Genomic Data Science

Seydanur Tikir  
Department of Biology  
Georgia State University  
145 Piedmont Ave SE Atlanta, GA 30303  
stikir1@gsu.edu

Keywords:  
Genomic Data  
Next Generation Seq  
Bioinformatics

Background

In 20th century, computational biologists were not concerned about not having powerful computing clusters. The discovery of next generation sequencing technologies introduced of a major threat against genomic data science a decade ago. The genome of a single person is stored as 100 gigabytes of data. Full sequence data has been reported and archived for thousands of humans as well as many thousands of species [1]. As the cost of sequencing decrease, the number of genomes is increasing exponentially [2].

In order to understand genomic variation and biological pathways extensively, it is necessary to analyse millions of genomes. On the other hand, the challenges that we have in Genomic Data Science is increasing as the genomic big data continues growing exponentially. The storage and analysis of big data, as well as the security and privacy issues has been a big concern in Genomic Data Science. On the other hand, the area of genomics provides Data Science with not only great challenges but also great promises.

Problem Statement

A single next-generation sequencer can generate 40 Gb data per day, which reveals a huge raw data along with thousands of sequencers worldwide. On the other hand, storing and analyzing the big data is even more challenging than it is to generate. In next generation sequencing technologies, each base that is sequenced is assigned to quality scores showing
the chance that the base is correct. The size of these quality scores, which are generally captured as image files, is currently much more greater the raw data. The cost of storage of such large data produce a major financial obstacle for small research groups. A key to solve this problem is improvement of ways to record and achieve quality values, which will upgrade the ease and speed of the analyses. Considering that the sequencing is much cheaper than storage, a migration to cloud computing on Genomics era is also a smart way so deal with this issue. [3]

When the challenges of Data Science is considered, the main issues that are considered are analyzing, sharing the gigantic data, as well as problems of speed, cost and hardware regarding this job. Having said that, I think that a bigger problem of Data Scientists is that we are spending our time to figure out how to analyze the data faster, rather than thinking about putting the tools we have into use efficiently. As a bioinformatician who is dealing with a huge size of Next Generation Sequencing Data, I am not able to compare my data with hundreds of other published data easily and regularly. Such an analysis would reveal marvelous discoveries in my research. The question here is “Can we routinely compare our data with the other published bulk?” When I ask that question to myself, the word “routinely” fires a light on my mind. We need to build new tools for automatic analysis and comparisons of the genomics data as new data published. Such an automatic system would allow data scientists to save their time to evaluate already analyzed data rather than analyzing them. Invention of such tools will be a shift to move computation to the data rather than moving the data to the computation.

Another challenge in the world of Big Genomic Data is data sharing and privacy. Genomic Data is generally integrated with different sources, especially in genome-wide association studies regarding medical researches. The public databases that provide human genomic data may store personal information of patients. [4] Though not being publicly available, it is possible to hack these databases. To overcome this issue, suggested policies and procedures should be modified.

The mission of data science needs to be carried through developments on research, education and economics in combination. A smart collaboration is required among data scientists and genomic researchers. There are different kinds of data scientists such as those who ask great questions having a good view of the big picture; and those who better analyze the data with great computation skills. Collaboration between two data scientists from such two groups will reveal more production than the collaboration between a dozen of data scientist from only a single group. In this regard, collaboration between data scientists and wet-lab scientist is of a further importance. I believe organization of workshops and holding discussion groups to produce solutions is a crucial key to handle challenges in Data Science.
Broader Impacts

The area of genomics provides Data Science with not only great challenges but also great promises. If we can achieve a considerable improvement on the issues such as speed, hardware and software problems in Big Genomic Data, we can analyze and compare the genomes much more efficiently. Such an improvement will advance our understanding of genomic variation and biological pathways extensively. The inventions in medicine will chiefly depend on our power to manage and analyze big genomic data.

References


