BAB I
Pendahuluan

1.1 Latar Belakang

Deoxyribose Nucleid Acid atau yang biasa disebut DNA adalah asam nukleotida dalam bentuk heliks ganda yang mengandung instruksi genetik yang menentukan perkembangan biologis dari seluruh bentuk kehidupan sel[1]. DNA biasa identik dengan pewarisan sifat, sehingga DNA sering digunakan untuk masalah-masalah yang berkaitan dengan genetika. DNA berbentuk polimer panjang nukleotida, mengkode barisan residu asam amino dalam protein dengan menggunakan kode genetik, sebuah kode nukleotida triplet[2]. DNA manusia terdiri dari sekitar 3 miliar pasangan basa dengan genom pribadi yang mewakili sekitar 100 gigabytes data, setara dengan 102.400 foto[3]. Satu jenis DNA manusia saja memiliki data sebesar itu. Sedangkan jenis manusia beragam sehingga DNA antar manusia pun berbeda. Dapat dibanyangkan akan sebesar apa data DNA yang akan digunakan hanya untuk mengetahui jenis manusia saja. Maka dari itu dibutuhkannya sistem-sistem dengan performansi yang handal untuk mendapatkan nilai akurasi yang sangat akurat dalam pengolahan datanya.

Saat ini, data-data biologi yang besar termasuk data DNA sangat mahal untuk disimpan, diproses dan dianalisa daripada dihasilkan[3]. Big Data hadir untuk menangani permasalahan pada data yang memiliki skala besar, salah satunya adalah data DNA. Big Data dapat memproses data-data dalam skala yang besar. Untuk memproses data-data tersebut, Big Data memiliki beberapa *framework* yang dapat mendukung pemrosesan data. Salah satu *framework* yang dapat digunakan untuk mengolah data DNA tersebut adalah hadoop. Hadoop merupakan sebuah *software framework* yang memungkinkan pemrosesan data berukuran besar secara terdistribusi dengan melibatkan berkluster-kluster komputer. Hadoop memiliki banyak teknik atau algoritma yang digunakan untuk pemrosesan data yang besar tersebut.

Pemrosesan data-data DNA membutuhkan pemrosesan yang khusus maka dari itu, hadir lah Bioinformatika yang mempelajari berbagai macam keterkaitan antara biologi dan teknologi informatika sehingga dapat menciptakan sebuah kajian atau teknik-teknik untuk meningkatkan berbagai macam aspek mulai dari kesehatan, pendidikan, ekonomi, teknologi, dll. Dalam bidang ilmu Bioinformatika, terdapat salah satu teknik yang biasa digunakan untuk pemrosesan DNA yaitu *Multiple Sequence Alignment* (MSA). MSA banyak digunakan untuk evolusi dan analisis filogenetik, homologi dan prediksi struktur domain. Macam-macam MSA sangat banyak seperti MAFFT, ProbCons dan T-COFFEE yang dapat digunakan untuk meningkatkan akurasi pemrosesan data[4].

T-COFFEE merupakan salah satu algoritma MSA yang memiliki nilai akurasi paling tinggi. Akan tetapi, T-COFFEE memiliki kelemahan dalam kecepatan pemrosesannya. T-COFFEE merupakan algoritma yang mengedepankan nilai akurasi dan mengorbankan kecepatan yang dibutuhkan dalam proses pengolahan data, sehingga membutuhkan waktu yang lama[5]. Untuk data berskala besar seperti data DNA maka sudah dipastikan akan membutuhkan waktu yang sangat lama. Oleh karena itu, diperlukannya paralelisasi pada hadoop dengan menggunkanan algortima T-COFFEE sehingga dapat mengurangi waktu yang dibutuhkan dalam pemrosesan data tersebut dengan tidak merubah nilai akurasinya.

1.2 Topik dan Batasannya

Berdasarkan latar belakang yang telah dijabarkan, rumusan masalah dalam penelitian ini diantaranya, yaitu:

1. Bagaimana mengidentifikasi proses paralelisasi hadoop menggunakan T-COFFEE pada data DNA?
2. Bagaimana potensi hadoop yang dapat mendukung teknik MSA untuk komputasi kecocokan DNA?
3. Bagaimana cara mengimplementasikan T-COFFEE pada paralelisasi hadoop?

Adapun batasan-batasan untuk penulisan tugas akhir ini diantaranya:

1. Data yang digunakan dalam penelitian ini merupakan data DNA yang diambil dari NCBI.
2. Framework Big Data yang digunakan yaitu Apache Hadoop.3.1.4 dan Apache Spark.3.0.0.
3. Parameter yang digunakan dalam penelitian untuk dianalisis adalah *execution time* dan *speed up*.

1.3 Tujuan

Tujuan dari penulisan tugas akhir ini diantaranya:

1. Untuk mengimplementasikan algoritma T-COFFEE pada paralelisasi hadoop.
2. Untuk mengukur performansi paralelisasi hadoop menggunakan algoritma T-COFFEE.
3. Untuk mengetahui potensi paralelisasi hadoop menggunakan teknik MSA T-COFFEE.