BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Kemajuan penelitian dalam bidang ilmu kehidupan telah menghasilkan sejumlah besar data biologis yang telah mengantarkan menuju era biologi komputasi (Galib, et al., 2015). Salah satu data biologis ialah sekuens DNA (Reece, 2011). Untuk mendapatkan sekuens DNA dilakukan proses sekuensing, dimana setiap molekul nukleotida dianalisis seluruh perangkat dan interaksinya. Setiap spesies memiliki keunikan sekuens DNA yang berbeda-beda, hal tersebut dapat diakibatkan karena terjadinya mutasi pada sekuens. Didalam sekuens DNA terdapat perulangan pola pendek dari kombinasi empat basa nitrogen (*Adenine*(A); *Guanine*(G); *Cytosine*(C); *Thymine*(T)) yang disebut sebagai motif. Pada bidang biologi komputasi, motif dijadikan sebagai bahan penelitian karena motif mengandung makna tertentu yang dapat merepresentasikan DNA (Galib, et al., 2015).

Pencarian motif pada sekuens DNA merupakan salah satu permasalah yang penting dalam bidang bioinformatika, karena hal tersebut dapat membantu ahli biologi untuk memahami lebih baik mengenai struktur dan fungsi dari molekul yang berada pada sekuens tersebut (Ashraf, et al., 2017). Salah satu kasus dari pencarian motif ialah pencarian daerah sekuens DNA yang cocok dengan faktor transkripsi pada proses transkripsi DNA menjadi protein (Davila, et al., 2007).

Permasalahan dalam pencarian motif dapat dikategorikan menjadi 3 jenis, yaitu *Simple Motif Search* (SMS), *Edit distance based* (EMS), *Planted Motif Search* (PMS). Tujuan dari SMS (Rajasekaran, et al., 2005) ialah untuk menemukan semua motif pada semua sekuens dengan panjang 1 sampai panjang yang ditentukan. Sedangkan tujuan dari EMS ialah untuk menemukan semua motif pada jumlah sekuens yang diinginkan (Pal & Rajasekaran, 2015). Kemudian PMS bertujuan untuk menemukan motif yang muncul pada setiap sekuens yang

Tyas Farrah Dhiba, 2018
PLANTED MOTIF SEARCH DALAM SEKUENS DNA MENGGUNAKAN ALGORITMA RANDOM PROJECTION PADA R HIGH PERFORMANCE COMPUTING PACKAGE
Universitas Pendidikan Indonesia | repository.upi.edu | perpustakaan.upi.edu

ada (Martinez, 1983). Pada penelitian ini peneliti akan fokus kepada permasalahan PMS.

Tyas Farrah Dhiba, 2018
PLANTED MOTIF SEARCH DALAM SEKUENS DNA MENGGUNAKAN ALGORITMA RANDOM PROJECTION PADA R HIGH PERFORMANCE COMPUTING PACKAGE
Universitas Pendidikan Indonesia | repository.upi.edu | perpustakaan.upi.edu

Dalam PMS terdapat dua masukan penting yaitu panjang motif yang diinginkan yang disimbolkan dengan "*l*" dan jumlah ketidakcocokan (*mismatches*) yang disimbolkan dengan "*d*" (Miklós, 2016). Semua sekuens masukan akan dibagi-bagi menjadi sub-sekuens (*l*-mers) sesuai dengan panjang *l*. Jika dengan mempertimbangkan nilai *d* pada setiap masukan sekuens terdapat *l*-mers yang sama, maka *l*-mers tersebut termasuk kedalam motif yang tertanam (*planted motif).* Misalkan terdapat tiga sekuens masukan, yaitu s_1 = ATTGCTGA, s_2 = GCATTGAA dan s_3 = CATGCTTG, pada sekuens masukan tersebut akan dicari motif yang tertanam dengan panjang $l = 4$ dan jumlah maksimal ketidakcocokan yang dipertimbangkan $d = 1$, maka motif tertanam yang ditemukan ialah ATTG dan TTGC. Dalam penerapan PMS terdapat dua cara pendekatan yaitu tepat (*exact*) dan perkiraan (*approximate*) (Ashraf, et al., 2017). Pada pendekatan yang tepat, ditemuakan semua kemungkinan motif yang muncul pada semua sekuens, sedangkan pada pendekatan perkiraan motif yang di temukan ialah motif yang memiliki jumlah yang cukup dari semua sekuens yang ada. PMS termasuk kedalam masalah NP-Hard, sehingga jika algoritma ini dijalankan dengan menggunakan pendekatan yang tepat maka waktu dihabiskan untuk mencari semua motif ialah exponensial. Sedangkan pada pendekatan secara perkiraan waktu yang dihabiskan akan lebih sedikit dari pada pendekatan yang tepat, karena pada pendekatan secara perkiraan tidak selalu menemukan semua motif yang ada.

Telah banyak algoritma PMS yang dikembangkan sesuai dengan pendekatan yang ada. Pada pendekatan secara tepat terdapat algoritma PMS1, PMS2, PMS3 (Rajasekaran, et al., 2005), RISSOTO (Pisanti, et al., 2006), PMSPrune (Davila, et al., 2007), PMS5 (Dinh, et al., 2011), qPMS7 (Dinh, et al., 2012) dan PMS8 (Nicole & Rajasekaran, 2014). Sedangkan algoritma yang dikembangkan dengan pendekatan secara perkiraan ialah CONSENSUS (Hertz & Stormo, 1999), MEME (Bailey, et al., 2010), WINNOWER, SP-STAR (Pevzner & Sze, 2000), Random Projection (Buhler & Tompa, 2002) dan RPB-DC (Galib, et al., 2015), RPPMD (Ashraf, et al., 2017). Dari semua algoritma tersebut, RPPMD mendapatkan nilai akurasi tertinggi. RPPMD merupakan algoritma hasil modifikasi dari algoritma Random Projection. Dalam Random Projection terdapat

konsep *bucketing* dimana semua *l*-mers di proyeksikan kedalam *bucket* tersebut sesuai jumlah (*k*) posisi *random* sehingga mempercepat proses pencarian motif.

Namun seiring dengan perkembangan zaman, data yang dihasilkan dari proses sekuensing menjadi semakin banyak, oleh karena itu para ilmuwan dituntut untuk dapat mengatasi permasalahan komputasi dengan data yang lebih besar (Liu, et al., 2015). Maka dari itu munculah konsep komputasi paralel, dimana maksud dari komputasi paralel ialah menyelesaikan suatu perkerjaan secara bersama-sama, baik itu menggunakan banyak *core* atau *node* atau komputer secara bersamaan. Dalam penelitian yang dilakukan oleh (Nicole & Rajasekaran, 2014) diterapkan model komputasi paralel untuk pencarian motif pada DNA dengan menggunakan banyak *core*.

Penerapan komputasi paralel salah satunya dapat menggunakan *package high perfomance computing* (HPC) pada bahasa pemograman R. *Package* yang dihasilkan dari pengembangan MPI (*Message Passing Interface*) dalam komputasi paralel dengan menggunakan bahasa pemograman R diantaranya ialah pbdMPI (Chen, et al., 2012). pbdMPI ialah salah satu dari *programming with big data in R* atau disingkat dengan pbdR yang berguna sebagai alat bantu dalam pendistribusian memori dan juga komunikasi data yang akan didistribusikan pada proses komputasi paralel.

Pada penelitian ini penulis akan merancang sebuah model dan mengimplementasikan algoritma Random Projection pada pbdMPI menggunakan bahasa pemograman R . Model yang telah dirancang tersebut akan digunakan untuk kebutuhan analisa data dalam skala besar. Pada penelitian ini model tersebut digunakan untuk menyelesaikan masalah pada PMS (*planted motif search*). Penulis berharap semoga model yang dibuat dapat mempercepat proses pencarian motif sehingga banyak motif yang dapat dibangkitkan.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan pada sub bab sebelumnya, maka terdapat beberapa masalah yang dirumuskan sebagai berikut:

1. Bagaimana model komputasi paralel yang dihasilkan dengan menggunakan algoritma Random Projection untuk *planted motif search* pada sekuens DNA?
2. Bagaimana penerapan algoritma Random Projection 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 *batch* sesuai model yang dibuat dan diimplementasikan?

1.3 Tujuan Penelitian

Sesuai dengan rumusan masalah yang telah dibuat, maka tujuan dari penelitian ini ialah sebagai berikut:

1. Merancang model komputasi paralel untuk *planted motif search* pada sekuens DNA menggunakan algoritma Random Projection.
2. Menerapkan model yang telah dibuat pada *Package High-Performance Computing* pbdMPI dalam bahasa pemograman R.
3. Melakukan eksperimen dari program yang telah dibuat dengan menggunakan data sekuens DNA yang ada.
4. Melakukan analisa terkait kecepatan dan akurasi dari hasil eksperimen.

1.4 Manfaat Penelitian

Adapun manfaat yang dapat diberikan sebagai hasil dari penelitian ini ialah sebagai berikut :

1. Mempermudah para ahli biologis untuk mencari motif yang terdapat dalam sekuens DNA.
2. Mempermudah ahli biologis dalam memahami struktur dan fungsi dari sekuens DNA.

3. Dapat menjadi literatur untuk pengembangan program dan model komputasi paralel.
4. Memberikan pengetahuan mengenai penerapan ilmu Bioinformatika, khususnya mengenai *planted motif search* pada sekuens DNA

1.5 Batasan Masalah

Dalam penelitian yang dilakukan terdapat beberapa batasan masalah sebagai berikut:

1. Program ini bekerja hanya untuk data dengan format file .fasta.
2. Jumlah *core* yang digunakan dalam eksperimen ini adalah 1 sampai 6 *cores*.
3. Sekuens DNA yang dikenali hanya asam amino adenine, guanine, cytosine dan thymne. Yang disimbolkan dengan huruf A,C,G,T

1.6 Sistematika Penelitian

Pada sub bab ini akan dijelaskan secara singkat kandungan yang ada pada penelitian setiap bab.

Bab I pendahuluan menjelaskan alasan peneliti dalam mengangkat masalah dan memberikan solusi terhadap masalah tersebut. Subbab yang terdapat pada bab ini ialah latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah dan sistematika penelitian.

Selanjutnya bab II kajian pustaka merupakan bab yang menjelaskan mengenai teori-teori yang mendukung dalam penelitian ini. Teori yang akan dijelaskan pada bab ini secara garis besar akan berhubungan dengan tiga topik utama yaitu *planted motif search,* algoritma Random Projection dan pbdMPI.

Kemudian bab III metodologi penelitian. Dalam bab ini dijelaskan langkah-langkah yang akan dilakukan oleh peneliti dalam melaksanakan penelitian ini. Dimulai dari penjelasan desain penelitian, fokus penelitian, spesifikasi alat dan

bahan yang digunakan untuk penelitian dan yang teakhir ialah mengenai metode penelitian.

Bab IV hasil dan pembahasan merupakan bab yang menjelaskan hasil dari penelitian dan percobaan yang telah dilakukan oleh peneliti. Semua pertanyaan mengenai masalah yang diangkat dalam penelitian ini dibahas pada bab ini. Beberapa hal di antaranya adalah tentang proses pengumpulan data, pengembangan model, implementasi sistem, studi kasus, desain eksperimen, dan analisa.

Selanjutnya bab V kesimpulan dan saran. Pada bab ini diutarakan hasil dari seluruh pembahasan serta saran yang diajukan oleh peneliti untuk penelitian yang akan dilakukan selanjutnya.