addition focused on genomic data integration [4]. **DeepBlue**, provides easy access to datasets produced within the BluePrint consortium, with a language which is quite similar to GMQL [5]. **FireCloud,** developed by the Broad Institute, offers an integrated platform supporting cancer research [6]. All these systems already support access to a huge number of open datasets, including ENCODE and TCGA.

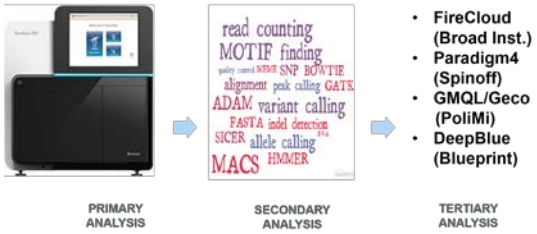

Figure 4 The landscape of genomic computing tools; few of them are dedicated to tertiary data analysis

This two-page abstract has set the stage for discussing why GeCo is an important project in the context of tertiary data analysis for genomics. In the full paper, we will describe some of the aspects of the GeCo project; we will focus on the GeCo API (not been presented so far). This is an important aspect of the project, as it guarantees usability of the system from multiple user and language interfaces, thereby allowing interoperability.

References

[1] Stephens, Z. D. et al.: Big Data: Astronomical or Genomical? PLoS Biol; 13 (7) (2015)

[2] Kaitoua, A., Pinoli, P., Bertoni, M., Ceri, S.: Framework for Supporting Genomic Operations, IEEE-TC, 10.1109/TC.2016.2603980 (2016)

[3] Masseroli, M. et al.: GenoMetric Query Language: A novel approach to large-scale genomic data management. Bioinformatics 31 (12), 10.1093/bioinformatics/btv048 (2015)

[4] Anonymous paper, Accelerating Bioinformatics Research with New Software gor Big Data to Knowledge (BD2K), Paradigm4 Inc. (2015)

[5] Albrecht, F. et al.: DeepBlue Epigenomic Data Server: Programmatic Data Retrieval and Analysis of the Epigenome, Nucleid Acids Research, 44/W1 (2016)

[6] https://software.broadinstitute.org/firecloud