Abstract

DNA (Deoxyribose Nucleid Acid) is a series of nucleotide acid proteins that exist in the organism body where DNA will be identical with inheritance. DNA in the organism body is in pairs, so biological analysis isneeded to match the similarity between the DNA data. DNA data is too large, so Big Data can be one of the answers to compute the similarity matching of data. Big Data is a technique that can do large-scale computation by having several framework that support searching for biology sequence similarities. Hadoop is a framework which is very appropriate for running Big Data with has a lot of tools that can be used. To do processing on DNA's data, Bioinformatics is here with some technique that can be used. One of the technique that can be used is MSA (Multiple Sequence Alignment) where one of the algorithms has a very high accuracy value, the T-COFFEE (Tree Based Consistency Objective Function for Alignment Evaluation) algorithm. T-COFFEE is an algorithm for multiple sequence which is very suitable for finding similarities in DNA data by focusing on very high accuracy values. Besides having a high accuracy value, T-COFFEE requires a very long time to process it so if T-COFFEE implemented in large amounts of DNA's data will be required a long time. Implementation of T-COFFEE on hadoop parallelization car reduce the time that T-COFFEE needed to computed with had the best speedup than not used hadoop parallelization.

Keywords: DNA, Big Data, MSA, T-COFFEE, hadoop paralelization