Chapter 26
Biological Big Data Analysis and Visualization: A Survey

Vignesh U
VIT University, India

Parvathi R
VIT University, India

ABSTRACT

The chapter deals with the big data in biology. The largest collection of biological data maintenance paves the way for big data analytics and big data mining due to its inefficiency in finding noisy and voluminous data from normal database management systems. This provides the domains such as bioinformatics, image informatics, clinical informatics, public health informatics, etc. for big data analytics to achieve better results with higher efficiency and accuracy in clustering, classification and association mining. The complexity measures of the health care data leads to EHR (Evidence-based HealthcaRe) technology for maintenance. EHR includes major challenges such as patient details in structured and unstructured format, medical image data mining, genome analysis and patient communications analysis through sensors – biomarkers, etc. The big biological data have many complications in their data management and maintenance especially after completing the latest genome sequencing technology, next generation sequencing which provides large data in zettabyte size.

INTRODUCTION

The chapter was initiated by requirement of higher and efficient methodologies to analyze big data in a faster manner. The deficiency has motivated us to investigate the problems in an existing technology and frame a feasible model for this big data analysis. On the other hand, there is a considerable interest in the development of new techniques using dynamic programming algorithms to work faster for bioinformatics methods. High throughput sequencing workflow systems provide easy and cost reduced perspective to genome sequencing with timely detection of functions, accurate and fast solutions for big

DOI: 10.4018/978-1-5225-8903-7.ch026

data in bioinformatics. The table 1 shows the detailed view of the different workflow systems that can support high throughput sequencing technologies which includes a big data incorporated in it for analysis.

Bioinformatics is an interdisciplinary area that deals with the biology, computer and statistics. It involves the major aspects of genomics and proteomics with the genome sequencing, which are very sensitive in nature as representing the individual letter for a single nucleotide in case of DNA sequencing. Since 1970, the biological databases are digitized and their sensitivity factors with efficiency are maintained in a perfect manner but due to the vast amount of increasing data the maintenance aspect and extraction of information from gene expression becomes so complex, thus the big data gives the better results for these problems in an accurate manner. The big data includes the analysis of following major characteristics, viz.

- **Scale of Data:** Representing the high amount in size
- **Streaming Data:** Maintaining the velocity for extraction process
- **Various Data Forms:** Variety in form of data included in database can also be easily analyzed
- **Uncertainty of Data:** Poor and inaccurate data can be identified

These characteristics are applied on the biological data to provide the information efficiently, accurately and in a faster manner by saving enormous time with big data concepts.

Big data is defined by the dimensions volume, variety, veracity, value and velocity. Most of big data management deals with the map reducing paradigm. In past two decades, data has tremendous growth day by day. Large amount of data is generated from various sources in structured form or unstructured form, which is difficult to analyze. The availability of tremendous data in volume paves the way for higher efficient analysis in an active interdisciplinary area like bioinformatics. In a survey, IBM indicates 2.5 Exabyte's (2003) and 2.72 zettabytes (2012-2015) created every day in social media, which is very hard to analyze. Other than these, the massive amount of data in genome analysis also difficult to handle. Most of these big data prefers Hadoop technology to process, which reads the raw sequence and produce

Table 1. High Throughput Sequencing Workflow Systems

Name	Illumina	Solid	Requirements	GUI	CLI	Online	Cloud
Ergatis	yes	yes	Linux, MAC OS X, Windows	yes	no	yes	Yes
Galaxy	yes	yes	Linux, MAC OS X	yes	no	yes	yes
Genboree Workbench	yes	yes	Linux, MAC OS X, Windows	yes	no	yes	Yes
GenePattern	yes	yes	Linux, MAC OS X, Windows	yes	no	yes	No
GeneProf	yes	yes	Linux (it is not tested on Others yet)	yes	no	yes	No
Kepler (bioKepler)	yes	yes	Linux, MAC OS X, Windows; > 1 GB RAM, 2 GHz CPU	yes	no	no	No
KNIME	yes	-	Linux, MAC OS X, Windows	yes	yes	no	Yes
LONI Pipeline	yes	yes	Linux, MAC OS X, Windows	yes	yes	no	No
Moa	yes	yes	Linux	yes	yes	no	No
Tavaxy	yes	yes	Linux	yes	no	yes	Yes
Taverna	yes	yes	Linux, MAC OS X, Windows	yes	yes	no	yes
Yabi	-	-	Linux	yes	yes	yes	yes