1. Pendahuluan

Latar Belakang

Asam dioksibirinukleat atau yang secara umum dikenal dengan DNA adalah molekul biologis yang menyimpan berbagai informasi genetika sebuah organisme. DNA merupakan asam nukleat, salah satu molekul yang hakiki bagi makhluk hidup yang ada. DNA umumnya memiliki dua untai polinukleotida yang masing-masing terdiri dari nukleotida dengan salah satu jenis basa nitrogen (guanina, adenine, timina atau sitosina) yang saling terikat menjadi satu rantai [1]. Dari rantai ini didapatkan kombinasi *string* yang berbeda-beda. Kombinasi *string* ini menyimpan banyak informasi mengenai organisme tersebut, contohnya seperti fungsi, keturunan, struktur dan lainnya [1]. Dalam bioinformatika, analisis kemiripan antar sequence dilakukan dengan metode *sequence alignment*. Informasi kemiripan pola sequence membawa kemiripan pada fungsi, keturunan, struktur dan informasi genetik [1]. Algoritma untuk melakukan s*equence alignment* diantaranya yaitu algoritma Needleman-Wunsch yang menghasilkan *global alignment* dan algoritma Smith-Waterman yang menghasilkan *local alignment*. *Global alignment* mencari nilai kecocokan tertinggi dari ujung awal DNA hingga ujung akhir DNA, sehingga hasil alignment akan sama panjang, sementara *local alignment* hanya mencari hasil kecocokan tertinggi berdasarkan bagian DNA tersebut [2].

Untuk kebutuhan membandingkan sequence dalam lebih dari 2 menggunakan metode *multiple sequence alignment* (MSA). Salah satu contohnya Clustal yang menggunakan algoritma Needleman-Wunch pada proses di dalamnya [3], ada juga T-COFFEE (*Tree Based Of Consistency Objective Function For Evaluation Alignment*) proses didalamnya menggunakan gabungan dari algoritma Smith-Waterman dan juga Needleman-Wunsch. Hasil T-COFFEE lebih akurat dibandingkan dengan Clustal karena menggunakan *local* dan *global alignment* [4]. Pada MSA dibangun sebuah *phylogenetic tree* (pohon evolusi) yang menunjukkan kedekatan tiap sekuens DNA yang ada dan merepresentasikan hubungan evolusi dalam organisme tersebut. Namun, algoritma yang digunakan untuk MSA membutuhkan banyak waktu komputasi yang lama [2]. Untuk mengatasi masalah ini, eksekusi algoritma MSA dapat dilakukan secara paralel terdistribusi. Hadoop merupakan sebuah *framework* yang membantu dalam melakukan proses secara terdistribusi dan paralel [5]. Di dalam *framework* Hadoop terdapat sebuah metode bernama *MapReduce* dimana metode ini dapat melakukan proses paralel dari data-data berukuran besar [5]. Algoritma MSA dapat memanfaatkan *framework* Hadoop dan juga metode *MapReduce* untuk mempercepat waktu komputasi walaupun data yang diproses berukuran besar [6]. Dalam melakukan melakukan proses terhadap DNA dapat digunakan *MapReduce* yang bertugas untuk melakukan *sequence alignment* pada tiap baris DNA dari hasil tersebut kemudian dapat digunakan untuk melihat kedekatan hubungan antar baris pada DNA.

Topik dan Batasannya

Berdasarkan hal-hal yang sebelumnya telah dibahas di latar belakang, berikut ini merupakan rumusan-rumusan masalah yang akan dibahas dalam penelitian ini:

1. Bagaimana membangun *phylogenetic tree* menggunakan model komputasi *MapReduce*?
2. Proses mana yang dapat diparalelkan dalam pembangunan *phylogenetic tree*?

Adapun batasan-batasan yang didetapkan dalam menjawab rumusan masalah diatas, yaitu:

1. Menggunakan *framework* Hadoop dengan *MapReduce* dan Hadoop File System (HDFS) .
2. Metode yang digunakan untuk membangun *phylogenetic tree* adalah NJ (Neighbor Joining).

Tujuan

Tujuan dibuatnya tugas akhir ini adalah untuk membangun *phylogenetic tree* dengan menggunakan metode NJ (*Neighbor Joining*) dengan Hadoop *MapReduce* sebagai model komputasi.

Organisasi Tulisan

Pada jurnal ini, terdiri dari lima bab utama, yang pertama merupakan Pendahuluan. Bab kedua merupakan Studi Terkait. Bab ini berisi teori-teori penunjang penilitian yang dilakukan, yang dibahas disini meliputi Bioinformatika, Algoritma-algoritma yang digunakan dalam pemrosesan sekuens biologi, Penjelasan mengenai *framework* Hadoop dan juga model komputasi *MapReduce*. Selanjutnya bab ketiga yaitu Sistem yang Dibangun, pada bab ini dijelaskan mengenai rancangan sistem yang akan dihasilkan. Bab keempat yaitu Evaluasi. Pada bab ini terdiri dari hasil rancangan sistem yang telah dibangun yang dilakukan sesuai rancangan sistem pada bab tiga. Terakhir yaitu bab kelima, berisi tentang kesimpulan dari hasil pengujian sistem yang dibangun dan juga saran untuk penelitian yang akan dilakukan selanjutnya.