Chapter 26
Big Data Analysis Techniques for Visualization of Genomics in Medicinal Plants

Hithesh Kumar
Siddaganga Institute of Technology, India

Vivek Chandramohan
Siddaganga Institute of Technology, India

Smrithy M. Simon
Siddaganga Institute of Technology, India

Rahul Yadav
Siddaganga Institute of Technology, India

Shashi Kumar
GenEclat Technologies, India

ABSTRACT

In this chapter, the complete overview and application of Big Data analysis in the field of health care industries, Clinical Informatics, Personalized Medicine and Bioinformatics is provided. The major tools and databases used for the Big Data analysis are discussed in this chapter. The development of sequencing machines has led to the fast and effective ways of generating DNA, RNA, Whole Genome data, Transcriptomics data, etc. available in our hands in just a matter of hours. The complete Next Generation Sequencing (NGS) huge data analysis work flow for the medicinal plants are discussed in the chapter. This chapter serves as an introduction to the big data analysis in Next Generation Sequencing and concludes with a summary of the topics of the remaining chapters of this book.

DOI: 10.4018/978-1-5225-3142-5.ch026

INTRODUCTION

Big Data refers to huge datasets, both qualitative and quantitative, whose analysis is a troublesome task by conventional methods. The analysis of such data is difficult due to its size that leads to the increase in computational space as well as time complexity meaning that the computational space and time required for the analysis increases relative to the size of the data (Chen, Mao, Zhang & Leung, 2014). Big Data can be of any kind, like for the analysis of the search performed by the software giants like Google, (Lazer, Kennedy, King & Vespignani, 2014) Yahoo, Facebook, etc. Various approaches are taken in order to analyze such kinds of data. The approaches include the implementation of algorithms; one majorly used is Machine Learning Algorithm. Other approaches include breaking the problem into different categories and solving independently in order to deal with the problem more efficiently (Weber, 2015).

In the context of Biology, it, however, relates to any kind of data related to any organism; its Sequence, Structure, etc. Big Biological Omics data is stored in databases like Sequence Read Archive (SRA), European Nucleotide Archive (ENA), DDBJ Sequence Read Archive (DRA), etc. The size of data stored in these databases is very large. For example, the data stored in ENA as of 2010 contained approximately 500 billion rows and assembled sequences comprise of around 50 trillion base pairs. The size of such kind of data ranges from Gigabytes to Terabytes (Fritz, Leinonen, Cochrane & Birney, 2011).

Big Data analysis has been possible since the advancement of Next Generation Sequencing Technologies. The Next Generation Sequencing Technologies has not only drastically increased the rate of Sequencing (Stephens et al., 2009), but also led to the advancements in the analysis of the produced data leading to Big Data Analysis as the data generated is quite large. The analysis of such data led to the revolution in characterization of various lethal and life costing diseases like Cancer, analysis of Mutation in genomes, etc., among others. The analysis of such Big Data can be performed in a matter of a day or even hours. The analysis methodology basically includes the search for the availability of data, quality checking of the data and the analysis as a whole. In the absence of the data required, the initial step is to sequence the particular organism using Sequencing machines (Howe et al., 2008).

The quality of data is then checked using tools like FASTQC. Quality Checking and necessary processing of the data is an important step in the process to obtain a good quality result during further analysis (Mousavi et al., 2014). One such example is the analysis of medicinal plant *Rauwolfia Serpentina* that belongs to the Apocynaceae family. *Rauwolfia Serpentina* is a medicinal plant found primarily in India and some neighboring countries. The use of this plant in the past was basically restricted to Ayurveda and has been used in India since the ancient times (Kline, 1954). It was and is still used in the treatment of snake bites, dysentery as it contains several alkaloids that has anti-venom properties, anti-inflammatory properties, etc. (Hutt & Houghton, 1998). It was also used for the treatment of some other diseases since the plant produces many different alkaloids helpful in the treatment of various other diseases like Dysentery (Chandramohan, Kaphle, Chekuri, Gangarudraiah & Siddaiah, 2015).

BACKGROUND

What Is Big Data?

When a large set of datasets becomes too complex or difficult to deal, such types of data are usually termed as Big Data. The challenges related with this field are Capture, Curation, Storage, Search, Sharing,