Chapter 3
Fragment Assembly Based Fast and Optimal DNA Sequencing

Raja G.
Koneru Lakshmaiah Education Foundation, India

Srinivasulu Reddy U.
https://orcid.org/0000-0002-6478-3839
National Institute of Technology Tiruchirappalli, India

ABSTRACT

Growth of healthcare systems has resulted in growth of personalized medicine. Genome sequencing is one of the major players that can enable personalized medicine. The huge computational requirement of this process has made this facility costly and unaffordable for many. DNA sequencing methods that can be performed at computationally low cost and with better performance are sought. The first model presents particle swarm optimization (PSO) and cuckoo search (CS) based models and analyzes their performance levels on sequencing DNA. The sequence assembly is performed using particle swarm optimization (PSO) and cuckoo search (CS). The work then analyzes the pros and cons of using PSO and CS to determine the most effective model. The second method presents the approximate matching model for DNA sequence assembly. The third technique proposes a MapReduce based highest exact matches which successfully exploits and maps between DNA sequences using parallel index method.

DOI: 10.4018/978-1-6684-6523-3.ch003

INTRODUCTION

Advanced technologies for leveraging data have resulted in generation of huge amounts of data, in which biological data also plays a major role. The sheer size of biological data being generated has made the processing abilities of the traditional data processing systems null and void. They require huge amount of processing and since processing and time are directly proportional to the cost, the cost of processing biological data is very large. Bioinformatics is an interdisciplinary field that works based on methods and software tools helpful for understanding biological data. This is an interdisciplinary field of science, which combines computer science, mathematics, statistics and engineering to study and process biological data.

The field of bioinformatics pursues to provide tools and analyses that facilitate better understanding of the molecular structures, by analyzing and correlating genomic and proteomic information. As increasingly large amounts of genomic information, including both genome sequences and expressed gene sequences, becomes available, more efficient, sensitive, and specific analyses become critical. Specifically, the area of bioinformatics includes Next Generation Sequencing (NGS), Virtual Screening, Genotyping, SNP (Single Nucleotide Polymorphism) discovery, Gene expression and Proteomics. Hence computationally, the following functionalities such as, Sequence mapping, Sequence analysis, Peak caller for CHIP-sequencing data, Identification of epistatic interactions of SNPs and Drug discovery can be carried out.

The field of bioinformatics has shown huge growth, and with it the sequencing techniques. The parallel sequencing technologies are known as the Next-Generation Sequencing (NGS). NGS techniques produce high-throughput genome fragments from the input DNA. These sequences are usually of short lengths. Ordering of the nucleic acid molecules in DNA can be used to uncover vital information that can effectively depict a person's hereditary properties. Measuring these sequences can aid in effective identification of anomalies contained in DNA and curing disease. Genome Sequencing is act of determining the nucleotide sequence of given DNA molecules from a short segment of a single molecule, such as a regulatory region or a gene, up to collections of entire genomes.

Although substantial reduction of cost could be observed in the genome sequencing domain, the costs are still considerable and beyond the reach of a common man. Variations in the genome sequencing costs are shown in Figure 1. Reducing these costs is of great interest, as they impact the scope and scale of genomic projects (Chial, 2008). Lowered costs can also lead to more genome