Chapter 47
Genome Sequence Analysis in Distributed Computing Using Spark

Sagar Ap.
VIT University, India

Pooja Mehta
VIT University, India

Anuradha J.
VIT University, India

B.K. Tripathy
VIT University, India

ABSTRACT

Integration of Computer Science with Bio Science has led to new field Computational Biology which created an opportunity in speeding up the process of analyzing the Bio-data. DNA sequence analysis especially finding the base pairs that helps in identifying the order of nucleotides present in all living beings, it also helps in forensics for DNA profiling and parenting testing. This sequence analysis has been a challenging task in Computational Biology due to large volumes of data and need of more computational resources. Using a distributed file system with distributed computation of tasks can be one of the solutions to above problem. In this paper, the authors use Spark a query engine for large-scale data processing in analyzing the DNA sequence and extracting the base pairs and also they try to improve base pair extraction with improvised algorithms.

1. INTRODUCTION

Bioinformatics is a branch of science, which represent the hybrid form of computer science and biology. As the digital data is increasing rapidly, the data of medical field is also increasing. Bioinformatics basically helps in maintaining the biological information. Computers store the data and by using different techniques of computer science we analyze the tremendous amount of genetic data information related to molecular biology to find the diseases in the human body.

DOI: 10.4018/978-1-5225-8903-7.ch047

Biology needs computational methods to answer all modern biology questions in realistic way to develop more realistic applications. Today analyzing the millions of data is not easy; to analyze this big data in distributed locality we need computational analysis methods. As finding the oriC and base pair in genome data string is time consuming but with help of computational techniques we can annotate easily. So as result half of the part can be completed by computer professionals and biologists can spend their time and money in other tasks. So in this research paper we are analyzing the genome sequence using the Spark distributed processing platform. We described the process of analysis and different algorithm used for the decoding the hidden message of genome sequence and extract the useful information.

Now Bioinformatics is using in every field including machine learning, data mining, pattern recognition algorithm and visualization. In genetic research it used in sequence alignment, genome assembly drug design, gene finding, drug discovery, gene expression and structure etc.

1.1. Application of Bioinformatics

The tremendous masses of genomic information produced by elite innovations are difficult to handle without a parallel improvement in computational assets empowering the capacity, administration and investigation of genomic data. Bioinformatics has gained a central part in the genomic time. The millions DNA arrangements sections delivered by new era sequencers are sorted and gathered with advanced bioinformatics programming. Once the sequences are collected, then it's an ideal opportunity to comprehend it. Explanation programming hunts down practical signs in the genomes to induce coding qualities in the succession and other kind of utilitarian non-coding sequences. There are different applications of bioinformatics as follows:

- Drug discovery;
- Preventive medicines;
- Gene therapy;
- Personalized medicines;
- Microbiology;
- Agriculture;
- Waste cleanup;
- Antibiotic resistance;
- Gene prediction and genome annotation;
- Human microbiome.

1.2. DNA

DNA (deoxyribonucleic acid) stores information in the form of four chemical bases: Adenine (A), Thymine (T), guanine (G), Cytosine (C). Genome is the repeating sequence of these four bases. Human DNA has more than 3 billion bases and 99 percent bases are same in most of the people. The sequence of those bases help in extracting the information from genomes.

The structure of DNA consist pair of A with T and C with G, which forms base pair. Each base merged with a phosphate molecule and sugar molecule this whole sum unit calls as Nucleotide. Nucleotides are set up in two strands that looks like spiral called as Double Helix. The structure of DNA shown in Figure 1.