Big Data in Bioinformatics

Author : Muniba Faiza

Categories : Cloud Computing

Date : March 13, 2016

With the ever increasing amount of biological data being generated with the advanced tools and techniques, a number of suitable ways have been simultaneously developed to handle this vast amount of data in order to make it presentable, accessible and arranged in a logical order to increase workability with the data. Due to the nature of the data being voluminous, Big Data management methods have shown their capabilities to make the biological data effectively managed both in terms of accessibility as well as cost.

> Big data describes a large volume of data, in bioinformatics and computational biology, it represents a new paradigm that transforms the studies to a large-scale research.

The high-throughput experiments in bioinformatics, and increasing trends of developing personalized medicines, etc., increasing a need to produce, store, and analyze these massive datasets in manageable time. The role of big data in bioinformatics is to provide repositories of data, better computing facilities and data manipulation tools to analyze data.

Parallel Computing is one of the fundamental infrastructures that manages big data tasks. It allows to execute algorithms simultaneously on a cluster of machines or supercomputers. Recently, Google has

proposed the MapReduce novel parallel computing model as new big data infrastructure. Similarly, Hadoop which is an open-source MapReduce package was introduced by Apache for distributed data management and is successfully applied in the field of bioinformatics. Hadoop also provides the cloud computing facilities for centralized data storage and provides remote access to them.

In the field of bioinformatics, the big data technologies / tools have been categorized into four:

1. Data storage and retrieval:

The sequencing data obtained has need to be mapped to specific reference genomes for further analysis. For this purpose, *Cloud Burst*, a parallel computing model is used. Parallel computing model facilitates the genome mapping by parallelizing the short-read mapping process to improve scalability of large sequencing data. They have also developed some new tools such as Contrail for assembling large genomes and Crossbow for identifying SNPs from sequence datasets. Similarly, various tools such as DistMap (a toolkit for distributed short read mapping on a Hadoop cluster), SeqWare (to access large scale whole genome datasets), Read Annotation pipeline (developed by DDBJ, cloud-based pieline to analyze NGS data) and Hydra (for processing large peptide and spectra databases) have been developed.

2. Error Identification:

It is necessary to identify errors in the sequence datasets, so many of the cloud based software packages have been developed to achieve this purpose. For example, SAMQA which identifies errors and ensures that large-scale genomic data meet the minimum quality standards, ART which simulates data for three major sequencing platforms,viz., Sequencing, Illumina and SOLiD, and CloudRS.

3. Data Analysis:

This feature of big data allows the researchers to analyze the data obtained by performing experiments. For example, GATK (Genome Analysis Toolkit) is a MapReduce-based programing framework which is used for large-scale DNA sequence analysis. It supports many data formats (SAM, BAM, and many others), ArrayExpress Archive of Functional Genomics data repository is an international collaboration for integrating high-throughput genomics data, BlueSNP is used to analyze the genome-wide association studies, and many more.

4. Platform Integration Deployment:

Since everyone does not have a good grasp on the computing and networking knowledge, therefore, novel methods are needed to integrate big-data technologies into user-friendly operations. For achieving this purpose, some of the software packages have been introduced. SegPig reduces the technological skills required to use MapReduce by reading large formatted files to feed analysis applications, CloVR is a sequencing analysis package distributed through a virtual machine, CloudBioLinux, and so on.

For further details click here.

Note:

An exhaustive list of references for this article is available with the author and is available on personal request, for more details write to muniba@bioinformaticsreview.com.

Sharing is caring. Spread the love!