BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Sekuensing genom dari banyak spesies memungkinkan ilmuwan untuk mempelajari seluruh perangkat gen beserta interaksinya (Campbell dan Reece, 2008). Dalam satu dekade terakhir para ilmuwan harus melakukan penelitian laboratorium selama 3 tahun untuk menganalisa DNA (Pahadia, Srivastava, Srivastava dan Patil, 2015). Salah satu kasus dari analisa DNA yang membutuhkan waktu dan tenaga dalam skala besar tersebut adalah untuk menganalisa penyakit yang disebabkan oleh pola genom yang berulang atau disebut dengan *genomic repeats* (Edgar dan Myers, 2005) seperti tiga pasang basa berulang yang dapat menyebabkan penyakit dalam kategori *trinucleotide repeat disorders* (Orr dan Zoghbi, 2007).

Upaya dari sekuensing telah menghasilkan banyak sekali data sehingga juga melahirkan bidang baru yang disebut dengan bioinformatika. Dengan data sekuens yang dihasilkan, ilmuwan dapat menganalisa kepentingan biologi dengan penerapan metode-metode komputasi yang memungkinkan analisa jauh lebih efisien dari segi waktu dan tenaga seperti yang dilakukan banyak laboratorium riset hari ini (Liu, dkk., 2015).

Dalam menganalisa masalah *genomic repeats* dilakukan analisa *string matching* atau *pattern matching* yang mana akan dicari sebuah pola dalam sebuah teks yang berukuran besar. Algoritma dasar untuk pencarian string atau pola ini adalah dengan cara mencocokan semua kemungkinan yang terdapat dalam data dari indeks pertama dalam teks hingga habis. Algoritma ini dikenal dengan *Brute Force (Naïve) Algorithm* yang memiliki kompleksitas dengan kemungkinan terburuknya adalah O(mn) yang akan sangat memakan waktu yang lama jika semakin banyak teks yang akan dijadikan objek pencarian string atau pola (Kindhi dan Sardjono, 2015).

Kebutuhan akan pencarian string atau pola dari data yang semakin besar membuat para ilmuwan membuat algoritma yang lebih efisien dari pada algoritma *brute force* yang mencocokkan satu persatu pola dengan teks yang akan mengakibatkan kompleksitas yang besar. Oleh karena itu beberapa algoritma *string matching* dikembangkan seperti algoritma Knut-Morris-Pratt (Knuth, Morris dan Pratt, 1977). Algoritma pencarian string paling terkenal ini akhirnya memberi inspirasi bagi ilmuwan lainnya untuk terus mengembangkan algoritma yang lebih efisien. Salah satu algoritma pengembangan dari Knuth-Morris-Pratt adalah algoritma Ukkonen (Ukkonen, 1985) yang digagas oleh ilmuwan dari Finlandia dan Fast Hybrid Pattern-Matching Algorithm (Franek, Jennings dan Smyth, 2005).

Namun seiring dengan perkembangan zaman dan semakin banyaknya data yang dihasilkan dalam upaya sekuensing juga menuntut para ilmuwan untuk dapat mengatasi permasalahan komputasi dengan data yang lebih besar (Liu, dkk., 2015). Maka dari itu pawa ilmuwan komputer membuat sebuah konsep komputasi paralel atau sistem terdistribusi yang memungkinkan sebuah pekerjaan komputasi dapat diselesaikan oleh banyak *core, node* atau komputer secara bersamaan. Salah satunya adalah konsep MapReduce (Dean dan Ghemawat, 2010) yang menjadi dasar teknologi pencarian Google dalam skala besar dan memungkinkan para ilmuwan juga menerapkan konsep MapReduce untuk berbagai kasus penelitian.

Contoh lainnya adalah dengan *Package High-Performance Computing* pada bahasa pemrograman R yang beberapa di antaranya adalah Rmpi (Yu, 2002) dan pbdMPI (Chen, Ostrouchov, Schmidt, Patel dan Yu, 2012) yang mengembangkan *parallel computing* dengan *MPI (Message Passing Interface)* pada bahasa pemrograman R. Beberapa contoh selanjutnya adalah randomForestSRC (Ishwaran dan Kogalur, 2007), dclone (dclone: Data Cloning in R, 2010), dll.

Berbagai *package* memiliki prosedur tersendiri dalam penggunannya seperti harus adanya kompilasi di *prompt/terminal* atau dapat dilakukan di dalam *console* R itu sendiri. Juga dengan konsep pemrogramannya seperti pemecahan data yang akan dianalisa, dsb. Skripsi ini akan mendesain model dan mengimplementasikan algoritma Knuth-Morris-Pratt sebagai algoritma *string matching* terbaik

(S.Vijayarani dan R.Janani, 2017) pada R *Package High-Performance Computing* pbdMPI agar dapat digunakan untuk skala data yang lebih besar di masa yang akan datang.

1.2 Rumusan Masalah

Sesuai latar belakang masalah yang telah diuraikan pada sub bab sebelumnya, maka munculah rumusan masalah sebagai berikut:

1. Bagaimana model *parallel computing* dengan menggunakan algoritma Knuth-Morrit-Pratt untuk pencarian *pattern* pada *string?*
2. Bagaimana penerapan algoritma Knuth-Morris-Pratt dapat bekerja pada R *Package High-Performance Computing* pbdMPI?
3. Bagaimana hasil keluaran dari program yang telah didesain dan diimplementasikan sebelumnya?
4. Bagaimana perbandingan kecepatan, akurasi dari tiap komputasi yang dikerjakan oleh berbagai jumlah *core* dan *iterator* sesuai model yang dibuat dan diimplementasikan?

1.3 Tujuan Penelitian

Setelah diketahui rumusan masalahnya, maka tujuan dari penelitian ini adalah:

1. Merancang model *parallel computing* untuk pencarian *pattern* pada *string* menggunakan algoritma Knuth-Morris-Pratt.
2. Implementasi model pada tujuan pertama pada *Package High-Performance Computing* pbdMPI di bahasa pemrograman R.
3. Melakukan eksperimen dari program yang telah dibuat pada data yang dikumpulkan dari laman FTP Ensembl.
4. Melakukan analisa terkait kecepatan dan akurasi dari hasil eksperimen.

1.4 Manfaat Penelitian

Adapun manfaat penelitiannya adalah sebagai berikut:

1. Mempermudah para peniliti di bidang biologi untuk menganalisa *genomic repeats* pada data yang telah tersedia.
2. Membuat sebuah program dan model komputasi paralel untuk dikembangkan oleh peneliti selanjutnya.
3. Memberikan pengetahuan tentang Bioinformatika, khususnya tentang analisa *genomic repeats*.

1.5 Batasan Masalah

Adapun batasan masalahanya adalah sebagai berikut:

1. Program ini bekerja untuk data dengan format standar NCBI/Ensembl.
2. Data yang digunakan pada penelitian ini adalah contoh sekuens DNA manusia dari publikasi nomor 88 di laman FTP Ensembl yang dapat diunduh di ftp://ftp.ensembl.org/pub/release-88/fasta/homo_sapiens/dna/
3. Jumlah *core* yang digunakan dalam eksperimen penelitian ini adalah 2, 4, dan 8 *cores.*

1.6 Sistematika Penulisan

Pada bagian sistematika penulisan ini akan diuraikan mengenai penjelasan tiap bab.

BAB I PENDAHULUAN

Bab ini menjelaskan mengapa dan bagaimana penelitian akan dilakukan. Diawali dengan latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini menjelaskan tentang teori pendamping atau pendukung untuk melakukan penelitian. Teori yang dijelaskan dalam bab ini yaitu mengenai *genomic repeats,* algoritma Knuth-Morris-Pratt dan pbdMPI.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan langkah-langkah penelitian yang akan dilakukan dimulai dari desain penelitian, fokus penelitian, alat dan bahan yang digunakan untuk penelitian dan yang terkahir adalah metode penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjabarkan hasil penelitian dan eksperimen yang telah dilakukan. Semua pertanyaan mengenai masalah yang diangkat dalam tema skripsi dibahas pada bab ini. Beberapa hal di antaranya adalah tentang proses pengumpulan data, pengembangan model, implementasi sistem, studi kasus, desain eksperimen, dan analisa.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dan saran bagi peneliti selanjutnya dari hasil penelitian yang telah dilakukan.