Implementasi Model Pemograman *Map-Reduce* pada Pengelompokan Sekuen DNA dengan Metode *Single Link*

Ferry Ramdhani(G64134005)*, Wisnu Ananta Kusuma

Abstrak/*Abstract*

Sekuen DNA merupakna rantaian urutan huruf-huruf yang mewakili struktru primer dari molekul DNA. Kendala dalam sekuen DNA adalah jumlah data yang besar sehingga memerlukan waktu komputasi yang lama. Penelitian ini bertujuan untuk melihat hasil pengelompokan sekuen DNA dan waktu komputasi dengan metode *single link*. Proses pengelompokan akan menerapkan model *map-reduce* untuk menangani besarnya ukuran data dari sekuen DNA. Penelitian ini akan menggunakan metode *k-mers frequency* untuk ekstraksi fitur.

Kata Kunci

k-mers frequency; *map-reduce*; sekuen DNA; *single link*.

***Alamat Email**: ferry.ramdhani.16@gmail.com

PENDAHULUAN

Latar Belakang

Bioinformatika merupakan salah satu cabang ilmu yang memiliki peranan penting dalam kemajuan ilmu biologi, salah satunya adalah analisis sekuen *deoxyribo nucleid acid* (DNA). DNA merupakan pembawa informasi genetik dari makhluk hidup. DNA merupakan rantai ganda dari nukleotida yang diikat dalam struktur *helix* dikenal dengan *double helix*. Terdapat 4 basa utama dalam setiap nukleotida yaitu: *adenine* (A), *cytosin* (C), *thymine* (T), dan *guanine* (G).

Sekuen DNA didapatkan dengan memotong DNA dari suatu organisme yang diuraikan dan dipotong-potong menjadi *reads*. *Reads* tersebut berisi urutan huruf-huruf yang mewakili struktur primer dari molekul DNA. Sekuen DNA dalam bentuk *file* akan disimpan dalam format FASTA. Dari *reads* tersebut, dapat dilihat kode genetik setiap makhluk hidup. Variasi urutan basa setiap makhluk hidup memiliki kemiripan. Oleh karena itu, untuk mengetahui kekerabatan antarspesies diperlukan pengelompokan berdasarkan kesamaan ciri fiturnya.

Proses *binning* pada sekuen DNA akan dilakukan untuk dikelompokan. Proses *binning* dapat dilakukan dengan menggunakan metode *unsupervised*, yaitu dengan pengelompokan. Pengelompokan adalah proses pembelajaran *unsupervised* terhadap suatu pattern untuk dijadikan beberapa kelompok berdasarkan kemiripan (Jain et al. 1999). Teknik pengelompokan digunakan untuk melihat kemiripan dengan melihat hasil dendogram dengan menggunakan metode hierarki. Metode hierarki juga dibagi menjadi beberapa macam seperti: *single linkage*, *complete linkage*, *average linkage*, dan *average group linkage*. Penelitian tentang *single link* pada sekuens DNA pernah dilakukan oleh Tamsin (2013). Pada penelitian tersebut, ekstraksi ciri yang digunakan adalah *feature vector* dan tingkat kemiripan menggunakan *cosine similarity*. Dari 8 studi kasus yang masing-masing terdiri dari 5 percobaan menggunakan 50 data sekuen DNA, didapatkan akurasi rata-rata 86.7% dengan nilai akurasi tertinggi 100% dan terendah 70%.

Sekuen DNA merupakan yang data yang sangat besar yang bahkan dapat mencapai ratusan megabase data (Kunin et al. 2008). Salah satu solusi untuk mengatasi hal tersebut adalah pemodelan *map-reduce*. *Map-reduce* merupakan model pemrograman yang penerapannya digunakan untuk memproses data yang berukuran besar (Dean and Ghemawat 2004). *Map-reduce* diimplementasikan dalam *platform* hadoop. Hadoop adalah *framework* yang menangani data berskala besar dan digunakan untuk kluster pada Linux dengan tujuan untuk analisis data (Taylor 2010). Selain itu, penelitian terkait juga pernah dilakukan oleh Rasheed and Rangwala (2013) yang menggunakan teknik pengelompokan dengan metode MC-MinH. Penelitian yang dilakukan berfokus pada evaluasi pengelompokan dan waktu komputasi. Hasilnya, sekuen hasil pengelompokan sama dengan sekuen yang ada pada metagenom.

Pada penelitian ini, penulis akan mengimplementasikan *hierachical clustering* yang sama dengan yang dilakukan oleh Tamsin (2013) dengan menggunakan *single link* dengan *k-mers* sebagai ekstrasi ciri. Namun, penulis akan menggunakan Hadoop dengan pemodelan *map-reduce*. Dalam penelitian, ini juga ingin dilihat evalusi dari hasil pengelompokan dan waktu komputasi.

Perumusan Masalah

Perumasan masalah pada penelitian ini adalah penerapan *map-reduce* untuk pengelompokan sekuen DNA dengan menggunakan metode *single link*. Hasil dari pengelompokan akan dihitung waktu komputasinya.

Tujuan

Tujuan dari penelitian ini adalah untuk mengimplementasikan pemodelan *map-reduce* untuk pengelompokan sekuen DNA dengan *single link*, melihat hasil pengelompokan sekuen DNA dan mengevaluasi hasil pengelompokan dengan metode *single link* dan waktu komputasi

Ruang Lingkup

Ruang lingkup penelitian adalah:

1. Penelitian ini menitik beratkan pada tahap pengelompokan.
2. Data sekuen DNA yang digunakan dengan format FASTA.
3. Data sekuen yang digunakan adalah DNA bakteri *complete sequence*.
4. Data hasil simulasi bersifat bebas *error*.
5. Bahasa pemrograman yang digunakan adalah Java.

Manfaat

Manfaat dari penelitian ini adalah sebagai pertimbangan untuk penerapan teknik pengelompokan dengan pemodelan *map-reduce* pada sekuen DNA.

TINJAUAN PUSTAKA

Map-reduce

Map-reduce merupakan pemodelan pemograman yang digunakan untuk memproses data yang berukuran besar pada lingkungan paralel (Dean and Ghemawat 2004). Pemodelan *map-reduce* mempunyai *run time system* yang menjadi salah satu kelebihan dari pemodelan ini. *Run time system* dapat menangani masalah pembagian data dan penanganan kesalahan sistem. *Map-reduce* memiliki dua tahap yang yaitu *map* dan *reduce*. *Map* dan *reduce* dijalankan secara terpisah namun sama-sama dilakukan secara paralel. Setiap langkah mempunyai atribut *key* dan *value* yang sepasang. *Map* berfungsi untuk memetakan input ke dalam beberapa node dan memasangkan *key* dan *value* pada tiap node. Komputer yang menjalankan tahap ini disebut *mappers*. *Reduce* berfungsi untuk menyatukan hasil dari *map* dengan melihat *key* yang diberikan pada tahap *map*. *Reduce* juga akan memblokir proses sampai data dari *map* sudah semua ditransfer (Taylor 2010). Komputer yang menjalankan tahap ini disebut *reducers*.

K-mers Frequency

K-mers frequency merupakan frekuensi untuk menghitung banyaknya kemunculan dari subsequence yang mungkin dari 4 kombinasi (A, T, G, dan C) dengan panjang k pada suatu sekuen DNA. Setiap kombinasi akan dihitung sesuai dengan kombinasi dari sekuen DNA sebanyak k. Subsequence yang akan dihitung menggunakan nilai k sama dengan 3 akan menghasilkan pola ciri sebesar 43 atau 64 base pair. Pola ciri yang didapatkan adalah AAA, AAC, AAT, AAG, ACA, ACC, ACT, ACG, ATA, ATC, ATT, ATG, AGA, AGC, AGT, AGG, CAA, CAC, CAT, CAG, CCA, CCC, CCT, CCG, CTA, CTC, CTT, CTG, CGA, CGC, CGT, CGG, TAA, TAC, TAT, TAG, TCA, TCC, TCT, TCG, TTA, TTC, TTT, TTG, TGA, TGC, TGT, TGG, GAA, GAC, GAT, GAG, GCA, GCC, GCT, GCG, GTA, GTC, GTT, GTG, GGA, GGC, GGT, GGG. Ilustrasi perhitungan k-mers frequency pada sekuen DNA dapat dilihat pada Gambar 1.

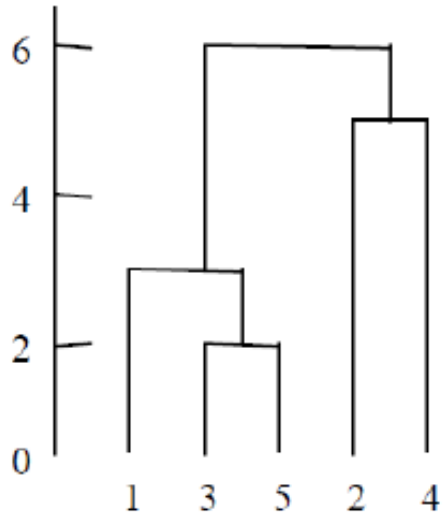
Gambar 1. Ilustrasi perhitungan *k-mers*

Single Link

Dalam analisis kluster pada dasarnya akan dilakukan pengelompokan secara alami terhadap sekelompok objek, dengan melakukan perbandingan terhadap masing-masing objek yang memiliki tingkat kesamaan atau jarak. Pengelompokan adalah proses pembelajaran *unsupervised* terhadap suatu *pattern* untuk dijadikan beberapa kelompok berdasarkan kemiripan (Jain et al. 1999). Kluster adalah koleksi dari *record* yang mirip dan tidak mirip dengan *record* dari kluster lain. Pengelompokan berbeda dengan klasifikasi, dalam hal ini tidak ada variabel target untuk dikelompokkan. Pengelompokan tidak mengklasifikasikan, meramalkan atau memprediksi nilai dari

sebuah variabel target. Algoritme pengelompokan digunakan untuk menentukan segmen keseluruhan himpunan data menjadi subgrup yang relatif sama atau kluster, dengan kesamaan record dalam kluster dimaksimumkan dan kesamaan *record* di luar kluster diminimumkan (Larose 2005).

Pengelompokan data dengan metode *single link* termasuk ke dalam metode *hierarchical agglomerative clustering*. *Hierarchical* melihat objek yang mirip yang nantinya akan dikelompokkan. Hasilnya dalam bentuk dendogram yang akan menampilkan gambaran antarkelompok yang terdekat. Ilustrasi dendogram dapat dilihat pada Gambar 2.

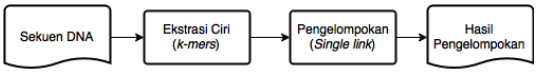


Gambar 2. Ilustrasi dendogram pada *single link*

Single link akan mengelompokkan objek berdasarkan jarak terdekat antaranggota. *Input* dari metode *single link* bisa berupa jarak atau tingkat kesamaan antara pasangan dari objek. Objek akan dikelompokkan berdasarkan jarak terdekat, dipilih nilai terkecil lalu digabungkan, dan hasilnya akan dibandingkan kembali dengan jarak pada kelompok lain.

METODE PENELITIAN

Penelitian ini melakukan pengelompokan dengan menggunakan teknik pengelompokan dengan metode *single link*. Tahapan yang dilakukan pada penelitian ini,ialah penyiapan data, ekstraksi ciri dengan *k-mers*, pengelompokan sekuen DNA dengan menggunakan *single link*, dan menganalisi hasil pengelompokan. Gambar 3 menunjukan tahapan proses tersebut.



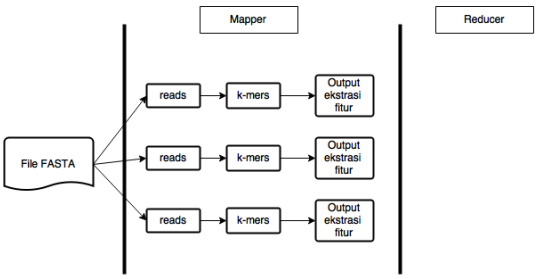
Gambar 3. Tahapan proses penelitian

Penyiapan Data

Data yang digunakan pada penelitian ini merujuk kepada penelitian Tamsin (2013). Data tersebut menggunakan 50 data sekuen DNA yang terdiri dari 10 data dari genus *Borellia*, 10 genus dari *Streptococcus*, 10 genus dari *Yersinia*, 10 genus dari *Methylobacterium*, dan 10 genus dari Bacillus. Semua data diperoleh dari situs National Center of Biotechnology Information, US National Library of Medicine dalam format FASTA.

Ekstraksi Ciri

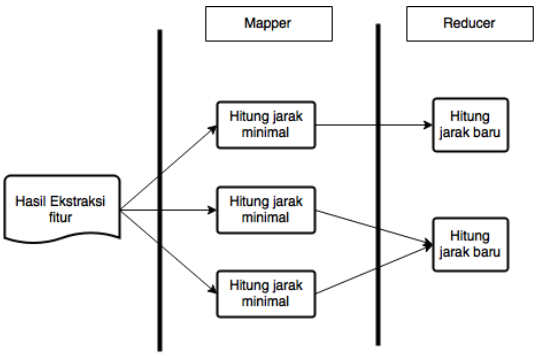
Pada tahap ini akan dilakukan ekstraksi ciri dari sekuen DNA dengan menggunakan *k-mers frequency*. Implementasi *map-reduce* pada tahap ini dengan memetakan sekuen DNA ke dalam *mappers*. Setiap mappers akan menghitung *k-mers frequency* dari reads yang sudah dipetakan. Hasilnya akan langsung menjadi output tanpa harus melakukan *reducers*. Ilustrasi proses *map-reduce* untuk ekstraksi ciri dapat dilihat pada Gambar 4.



Gambar 4. Illustrasi *k-mers frequency* dengan model *map-reduce*

Pengelompokan

Ilustrasi untuk pemodelan *map-reduce* dengan *single link* dapat dilihat pada Gambar 5. Ilustrasi untuk pemodelan



Gambar 5. Illustrasi *single link* dengan model *map-reduce*

Tabel 1. Rencana Jadwal Penelitian

Kegiatan	Juni				Juli				Agustus				September				Oktober			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Pengumpulan data dan anlisis kebutuhan sistem	■	■	■																	
Perancangan dan pemodelan				■	■	■	■	■	■	■	■	■								
Implementasi Sistem									■	■	■	■								
Penulisan skripsi final	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■				
Seminar																	■			

map-reduce dengan *single link* dapat dilihat pada Gambar 5.Pada tahap ini dilakukan pengelompokan dengan menggunakan *single link*. Pengelompokan dilakukan dengan data tingkat kesamaan yang didapat dari ekstraksi fitur, dimulai dengan pengelompokan menggunakan 1 data tiap genus, hingga 9 data sekuen setiap genus. Hasil dari ekstraksi akan dihitung jarak terdekatnya dan dipilih sebagai kluster, selanjutnya akan dihitung kembali jarak baru dengan memilih jarak terdekat dari kluster yang telah terpilih.

Evaluasi

Hasil dari pengelompokan *single link* akan dievaluasi dengan menghitung akurasi yang diperoleh. Akurasi akan dihitung menggunakan persamaan sebagai berikut:

$$Akurasi = \frac{\Sigma\ data\ benar}{\Sigma\ jumlah\ data} \times 100\% \qquad (1)$$

Jadwal Kegiatan

Penelitian ini akan dilakukan selama 4.5 bulan dengan rincian kegiatan seperti tercantum pada Tabel 1.

DAFTAR PUSTAKA

Dean, J and S Ghemawat (2004). "MapReduce: simplified data processing on large clusters. OSDI'04 Proceedings of the 6th conference on Symposium on Opearting Systems Design and Implementation" dalam: *International Journal of Enggineering Science Invention*, pp. 10–100. URL: http://static.googleusercontent.com/media/research.google.com (diunduh pada 2015-05-10).

Jain, AK, MN Murty, and PJ Flynn (1999). *Data Clustering: a Review*. New York (US): ACM Computing Surveys.

Kunin, V, A Copeland, A Lapidus, K Mavromatis, and P Hugenholtz (2008). "A Bioinformatician's Guide to Metagenomics" dalam: *Microbiology and Molecular Biology Reviews* 72 (4), pp. 557–578.

Larose, DT (2005). *Discovering Knowledge in Data: an Introduction to Data Mining*. Jersey (US): J Wiley.

Rasheed, Z and H Rangwala (2013). "A map-reduce framework for clustering metagenomes" dalam: *HiCOMB 2013 Online Proceedings Twelfth IEEE International Workshop on High Performance Computational Biology*, pp. 549–558. URL: http://www.hicomb.org/HiCOMB2013/papers/HICOMB2013-06.pdf (diunduh pada 2015-05-10).

Tamsin, AH (2013). "Pengelompokan Sekuen DNA Menggunakan Algoritme Single Link dan Feature Vectors". Skripsi. Departemen Ilmu Komputer, Institut Pertanian Bogor. 37 pp.

Taylor, RC, ed. (2010). *An Overview of the Hadoop Framework and its Current Applications in Bioinformatics*. The 11th Annual Bioinformatics Open Source Conference (BOSC).