Jurnal Ilmu Komputer Agri-Informatika

Tersedia secara online di:
http://journal.ipb.ac.id/index.php/jika

Volume 3 Nomor 1 halaman 9 - 18
ISSN: 2089-6026

Klasifikasi Metagenom dengan Metode *Naïve Bayes Classifier*

Metagenome Classification Using Naïve Bayes Classifier Method

DIAN KARTIKA UTAMI, WISNU ANANTA KUSUMA*, AGUS BUONO

Abstrak

Studi metagenom merupakan langkah penting pada pengelompokan taksonomi. Pengelompokan pada metagenom dapat dilakukan dengan menggunakan metode *binning*. *Binning* diperlukan untuk mengelompokkan *contigs* yang dimiliki oleh masing-masing kelompok spesies filogenetik. Pada penelitian ini, *binning* dilakukan dengan menggunakan pendekatan komposisi berdasarkan *supervised learning* (pembelajaran dengan contoh). Metode *supervised learning* yang digunakan yaitu *Naïve Bayes Classifier*. Adapun metode yang digunakan untuk ekstraksi ciri adalah dengan melakukan perhitungan frekuensi *k-mer*. Klasifikasi pada metagenom dilakukan berdasarkan tingkat takson *genus*. Dari proses klasifikasi yang dilakukan, akurasi yang diperoleh dengan menggunakan fragmen pendek (400 bp) adalah 49.34 % untuk ekstraksi ciri 3-*mer* dan 53.95 % untuk ekstrasi ciri 4-*mer*. Sementara itu, untuk fragmen panjang (10 kbp), akurasi mengalami peningkatan yaitu 82.23 % untuk ekstraksi ciri 3-*mer* dan 85.89 % untuk esktraski ciri 4-*mer*. Dari hasil tersebut dapat disimpulkan bahwa akurasi semakin tinggi seiring dengan semakin panjangnya ukuran fragmen. Selain itu, penelitian ini juga menyimpulkan bahwa metode ekstrasi ciri yang memberikan hasil paling maksimal adalah dengan menggunakan ekstraksi ciri 4-*mer*.

Kata Kunci: *metagenom, k-mer, Naïve Bayes Classifier, binning, klasifikasi*

Abstract

Metagenome study is an important step in the taxonomic grouping. Grouping can be conducted using binning method. Binning is required to determine contigs of each phylogenetic species groups. In this study, binning is done using supervised learning based composition approach. We used NaïveBayes Classifier method for performing supervised learning and employed counting of k-mer frequencies for extracting feature. The classification process was conducted at genus-level taxon. The results showed that using short fragments (400 bp), our method could obtain the accuracy of 49.34 % and 53.95 % with features of 3-mers dan 4-mers frequencies, respectively. Meanwhile, the accuracy of our method was significally increased when classifying long fragments (10 kbp). Our method could obtain the accuracy of 82.23% with 3-mers frequencies feature and 85.89% with 4-mers frequencies feature. It can be concluded that the accuracy of our classifier was increased by increasing the size of fragments. Moreover, in this research, the 4-mers frequencies feature gave the best results for classifying metagenome fragments.

Keywords: *metagenome, k-mer, Naïve Bayes Classifier, binning*

PENDAHULUAN

Studi metagenom merupakan langkah penting pada pengelompokan taksonomi(Higashi *et al*. 2012). Pengelompokan metagenom menggunakan proses yang disebut *binning*. *Binning* diperlukan untuk mengelompokan *contigs* yang dimiliki dari masing-masing kelompok spesies filogenetik.

Metode *binning* terdiri atas dua pendekatan yaitu berdasarkan komposisi dan homologi. Metode *binning* berdasarkan komposisi melakukan perhitungan frekuensi ciri yang muncul dari pasangan basa (*base pair*) yang membentuk sekuens metagenom. Ciri komposisi

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Penulis korespondensi: Tel/Faks: 0251-8625584; Surel: ananta@ipb.ac.id

digunakan sebagai masukkan pembelajaran dengan contoh (*supervised learning*) atau pembelajaran secara observasi (*unsupervised learning*). Metode *binning* berdasarkan komposisi dengan *supervised learning* yang telah dilakukan antara lain *Naïve Bayes Classifier* (Rosen *et al*. 2008), *Support Vector Machine* (Kim *et al*. 2010), *PhyloPythia* (McHardy *et al.* 2007), dan *Phymm* (Brady dan Salzberg 2009). Metode *binning* berdasarkan komposisi dengan *unsupervised learning* terdiri atas *Growing Self Organizing Map* (Chan *et al.* 2008 dan Overbeek *et. al.*, 2013), SOC atau *Self Organizing Clustering* (Amano *et al.* 2003; Amano *et al.* 2007), Kohonen SOM atau Kohonen *Self Organizing Map* (Abe *et al.* 2003).

Metode *binning* berdasarkan homologi melakukan pencarian penjajaran sekuens dengan membandingkan fragmen metagenom dengan sekuens referensi yang terdapat pada basis data yang digunakan, misal *National Centre for Biotechnology Information* (NCBI). Hasilnya disimpulkan pada tiap level taksonomi. Hal tersebut menyebabkan pendekatan dengan homologi membutuhkan banyak waktu dalam proses pengelompokan.

Beberapa penelitian yang telah dilakukan, diantaranya Kusuma dan Akiyama (2011) mengusulkan metode klasifikasi fragmen metagenom dengan menggunakan *Support Vector Machine* (SVM) dan *characterization vector* sebagai fiturnya. Untuk mengevaluasinnya, Kusuma dan Akiyama (2011) mengimplementasikannya pada *dataset* kecil yang mempresentasikan komunitas mikrob kecil. Untuk data latih digunakan 10 organisme, sedangkan untuk data uji digunakan 9 organisme yang mempresentasikan organisme baru. Organisme yang digunakan pada data uji ialah organisme yang berbeda dengan data latih, namun termasuk ke dalam genus yang sama. Hasil akurasi yang diperoleh dari penelitian ini cukup tinggi yaitu 73% untuk panjang fragmen 500 bp sampai dengan 87% untuk panjang fragmen 10 kbp. Namun, ketika metode ini diterapkan pada *dataset* berukuran besar yaitu 374 organisme, akurasi yang diperoleh menurun secara signifikan, yaitu sebesar 30% untuk panjang fragmen 1 kbp pada level genus.

Untuk memperbaiki akurasi penelitian Kusuma dan Akiyama (2011), Arini (2013) mengusulkan penelitian yaitu klasifikasi fragmen metagenom menggunakan metode *Support Vector Machine* (SVM) yang didasari oleh penelitian Kusuma dan Akiyama (2011). *Dataset* yang digunakan yaitu 381 organisme sebagai data latih dengan rata-rata pembacaan sebanyak 320 000 dan 200 organisme sebagai data uji dengan rata-rata pembacaan sebanyak 100 000, kemudian data latih dan data uji diekstraksi ciri dengan metode *k-mer*. Untuk panjang fragmen yang digunakan yaitu 400 bp, 800 bp, 1 kbp, 3 kbp, 5 kbp dan 10 kbp yang menghasilkan akurasi antara 65.3% sampai dengan 95.4% pada level genus.

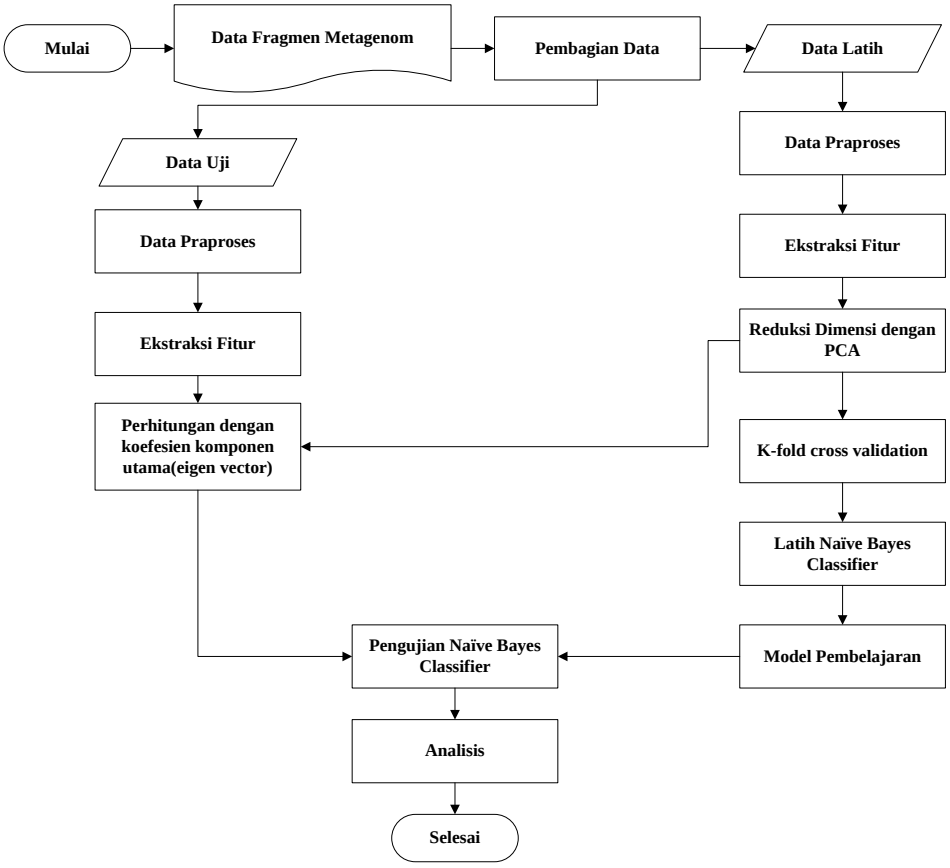
Penelitian lain dilakukan oleh Rahmawati (2013) dengan menggunakan metode *Naïve Bayes Classifier*. Penelitian ini menghasilkan akurasi 74% dengan menggunakan data latih sebanyak 381 organisme dengan jumlah kelas sebanyak 48 genus dan rata-rata jumlah pembacaan sebanyak 9 600, serta panjang fragmen yang digunakan yaitu 500 bp. Namun Rahmawati (2013) belum melakukan elaborasi pengaruh panjang fragmen terhadap akurasi.

Penelitian ini bertujuan untuk memperbaiki penelitian Rahmawati (2013) sehingga model yang dihasilkan dapat mengklasifikasi fragmen metagenom dengan panjang fragmen yang berbeda, yaitu 400 bp, 800 bp, 1 kbp, 3 kbp, 5 kbp dan 10 kbp. Fragmen yang digunakan pada penelitian ini berasal dari 381 organisme sebagai data latih dan 200 organisme sebagai data uji. Fitur yang digunakan adalah frekuensi *k-mer*. Untuk mendapatkan dimensi matriks ciri yang kecil maka digunakan 3-*mer* dan 4-*mer* yang terlebih dulu direduksi menggunakan Principal Component Analysis (PCA).

METODE

Penelitian ini dilakukan dengan mengikuti beberapa tahapan proses, yaitu pengumpulan data fragmen metagenom, pembagian data (data latih dan data uji), praproses data, ekstraksi fitur dengan metode *k-mer*, reduksi dimensi dengan PCA, *k-fold cross validation*, pelatihan

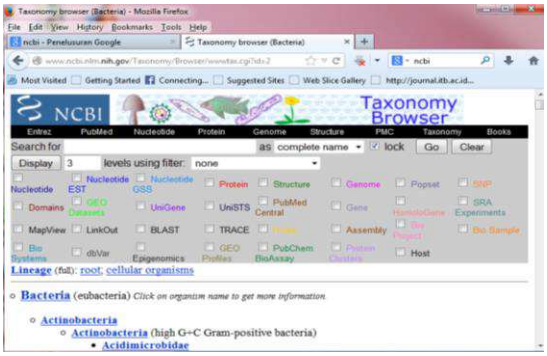
dan pengujian dengan *Naïve Bayes Classifier*, dan tahap analisis. Tahapan pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1 Metode Penelitian

Pengumpulan data

Data yang digunakan pada penelitian ini adalah data metagenome yang diunduh dari situs *National Centre for Biotechnology Information* (NCBI)(Gambar 2). NCBI merupakan salah satu institusi yang menjadi rujukan atau sumber informasi perkembangan biologi molekuler (Federhen 2012). Data metagenom yang digunakan ini merupakan *simulated data* berupa fragmen-fragmen DNA yang dibangkitkan dengan menggunakan aplikasi MetaSim (Richter *et al*. 2008) dari sekuens genom lengkap bakteri yang diunduh dari NCBI. MetaSim (Gambar 3) mensimulasikan kinerja *sequencer* untuk menghasilkan fragmen metagenom yang direpresentasikan sebagai string dengan format data berupa FNA (FASTA *Nucleic Acid*) (Richter *et al*. 2008).



Gambar 2 Situs NCBI

Gambar 3 Simulator MetaSim

Pembagian Data

Data penelitian yang digunakan dalam penelitian ini berupa *dataset* besar yang terdiri atas 381 organisme sebagai data latih dan 200 organisme sebagai data uji yang termasuk kedalam 48 *genus*. Data latih dan data uji menggunakan panjang fragmen yang seragam, yaitu 400 bp, 800 bp, 1 kbp, 3 kbp, 5 kbp dan 10 kbp. Satuan bp (*base pair*) adalah banyaknya atau panjangnya unsur basa a*denine* (A), *thymine* (T), *guanine* (G) dan *cytosine* (C) suatu DNA. Banyaknya pembacaan yang digunakan untuk data latih yaitu 200 000 pembacaan, sedangkan untuk data latih 100 000. Data *genus* yang digunakan seperti ditunjukkan pada Tabel 1.

Tabel 1 Genus berdasarkan NCBI *Taxonomy Browser*

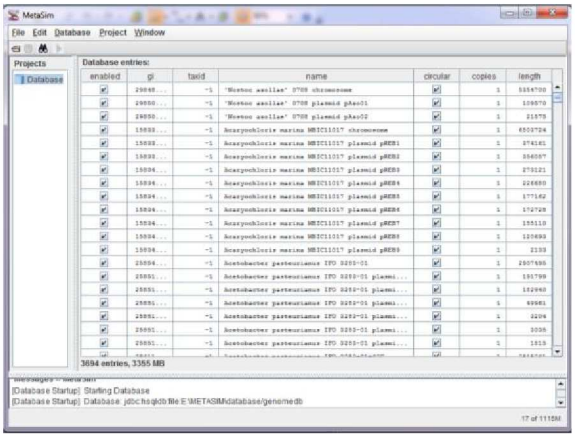
No	Genus	No	Genus	No	Genus
1	Bacillus	17	Ehrlichia	33	Pyrobaculum
2	Bacteroides	18	Francisella	34	Pyrococcus
3	Bartonella	19	Frankia	35	Rickettsia
4	Bordetella	20	Geobacter	36	Shewanella
5	Borrelia	21	Haemophilus	37	Shigella
6	Bradyrhizobium	22	Helicobacter	38	Staphylococcus
7	Brucella	23	Lactobacillus	39	Streptococcus
8	Burkholderia	24	Leptospira	40	Streptomyces
9	Campylobacter	25	Listeria	41	Sulfolobus
10	Candidatus	26	Methanococcus	42	Synechococcus
11	Chlamydophila	27	Methanosarcina	43	Thermoanearobacter
12	Chlorobium	28	Methylobacterium	44	Thermotoga
13	Clostridium	29	Mycobacterium	45	Vibrio
14	Corynebacterium	30	Mycoplasma	46	Wolbachia
15	Cupriavidus	31	Pseudomonas	47	Xanthomonas
16	Dehalococcoides	32	Psychobacter	48	Yersinia

Sumber : NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi)

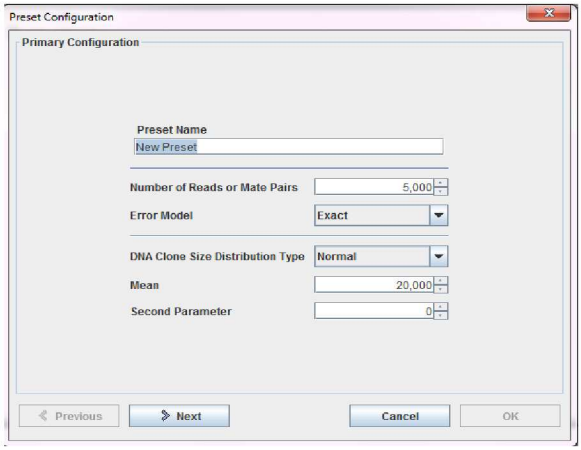
Data Praproses

Pada tahap praproses data, sekuens DNA metagenome yang sudah dipilih dari situs NCBI, akan diuraikan fragmennya menggunakan simulator MetaSim. Adapun tahapan yang dilakukan menggunakan MetaSim yaitu:

1 Basisdata bakteri yang berupa format FNA (FASTA *Nucleic Acid*) dihubungkan terlebih dahulu dengan simulator MetaSim. Jika basisdata telah berhasil dihubungkan, maka pada MetaSim akan dapat dilihat tampilan seperti yang ditunjukkan pada Gambar 4. Tahap selanjutnya yaitu pemilihan data latih dan data uji yang akan digunakan pada klasifikasi. Selanjutnya masing-masing data latih dan data uji disimpan dengan membuat nama *project*.
2 Tahap selanjutnya adalah mengatur *preset*. Pengaturan pada *preset* dilakukan untuk mengatur panjang fragmen dan banyaknya pembacaan yang akan digunakan. Pengaturan *preset* dapat dilihat pada Gambar 5.
3 Tahap terakhir adalah menjalankan simulator sesuai dengan *preset* yang telat diatur.



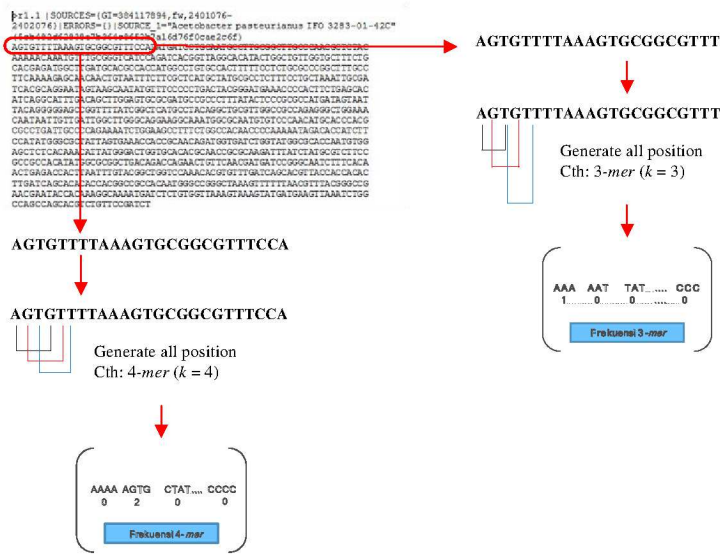
Gambar 4 Tampilan MetaSim dan Basisdata



Gambar 5 *Preset*

Ekstraksi Fitur

Ekstraksi fitur merupakan tahapan untuk mendapatkan suatu fitur dari data latih dan data uji dengan menggunakan frekuensi *k-mer*. Pola kemunculan *k* dalam sekuens dihitung menggunakan empat basa utama yaitu *Adenine* (A), *Thymine* (T), *Guanine* (G) dan *Cytosine* (C) yang dipangkatkan dengan rangkaian pasangan basa yang ingin digunakan (pola kemunculan: 4^k, dengan $k \geq 1$) (Kusuma 2012).



Gambar 6 Ekstraksi Fitur *k-mer*

Reduksi Dimensi Dengan PCA

Penggunaan metode ekstraksi fitur *k-mer* dengan pola $k = 3$ menghasilkan *array* 64 kombinasi x *n* jumlah pembacaan data. Dimensi data tersebut perlu direduksi tanpa adanya pengurangan karakteristik data secara signifikan sehingga lebih mudah untuk menginterpretasikannya.

Teknik reduksi data yang digunakan pada penelitian ini adalah *principal component analysis* (PCA). PCA adalah teknik yang digunakan untuk menyederhanakan suatu data dengan cara mentransformasi linier sehingga terbentuk sistem koordinat baru dengan varians maksimum. Analisis komponen utama merupakan teknik statistik yang dapat digunakan untuk menjelaskan struktur variansi-kovariansi dari sekumpulan variabel melalui variabel baru dimana variabel baru ini saling bebas, dan merupakan kombinasi linier dari variabel asal (Johnson 2002). Selanjutnya variabel baru ini dinamakan komponen utama (principal component). Salah satu tujuan dari analisis komponen utama adalah mereduksi dimensi data asal yang semula terdapat *p* variabel bebas menjadi *k* komponen utama (dimana $k \leq p$).

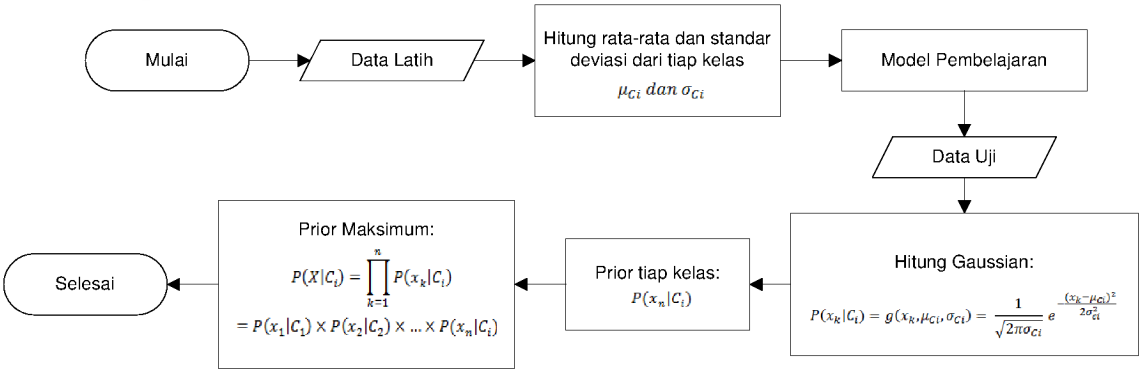
Adapun kriteria pemilihan *k* menurut Johnson (2002) yaitu proporsi kumulatif keragaman data asal yang dijelaskan oleh *k* komponen utama minimal 80 %, dan proporsi total variansi populasi bernilai cukup besar. Pada penelitian ini proporsi kumulatif keragaman data asal yang dipilih adalah sebesar 95%.

K-fold cross validation

Setelah mereduksi data menggunakan PCA dengan proporsi kumulatif 0.95, data tersebut akan dilatih dengan *Naïve Bayes Classifier*. Pelatihan data set dilakukan dengan menggunakan *k-fold cross validation* yang digunakan untuk membagi data menjadi data latih dan data uji. Pada penelitian ini *k* yang digunakan adalah 5. Data akan dibagi menjadi 5 bagian di mana 4 bagian akan menjadi data latih, dan 1 bagian sisanya akan digunakan untuk validasi. Data yang digunakan pada 5-fold *cross-validation* ini adalah data latih dengan jumlah fragmen 200 000.

Naïve Bayes Classifier

Naïve Bayes classifier didasari pada teorema Bayes dengan asumsi bahwa setiap ciri dalam klasifikasi tidak tergantung satu sama lain (Han dan Kamber, 2001). *Naïve Bayes classifier* berfungsi untuk menghitung peluang dari suatu kelas dari masing-masing kelompok atribut yang ada dan menentukan kelas mana yang paling optimal. Proses klasifikasi data dengan NBC diilustrasikan pada Gambar 7.



Gambar 7 Tahapan Klasifikasi dengan *Naïve Bayes Classifier*

Model

Pada proses pelatihan *Naïve Bayes Classifier,* sebelumnya model yang berupa bagian dari data latih yang dengan akurasi tinggi akan divalidasi kembali dengan data uji menggunakan *Naïve Bayes Classifier*.

Pengujian *Naïve Bayes Classifier*

Pengujian akan mengklasifikasikan data uji sebanyak 200 organisme baru yang masing-masing telah diketahui genusnya. Hasil klasifikasi dari model yang dibangun ini akan dievaluasi dengan menghitung persentase fragmen dari mikroorganisme yang diklasifikasikan dengan benar.

Analisis

Dari hasil pelatihan dan pengujian *Naïve Bayes Classifier* dapat dianalisis kinerja dari model klasifikasi yang dibangun menggunakan *Naïve Bayes Classifier*. Akurasi untuk hasil klasifikasi dihitung dengan persamaan (1).

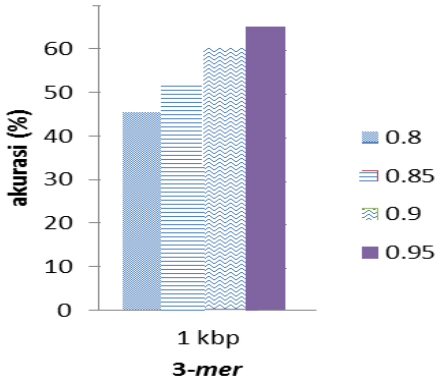
$$Akurasi = \frac{\sum datauji benar}{\sum datauji} \times 100\% \quad (1)$$

HASIL DAN PEMBAHASAN

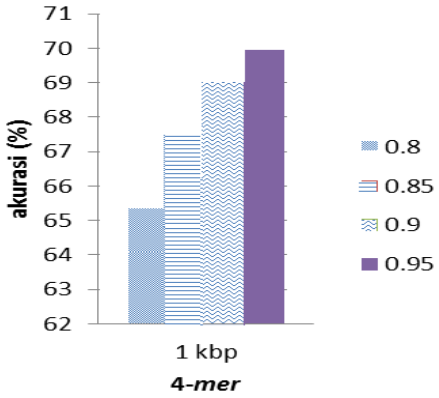
Analisis

Analisis dilakukan atas hasil klasifikasi di mana fitur yang digunakan telah direduksi menggunakan PCA. Untuk mendapatkan hasil terbaik, pengujian dilakukan pada beberapa nilai proporsi kumulatif. Selain itu disajikan pula hasil pengujian untuk beberapa panjang fragmen dan *k-mer* yang berbeda.

Pada penelitian ini, reduksi dimensi digunakan untuk menyederhanakan suatu data dengan cara melakukan transformasi linier sehingga terbentuk sistem koordinat baru dengan varian maksimum tanpa mengurangi karakteristik data secara signifikan. Teknik reduksi dimensi yang digunakan adalah PCA. Pada penelitian ini, untuk mendapatkan hasil terbaik, dievaluasi hasil klasifikasi pada beberapa nilai proporsi kumulatif, yaitu 0.8, 0.85, 0.9 dan 0.95. Hasil pengujian dengan panjang fragmen 1kbp dan ektrasi fiturnya 3-*mer* dan 4-*mer* dapat dilihat pada Gambar 8 dan Gambar 9.



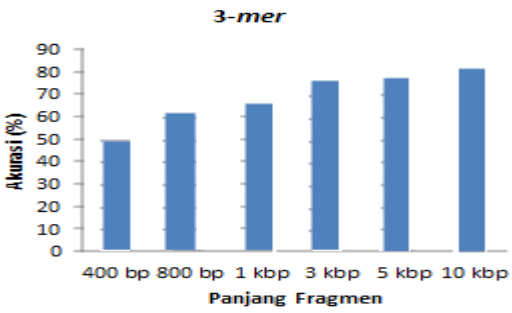
Gambar 8 *threshold* 3-*mer*



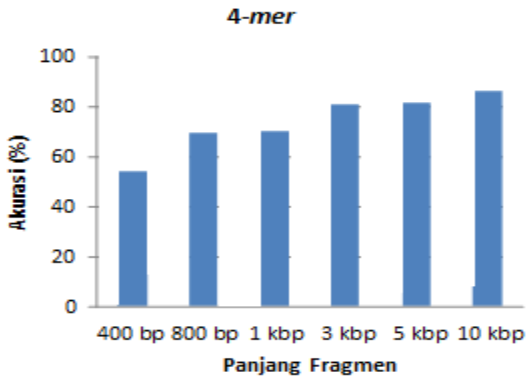
Gambar 9 *threshold* 4-*mer*

Berdasarkan hasil pada Gambar 8 dan Gambar 9, maka dapat disimpulkan bahwa akurasi yang paling baik berada pada threshold 0.95. Dengan threshold 0.95, ektrasi fitur 3-*mer* memperoleh akurasi sebesar 65.86% dan ekstrasi fitur 4-mer memperoleh akurasi sebesar 70.04%.

Selanjutnya nilai proporsi kumulatif 0.95 ini digunakan untuk menguji akurasi model klasifikasi pada panjang fragmen yang berbeda, yaitu mulai dari 400 bp, 800 bp, 1 kbp, 3 kbp, 5 kbp dan 10 kbp untuk masing-masing ekstraksi fitur 3-*mer* dan 4-*mer*. Akurasi yang diperoleh oleh model ini dapat dilihat pada Gambar 10 dan Gambar 11.



Gambar 10 Akurasi ekstraksi 3-*mer*



Gambar 11 Akurasi ekstraksi *4-mer*

Hasil percobaan menunjukkan bahwa pada proporsi kumultif sebesar 0.95 dan panjang fragmen 400 bp dihasilkan akurasi sebesar 49.34 % untuk ekstrasi fitur 3-*mer* dan 53.95 %

untuk ekstrasi fitur 4-*mer*. Adapun untuk panjang fragmen 10 kbp, akurasi mengalami peningkatan yaitu sebesar 82.23 % untuk ekstrasi fitur 3-*mer* dan 85.89 % untuk ekstraksi fitur 4-*mer*. Berdasarkan hasil akurasi yang diperoleh dapat dilihat bahwa panjang fragmen mempengaruhi hasil klasifikasi. Semakin panjang fragmen yang digunakan, akurasi yang diperoleh semakin tinggi. Hal ini menunjukkan bahwa semakin panjang fragmen semakin banyak informasi komposisi yang diperoleh. Dengan demikian similaritas antar fragmen semakin mudah untuk ditentukan. Namun demikian, ada kecenderungan bahwa mulai panjang fragmen 5 kbp, model ini memiliki akurasi yang cenderung konstan. Penggunaan panjang fragmen 10 kbp tidak memberikan pengaruh signifikan terhadap kenaikan akurasi. Dilihat dari hasil klasifikasi yang ditunjukkan Gambar 10, kecenderungan tersebut tampaknya berlaku bagi panjang fragmen lebih panjang dari 10 kbp.

Hasil penelitian (Gambar 8-12) menunjukkan bahwa metode ekstraksi fitur 4-*mer* memiliki akurasi yang lebih tinggi dibandingkan metode ekstraksi fitur 3-*mer*. Ini menunjukkan bahwa pola kemunculan $k \geq 1$ mempengaruhi akurasi yang diperoleh pada klasifikasi fragmen metagenom. Semakin besar *k* yang digunakan dalam ekstrasi fitur maka akan menghasilkan akurasi yang lebih baik. Memperbesar nilai *k* berarti memperbanyak kombinasi substring unik yang berperan dalam menentukan similaritas antar fragmen. Sebagai contoh jika kita menggunakan k=3 maka akan diperoleh 4^3 = 64 kombinasi 3-mers (AAA, AAT, AAG, AAC, ATA, ... , GGG). Sedangkan jika kita menggunakan k=4 maka akan diperoleh 4^4 = 256 kombinasi 4-mers (AAAA, AAAT, AAAG, AAAC, AATA, ..., GGGG). Dengan demikian semakin besar nilai *k*, semakin banyak variabel penciri yang dimiliki untuk membedakan fragmen yang satu dengan fragmen yang lain.

SIMPULAN DAN SARAN

Simpulan

Simpulan Berdasarkan hasil yang diperoleh dari penelitian yang telah dilakukan, dapat disimpulkan bahwa metode klasifikasi dengan menggunakan *Naïve Bayes Classifier* dan ekstraksi fitur *k-mer* yang diusulkan berhasil mengklasifikasikan fragmen metagenom pada level genus. Akurasi terbaik diperoleh dengan menggunakan 4-*mers* dan panjang fragmen 10 kbp, yaitu sebesar 85,9 %. Nilai akurasi ini bervariasi dengan bervariasinya panjang fragmen. Semakin panjang ukuran fragmen, akurasi model klasifikasi akan semakin besar. Namun demikian, setelah panjang 5 kbp, penambahan panjang ukuran fragmen tidak memberikan peningkataan akurasi yang signifikan. Selain itu, penelitian ini juga menyimpulkan bahwa semakin besar nilai *k* yang digunakan untuk mendapatkan matriks fitur yang direpresentasikan dengan frekuensi *k-mers*, akurasi dari model klasifikasi yang dibangun akan semakin tinggi. Pada penelitian ini, nilai akurasi model klasifikasi dengan menggunakan 4-*mers* lebih tinggi dari akurasi model klasifikasi yang menggunkan 3-*mers.*

Saran

Pada penelitian ini, model klasifikasi dibangun untuk mengelompokkan fragmen metagenom dengan panjang fragmen yang seragam. Pada kenyataannya, terdapat tipe *sequencer* yang menghasilkan fragmen dengan panjang yang tidak seragam. Untuk itu pada penelitian selanjutnya proses klasifikasi perlu diakukan pada panjang fragmen yang tidak seragam. Selain itu, pada penelitian ini data fragmen yang digunakan masih dihasilkan dari perangkat lunak simulasi yang bisa diatur kondisinya sedemikian rupa sehingga dapat diperoleh bebas dari *sequencingerror.* Pada penelitian selanjutnya, perlu dikembangkan model klasifikasi yang mampu mengklasifikasikan fragmen metagenom yang diambil secara riil dari lingkungan, seperti sampel dari *Sargasso Sea*, yang mungkin mengandung *sequencing error.*

DAFTAR PUSTAKA

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuku T, Ikemura T. 2003. *Informatics for Unveiling Hidden Genome Signatures*. *Genome Research*. 179(4):693-701. doi:10.1101/gr.634603.

Amano K, Nakamura H, Ichikawa H. 2003. Self-Organizing Clustering : A Novel Non-Hierarchical Method for Clustering Large Amountof Sequece DNAs. *Genome Informatics*. 14: 575-576.

Amano K, Nakamura H, Ichikawa H, Numa H, Kobayashi KF, Nagamura Y, Onodera N. 2007. *Self-Organizing Clustering : Non-Hierarchical Clustering for Large-Scale Sequence DNA Data*. *IPSJ Digital Courier*. 2(2):523-527.

Arini. 2013. *Metagenome Fragment Binning Using Support Vector Machine (SVM) Method* [skripsi]. Bogor-Indonesia : Institut Pertanian Bogor.

Brady A, Salzberg SL. 2009. *Phymm and PhymmBL : Metagenomic Phylogenetic Classification with Interpolated Markov Models*. *Nat Methods*. 6 (9) : 673 – 676. doi : 10.1038/nmeth.1358.

Chan CK, Hsu AL, Tang SL, Halgamuge SK. 2008. Using Growing Self-Organizing Maps to Prove the Binning Process in Environmental Whole-Genome Shotgun Equencing. *Journal of Biomedicine and Biotechnology*. 2008. doi:10.1155/2008/513701.

Federhen S. 2012. The NCBI Taxonomy Database. *Nucleic Acids Research*. 40: 136- 143. doi:10.1093/nar/gkr1178.

Higashi S, Barreto André da MS, Cantão ME, de Vasconcelos ATR. 2012. *Analysis Of Composition-Based Metagenome Classification*. BMC Genomic 2012, 13(Supply 5):S1. http://www.biomedcentral.com/1471-2164/13/S5/S1

Han Jiawei, Kamber Micheline, Pei Jian. Data Mining: Concept and Technique.3rd edition. ISBN 978-0-12-381479-1.

Johnson RA and Wichern, DW. 2002.*Applied Multivariate Statistical Analysis*,5th End. New Jersey: Prentice Hall.

Kim Jongwoo, Le DX, Thoma GR. 2010. *Naïve Bayes and SVM classifiers for classifying Databank Accession Number sentences from online biomedical articles*. Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7534, 75340U. doi: 10.1117/12.838961.

Kusuma WA, Akiyama Y. 2011. *Metagenome fragment binning based on characterization vectors*. Di dalam: Prosiding International Conferences on Bioinformatics and Biomedical Technology; 2011; Sanya, China. China (CH).

Kusuma WA. 2012. *Combined Approaches for Improving the Performance of de novo DNA Sequence Assembly and Metagenomic Classification of Short Fragments from Next Generation Sequencer* [tesis]. Tokyo (JP) : Tokyo Institute of Technology.

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. 2007. *Accurate phylogonetic classification of variabel-length DNA fragments*. *Nature Methods*. 4(1):63–72. doi: 10.1038/nmeth976.

Overbeek MV, Kusuma WA, Buono A. 2013. *Clustering Metagenome Fragment Using Growing Self Organizing Map*. In proc. ICACSIS 2013.

Rahmawati V. 2013. *Comparison of Feature Extraction Methods Spaced K-Mers and K-mers in Fragmen Metagenome Classification using Naive Bayes Classifier* [skripsi]. Bogor-Indonesia : Institut Pertanian Bogor.

Richter DC, Felix Ott1, Auch AF, Schmid R, Huson DH. 2008. *MetaSim—A Sequencing Simulator for Genomics and Metagenomics*. PLoS ONE 3(10): e3373.doi:10.1371/journal.pone.0003373.

Rosen G,Garbarine E, Caseiro D, Polikar R,Sokhansanj B. 2008. *Metagenome Fragment Classification Using N-Mer Frequency Profiles*. Hindawi Publishing Corporation. Advances in Bioinformatics. Volume 2008, Article ID 205969, 12 pages. doi:10.1155/2008/205969.