SELEKSI FITUR YANG BERPENGARUH MENGGUNAKAN NILAI MEAN PADA KLASIFIKASI FRAGMEN *METAGENOME*

Arini Aha Pekuwali[1], Wisnu Ananta Kusuma[2] dan Agus Buono[3]
[1]*Program Studi Teknik Informatika, Universitas Kristen Wira Wacana Sumba, Jl. R. Soeprapto 35 Waingapu, Sumba Timur, Nusa Tenggara Timur*
Email: arini.pekuwali29@gmail.com
[2,3] *Departemen Ilmu Komputer, Institut Pertanian Bogor, Jl. Meranti Wing 20 Level 5 IPB Dramaga, Bogor, Jawa Barat*
Email[2]: ananta@apps.ipb.ac.id
Email[3]: pudesha@gmail.com

ABSTRAK

Pekuwali (2018) telah melakukan penelitian klasifikasi fragmen metagenome menggunakan *spaced k-mers*. Optimasi susunan fitur menggunakan Algoritma Genetika. Pekuwali (2018) menyimpulkan bahwa susunan fitur terbaik atau disebut kromosom adalah 111111110001 dengan nilai *fitness* 85,42. Kromosom 111111110001 menghasilkan 336 fitur pengekstraksi fragmen DNA. Penelitian kali ini bertujuan untuk mengetahui fitur mana saja yang berpengaruh dalam pengklafikasian dan akurasi yang dihasilkan. Metode yang digunakan adalah nilai Mean. Metode nilai mean dipilih karena sebaran data normal atau mendekati normal. Penelitian ini menyimpulkan bahwa fitur yang berpengaruh dalam pengklasifikasian adalah fitur 22 sampai 27 dengan akurasi sebesar 78,83% dan fitur 38 sampai 43 dengan akurasi sebesar 79,67%.

Kata kunci: Seleksi Fitur Yang Berpengaruh, Nilai Mean, Klasifikasi *Metagenome*

ABSTRACT

Pekuwali (2018) has conducted research on the classification of metagenome fragments using k-mers spaces. Optimize the arrangement of features using Genetic Algorithms. Pekuwali (2018) states that the best feature called chromosome is 111111110001 with a fitness value of 85.42. Chromosome 111111110001 produced 336 features of extracting DNA fragments. This research aims to find out which features are successful in classification and the resulting accuracy. The method used is the Mean value. The value method is chosen because the data distribution is normal or taken normally. This study concludes that the features needed for classification are features 22 to 27 with an accuracy of 78.83% and features 38 to 43 with an accuracy of 79.67%.

Keywords: Influential Feature Selection, Mean Value, Metagenome Classification

1. PENDAHULUAN

Metagenomika adalah studi yang mempelajari keseluruhan informasi genetik tentang organisme-organisme dari sampel yang diambil langsung dari lingkungan. Lingkungan yang dimaksud seperti tanah, air, isi perut manusia, bangunan, limbah dan lain-lain yang merupakan tempat mikroba berkembang biak (Bouchot *et al*, 2013). Tahapan dalam metagenomika diawali dengan *deoxyribonucleic acid* (DNA) *sequencing* terhadap sampel *metagenome*. Karena diambil langsung dari lingkungan, fragmen yang dihasilkan mengandung berbagai mikroorganisme. Kondisi seperti ini memungkinkan terjadinya kesalahan perakitan terhadap fragmen *metagenome*, yaitu disambungkannya fragmen antara spesies yang satu dengan fragmen dari spesies yang lain. Permasalahan ini dapat diselesaikan menggunakan metode *binning*.

Ada 2 pendekatan *binning*, yaitu homologi dan komposisi. *Binning* berbasis komposisi mengelompokan fragmen-fragmen dari berbagai organisme berdasarkan tingkat taksonominya menggunakan teknik-teknik dalam *machine learning*, yaitu ekstraksi fitur dan pengklasifikasian atau pengklusteran. Pekuwali (2018) telah melakukan penelitian dengan menggunakan *spaced k-mers* frekuensi sebagai metode pengekstraksi fitur dan *Naive Bayesian Classifier* (NBC). Penggunaan metode *spaced k-mers* frekuensi menghasilkan model posisi *match* (1) dan *don't care* (0). Oleh karena itu, Algoritma Genetika digunakan untuk mengoptimasi model posisi yang menghasilkan akurasi pengklasifikasi tertinggi. Kromosom 111111110001 yang berarti [111 1111 10001] dengan nilai *fitness* 85,42 terpilih menjadi kromosom terbaik. *Fitness* kromosom merupakan nilai akurasi klasifikasi.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

Kromosom 111111110001 menghasilkan 336 fitur yang digunakan untuk melakukan ekstraksi terhadap 10.000 fragmen DNA latih dan 4.500 fragmen DNA uji. Setiap fragmen memiliki panjang 500 bp. Pekuwali (2015) menyatakan bahwa tidak semua fitur dapat mengklasifikasikan fragmen *metagenome* dengan akurat. Oleh karena itu, penelitian kali ini ingin mengetahui fitur mana saja yang berpengaruh dalam pengklafikasian dan akurasi yang dihasilkan. Metode klasifikasi yang digunakan adalah Naive Bayes Classifier (NBC).

Rumusan masalah

Adapun masalah yang peneliti angkat adalah:

1. Bagaimana sebaran data dari ekstraksi fitur menggunakan susunan fitur atau kromosom [111 1111 10001] ?
2. Fitur mana sajakah yang berpengaruh dalam proses klasifikasi fragmen *metagenome*?
3. Berapa akurasi yang dihasilkan oleh fitur yang berpengaruh dalam proses klasifikasi fragmen *metagenome* menggunakan metode Naive Bayes Classifier?

Tujuan

Adapun tujuan dari penelitian ini, yaitu:

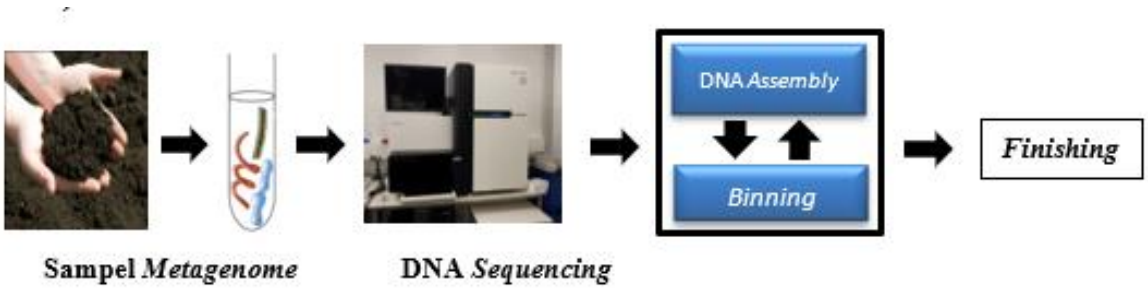
1. Mengetahui sebaran data dari ekstraksi fitur menggunakan susunan fitur atau kromosom [111 1111 10001].
2. Mengetahui fitur yang berpengaruh dalam proses klasifikasi fragmen *metagenome*.
3. Mengetahui akurasi yang dihasilkan oleh fitur yang berpengaruh dalam proses klasifikasi fragmen *metagenome* menggunakan metode Naive Bayes Classifier.

2. MATERI DAN METODE

Metagenomika

Riesenfeld et al. (2004) menyatakan bahwa metagenomika merupakan suatu teknik yang secara khusus ditujukan untuk mengumpulkan gen-gen secara langsung dari suatu lingkungan, diikuti dengan menganalisis informasi genetika yang terkandung di dalamnya. Berbeda dengan teknik analisis genom bakteri pada umumnya, teknik metagenome dilakukan dengan langsung mengekstraksi DNA genom dari lingkungan dan tidak memerlukan proses pengkulturan bakteri pada media buatan (Handelsman, 2007). Hal ini dilakukan untuk mempelajari bakteri yang tidak dapat dikultur yang diperkirakan terdapat lebih dari 99% dari populasi bakteri di seluruh lingkungan (Riesenfeld et al., 2004).

Tahapan pada metagenomika dimulai dengan pengambilan sampel dari lingkungan. Lingkungan dalam hal ini bisa berupa air laut, tanah, isi perut manusia, dan lain-lain. Sampel yang telah diambil selanjutnya diekstrak, kemudian hasil ekstrak dimasukkan ke dalam mesin sequencer. Mesin sequencer menghasilkan fragmen-fragmen DNA. Fragmen-fragmen DNA dirakit (assembly), namun bisa saja terjadi kesalahan pada tahap ini, seperti tersambungnya fragmen organisme A dengan fragmen dari organisme B. Kesalahan perakitan ini menghasilkan interspecies chimeras. Untuk menghindari terbentuknya interspecies chimeras, maka pada tahapan ini harus dilakukan perakitan (assembly) dan binning secara simultan. Binning adalah pengelompokan fragmen-fragmen metagenome sesuai tingkat taksonominya (Meyerdierks & Glockner, 2010).



Gambar 1. *Binning* sampel *metagenome* (Kusuma, 2012)

Fragmen-fragmen yang sudah tersambung dan membentuk urutan (*sequence*) DNA disebut *contigs* (*continuous sequence*), adapun *sequence* DNA yang sudah lebih panjang *contigs* tapi masih mengandung *gap* yang disebut *scaffolds*. Kemudian *scaffolds* dianotasi, yaitu pengidentifikasian *sequence* DNA.

Pendekatan berbasis homologi dan komposisi

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

Pendekatan berbasis homologi bekerja dengan cara membandingkan data *metagenome* dengan data yang ada pada basis data menggunakan BLAST (Altschul *et al.*, 1990) atau MEGAN (Huson *et al.*, 2007). Pada dasarnya pendekatan ini digunakan untuk mencari kemiripan antar *sequence* DNA melalui proses pensejajaran *sequence*. Cara kerja BLAST adalah melakukan pensejajaran *sequence* secara *local alignment* dengan mencari daerah pendek yang berdekatan atau yang bagian nukleotida yang *match* antara *metagenome sequence* dengan *sequence* di basis data (Gambar 2).

Pendekatan berbasis homologi terus berkembang, sehingga oleh Ma *et al.* (2002) mengusulkan *PatternHunter*. Metode ini diusulkan untuk memodifikasi aplikasi BLAST. Ide dasar dari metode ini adalah *seed* yang dipakai tidak perlu semua *match*. Kombinasi dari *match* dan *don't care* (*spaced seed*) dapat mempercepat proses pencarian nilai kesamaan dari 2 *sequence* DNA. Pada *spaced seed*, *match* disimbolkan dengan angka 1, sedangkan *don't care* disimbolkan dengan angka 0. Keadaan yang disebut *match* adalah nukleotida yang ada pada *metagenome sequence* wajib sama dengan nukleotida yang ada pada *sequence refrence* di basis data. Sedangkan keadaan *don't care* adalah nukleotida yang ada pada *metagenome sequence* tidak wajib sama dengan nukleotida yang ada pada *sequence refrence* di basis data (Gambar 3).

```
*  1  1  1  1  1  *  *  1  1  1  1  1  *  *  1  1  1  1  1
A  G  C  G  A  T  G  A  C  A  G  T  A  T  C  T  A  G  A  C
   |  |  |  |  |           |  |  |  |  |           |  |  |  |  |
T  G  C  G  A  T  A  G  C  A  G  T  A  A  G  C  A  G  A  T
```

Seed: 11111

Gambar 1. Mencari nilai *similarity* antara *sequence metagenome* dengan referensi *sequence* menggunakan pesejajaran *sequence* dengan *seed* (Kusuma, 2012)

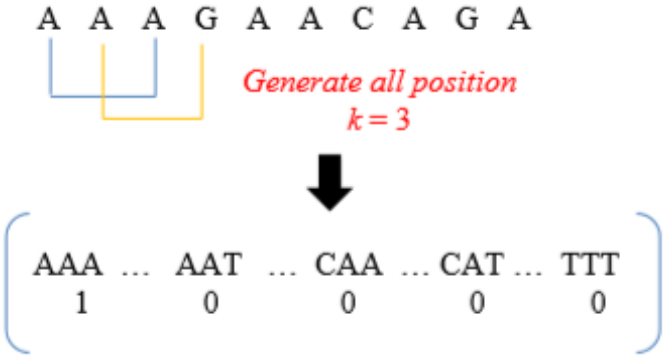
```
*  1  1  0  1  1  *  *  1  1  0  1  1  *  *  1  1  0  1  1
A  G  C  T  A  T  G  A  C  A  A  T  A  T  C  T  A  G  A  C
   |  |  |  |  |           |  |  |  |  |           |  |  |  |  |
T  G  C  C  A  T  A  G  C  A  T  T  A  A  G  C  A  G  A  T
```

Spaced Seed: 11011

Gambar 2. Mencari nilai *similarity* antara *sequence metagenome* dengan referensi *sequence* menggunakan pesejajaran *sequence* dengan *spaced seed* (Kusuma, 2012)

Berbeda dengan pendekatan homologi, pendekatan komposisi tidak perlu mensejajarkan *metagenome sequence*, tetapi dengan menghitung frekuensi kemunculan fitur dari *metagenome sequence*. *K-mers* frekuensi adalah salah satu metode pengekstraksi fitur yang dapat digunakan untuk menghitung frekuensi kemunculan fitur. Nilai *k* pada *k-mers* frekuensi adalah panjang fragmen *metagenome* (Choi & Cho, 2002).

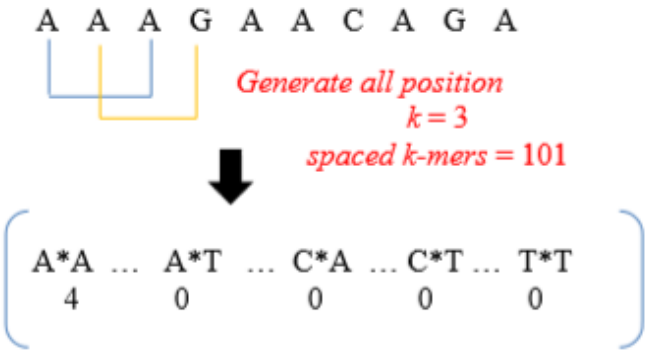
Pola kemunculan *k* dalam *sequence* dihitung menggunakan 4 basa utama, yaitu *Adenine* (A), *Cytosine* (C), *Guanine* (G) dan *Thymine* (T) yang dipangkatkan dengan rangkaian pasang basa yang ingin digunakan (pola kemunculan: 4^k, dengan $k \geq 1$). Gambar 4 menampilkan cara kerja *k-mers* frekuensi.



Gambar 4. Cara kerja *k-mers* frekuensi (Kusuma, 2012)

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

Metode yang berbasis pendekatan komposisi adalah *spaced k-mers* frekuensi (Kusuma, 2012). *Spaced k-mers* mengadopsi ide dasar *PatternHunter* ke dalam metode pengekstraksi fitur *k-mers* frekuensi. Gambar 5 menampilkan cara kerja *spaced k-mers* frekuensi.

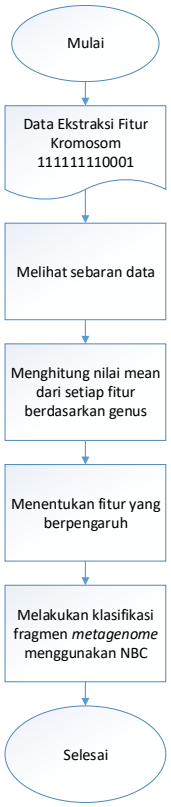


Gambar 3. Cara kerja *spaced k-mers* frekuensi (Kusuma, 2012)

METODOLOGI PENELITIAN

Tahapan-tahapan dalam penelitian ini dibagi dalam 5 bagian (Gambar 6), yaitu:

1. Pengumpulan data ekstraksi fitur menggunakan kromosom 111111110001
2. Melihat sebaran data ekstraksi fitur menggunakan kromosom 111111110001
3. Menghitung nilai mean dari masing-masing fitur berdasarkan genus
4. Menentukan fitur yang berpengaruh dalam proses klasifikasi
5. Menghitung nilai akurasi klasifikasi fitur berpengaruh menggunakan NBC



Gambar 6. Tahapan penelitian

Data ekstraksi fitur kromosom 111111110001

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

Kromosom 111111110001 menghasilkan 336 fitur. 336 fitur tersebut digunakan untuk mengekstraksi fragmen *metagenome*. Berikut ini daftar fragmen *metagenome* yang diekstraksi:

Tabel 1. Data latih

No	*Species*	Genus	Jumlah Fragmen	Panjang Fragmen
1	*Agrobacterium radiobacter K84 chromosome 2*		1000	500 bp
2	*Agrobacterium tumefaciens str. C58 chromosome circular*	*Agrobacterium*	1000	500 bp
3	*Agrobacterium vitis S4 chromosome 1*		1000	500 bp
4	*Bacillus amyloliquefaciens FZB42*		1000	500 bp
5	*Bacillus anthracis str. Ames Ancestor*	*Bacillus*	1000	500 bp
6	*Bacillus cereus 03BB102*		1000	500 bp
7	*Bacillus pseudofarmus OF4 chromosome*		1000	500 bp
8	*Staphylococcus aureus subsp. Aureus JH*		1000	500 bp
9	*Staphylococcus epidermis ATCC 12228*	*Staphylococcus*	1000	500 bp
10	*Staphylococcus haemolyticus JCSC1435*		1000	500 bp

Tabel 2. Data uji

No	*Species*	Genus	Jumlah Fragmen	Panjang Fragmen
1	*Agrobacterium radiobacter K84 chromosome 1*		500	500 bp
2	*Agrobacterium tumefaciens str. C58 chromosome linear*	*Agrobacterium*	500	500 bp
3	*Agrobacterium vitis S4 chromosome 2*		500	500 bp
4	*Bacillus thuringiensis str Al Hakam*		500	500 bp
5	*Bacillus subtilis subsp. Subtilis str 168*	*Bacillus*	500	500 bp
6	*Bacillus pumilus SAFR-032*		500	500 bp
7	*Staphylococcus carnosus*		500	500 bp
8	*Staphylococcus saprophyticus subsp. Saprophyticus ATCC 1530S*	*Staphylococcus*	500	500 bp
9	*Staphylococcus Lugdunensis HKU09-01*		500	500 bp

Sebaran data

Sebaran data dilihat menggunakan software minitab. Tujuan melihat sebaran data adalah mempermudah menentukan metode yang akan digunakan. Jika sebaran data normal, maka metode yang digunakan adalah mean. Jika sebaran data tidak normal, maka data perlu dinormalisasi terlebih dahulu.

Nilai mean dari data ekstraksi setiap fitur menurut genus

Data ekstraksi setiap fitur dikelompokan menurut genus. Kemudian dilakukan penghitungan nilai mean. Software yang digunakan MS. Office Excel. Nilai mean yang dihasilkan dibuat dalam sebuah grafik.

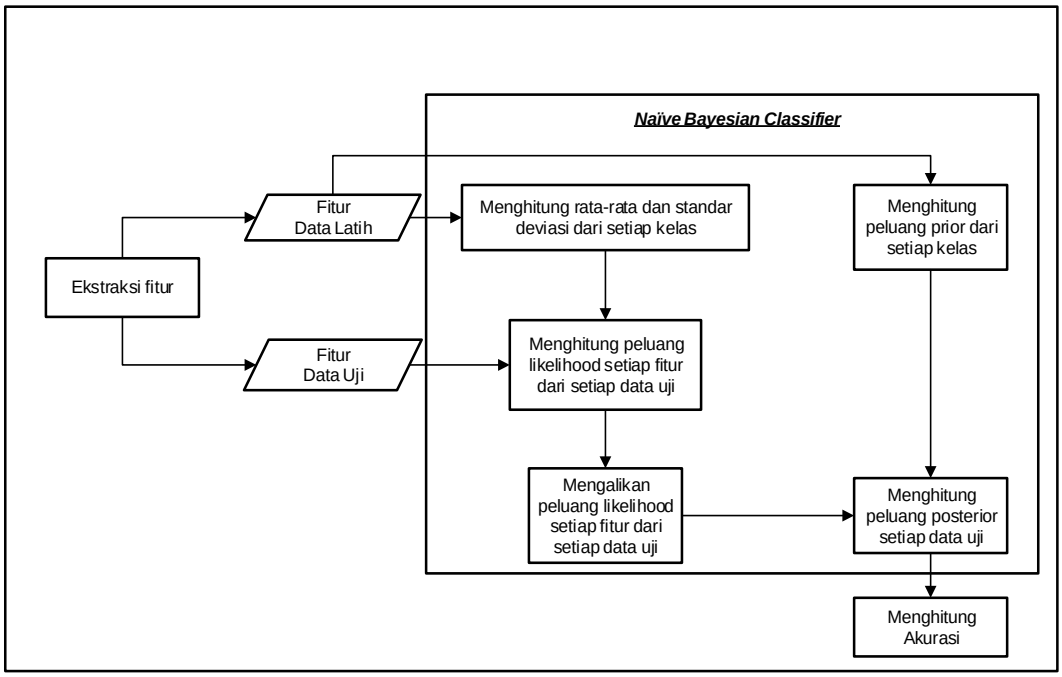
Fitur yang berpengaruh

Fitur yang berpengaruh dapat dilihat dari tabel nilai mean. Apabila nilai mean menunjukkan perbedaan signifikan antar genus sebanyak 6 fitur berturut-turut, maka fitur pada area grafik tersebut merupakan fitur berpengaruh dalam proses klasifikasi.

Klasifikasi fragmen metagenome menggunakan NBC

Proses klasifikasi fragmen metagenome menggunakan NBC dapat dilihat pada Gambar 7.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)



Gambar 7. Proses klasifikasi menggunakan NBC

Ekstraksi fitur dilakukan menggunakan fitur yang berpengaruh dalam klasifikasi. Data yang diekstraksi adalah data latih (Tabel 1) dan data uji (Tabel 2).

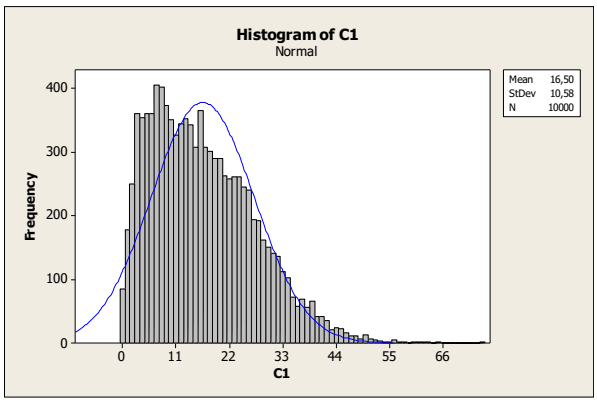
3. HASIL DAN PEMBAHASAN

Ekstraksi fitur dilakukan untuk data latih dan data uji. Metode yang digunakan dalam ekstraksi fitur dengan cara menghitung frekuensi kemunculan spaced k-mers. Model *match* (1) dan *don't care* (0) dibentuk dari hasil inisialisasi GA. Model posisi selanjutnya disebut kromosom. *Don't care* (0) berarti membolehkan pasangan basa apapun mengisi bit tersebut (Ma et al., 2002).

Tabel 3. Matriks komposisi yang terbentuk dengan kromosom 111111110001

Fitur Fragmen	AAA (1)	…	TTT (64)	AAAA (65)	...	TTTT (320)	A000A (321)	…	T000T (336)	Genus
F1	6	…	17	0	…	8	12	…	34	Agrobacterium
F2	10	…	14	4	…	6	2	…	23	Agrobacterium
…	…	…	…	…	…	…	…	…	…	…
F10000	16	…	36	7	…	13	337	…	74	Staphylococcus

Ekstrak fitur menggunakan kromosom 111111110001 menghasilkan data dengan sebaran normal (Gambar 8).



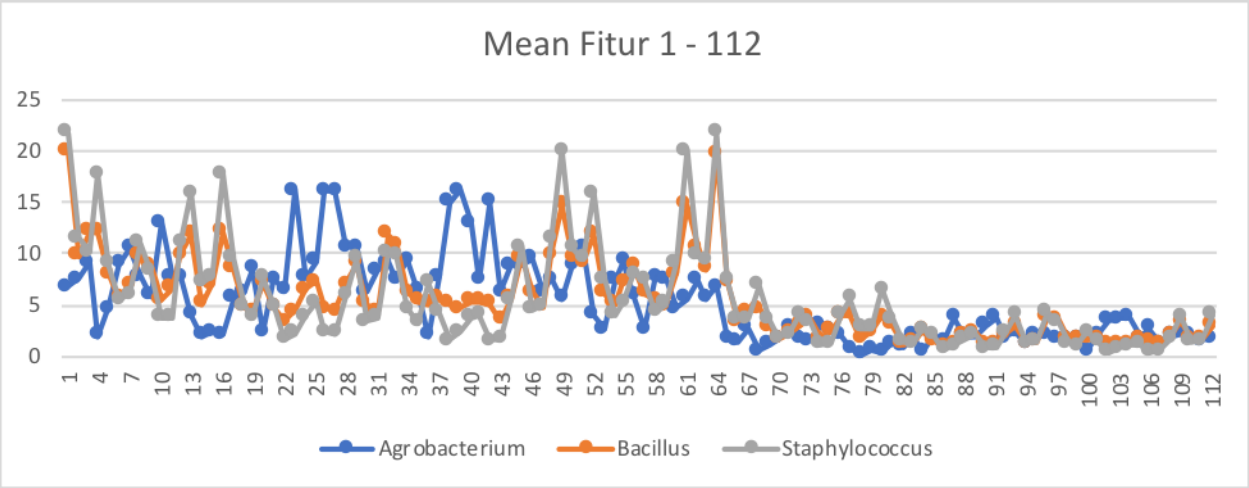
Gambar 8. Bentuk sebaran data hasil ekstraksi fitur

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

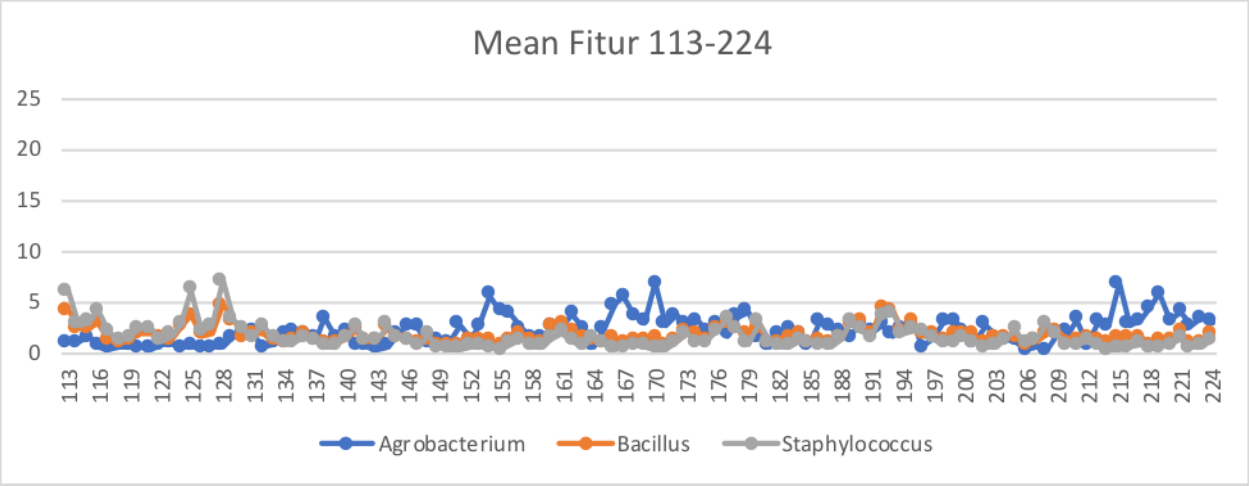
Nilai mean dari setiap fitur setelah dikelompokan menurut genus dapat dilihat pada Tabel 4. Nilai mean pada Tabel 4 disajikan dalam grafik Gambar 9a, 9b dan 9c.

Tabel 4. Nilai mean dari setiap fitur menurut genus

Fitur	Genus Agrobacterium	Genus Bacillus	Genus Staphylococcus
1	6,7197	19,9200	21,7040
2	7,5003	9,9700	11,4730
...	...	...	...
336	19,2870	47,8570	57,6060

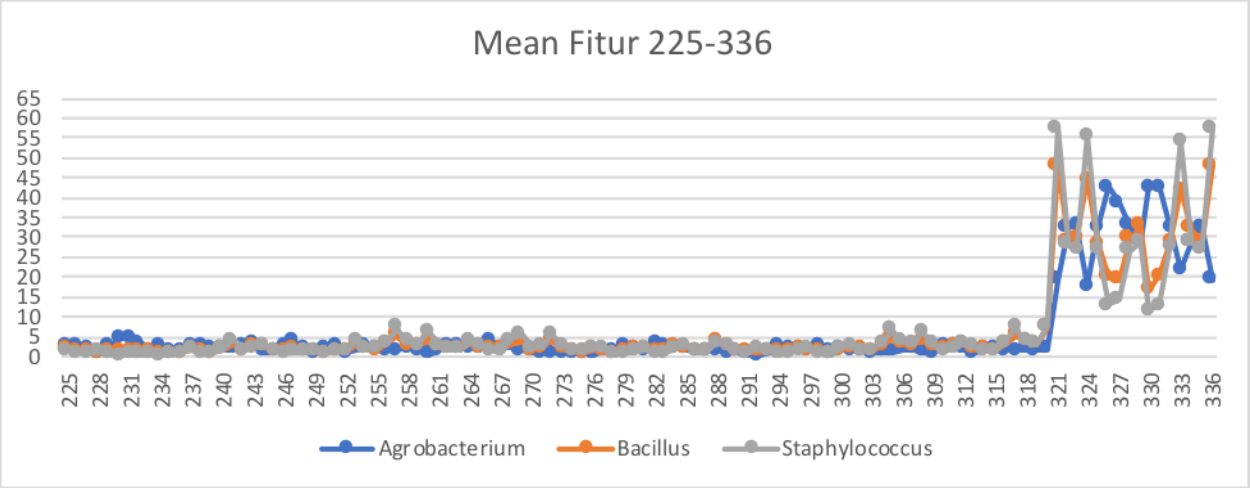


Gambar 9a. Mean fitur 1 - 112



Gambar 9b. Mean fitur 113 - 224

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)



Gambar 9c. Mean fitur 225 - 336

Pengekstraksian fitur menggunakan kromosom 111111110001, menghasilkan karakteristik kelas. Karakteristik kelas didapatkan melalui nilai *mean* dari kemunculan fitur untuk setiap genus. Dari 336 fitur yang dihasilkan diketahui bahwa pada fitur 22 sampai fitur 27 (Tabel 5) memiliki nilai *mean* yang memiliki perbedaan signifikan untuk setiap genus. Pada fitur 38 sampai 43 (Tabel 6) juga menunjukkan hal yang sama.

Tabel 5. Nilai mean fitur 22 sampai 27 menurut genus

Fitur	Genus Agrobacterium	Genus Bacillus	Genus Staphylococcus
22 (AACC)	6,4053	3,4182	1,6510
23 (AACG)	16,1107	4,4232	2,2866
24 (AACT)	7,7233	6,5272	3,8663
25 (AAGA)	9,2626	7,2920	5,2503
26 (AAGC)	16,0810	4,6527	2,4130
27 (AAGG)	15,9713	4,4600	2,2490

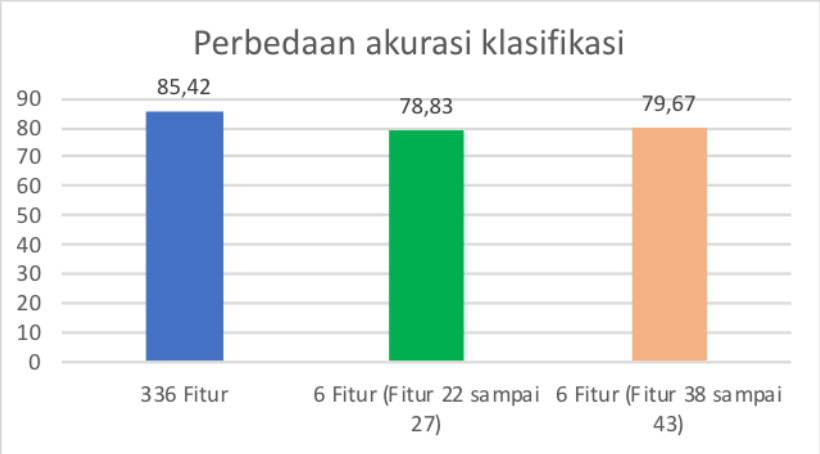
Tabel 6. Nilai mean fitur 38 sampai 43 menurut genus

Fitur	Genus Agrobacterium	Genus Bacillus	Genus Staphylococcus
38 (CACC)	14,9570	5,0372	1,5690
39 (CACG)	16,1240	4,7082	2,3533
40 (CACT)	13,0170	5,3375	3,7776
41 (CAGA)	7,4170	5,2850	4,0576
42 (CAGC)	14,9353	5,0992	1,6053
43 (CAGG)	6,3060	3,4975	1,6613

Telah dilakukan pengklasifikasian yang menggunakan fitur yang berpengaruh pada proses klasifikasi. Klasifikasi menggunakan fitur 22 sampai 27 menghasilkan akurasi sebesar 78,83%. Sedangkan saat menggunakan fitur 38 sampai 43 menghasilkan akurasi sebesar 79,67%. Pada Gambar 10 dapat dilihat bahwa dengan menggunakan 6 fitur menghasilkan akurasi yang tidak jauh berbeda dengan yang melibatkan 336 fitur (Pekuwali, 2018).

Hasil ini menunjukkan bahwa ada fitur yang berpengaruh dalam pengklasifikasian dan juga ada fitur yang tidak terlalu berpengaruh saat pengklasifikasian dilakukan. Sehingga dapat disimpulkan tidak semua fitur dapat mengklasifikasikan fragmen metagenome dengan akurat.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)



Gambar 10. Perbedaan akurasi klasifikasi

4. KESIMPULAN

Ekstrak fitur menggunakan kromosom 111111110001 menghasilkan data dengan sebaran normal. Pengekstraksian fitur menggunakan kromosom 111111110001, menghasilkan karakteristik kelas. Karakteristik kelas didapatkan melalui nilai *mean* dari kemunculan fitur untuk setiap genus. Dari 336 fitur yang dihasilkan diketahui bahwa pada fitur 22 sampai fitur 27 memiliki nilai *mean* yang memiliki perbedaan signifikan untuk setiap genus. Pada fitur 38 sampai 43 juga menunjukkan hal yang sama. Klasifikasi menggunakan fitur 22 sampai 27 menghasilkan akurasi sebesar 78,83%. Sedangkan saat menggunakan fitur 38 sampai 43 menghasilkan akurasi sebesar 79,67%.

DAFTAR PUSTAKA

Altschul, S., Gish, W., Miller, W., Myers, E. dan Lipman D. (1990). "Basic local alignment search tool". *Journal of Molecular Biology*. 215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

Bouchot, J.L., Trimble, W.L., Ditzler, G., Lan, Y., Essinger, S., dan Rosen, G. (2013). *Advances in machine learning for processing and comparison of metagenomic data* [internet]. [diunduh 15 September 2014]. Tersedia pada: http://www.sessinger.com/Publications/*metagenomics_advances_ML_preprint.pdf*

Choi, J.H. dan Cho, H.G. (2002). "Analysis of common k-mers for whole genome sequence using SSB-tree". *Genome Information*. 13: 30-41

Handelsman, J. (2007). *The new science of metagenomics: Revealing the secrets of our microbial planet.* The National Academics Press, Washington, USA.

Huson, D.H., Auch, A.F. dan Schuster, S.C. (2007). "MEGAN analysis of metagenomic data". *Genome Research*. 17(3):337-386. doi: 10.1101/gr.5969107.

Kusuma, W.A. (2012). *Combined approaches for improving the performance of denovo DNA sequence assembly and metagenomic classification of short fragments from next generation sequencer* [disertasi]. Tokyo Institute of Technology, Tokyo, Japan.

Ma, B., Tromp, J., dan Li, M. (2002). PatternHunter: faster and more sensitive homology search. *Bioinformatics*. 18(3):440-445.

Meyerdierks, A. dan Glockner, F.O. (2010). "Metagenome Analysis". *Advances in Marine Genomics* 1. doi:10.1007/978-90-481-8639-6_2.

Pekuwali, A.A. (2015). *Optimasi Pengekstraksi Fitur Spaced K-Mers Frekuensi Menggunakan Algoritme Genetika Pada Pengklasifikasian Fragmen Metagenome* [tesis]. Institut Pertanian Bogor, Bogor, Indonesia.

Pekuwali, A.A, Kusuma, W.A. dan Buono, A. (2018). "Optimization of Spaced K-mer Frequency Feature Extraction using Genetic Algorithms for Metagenome Fragment Classification". *Journal of ICT Research and Applications*, 12(2), 123-137 DOI: 10.5614/itbj.ict.res.appl.2018.12.2.2.

Riesenfeld, C.S., Schloss, P.D. dan Handelsman, J. (2004). "Metagenomics: genomic analysis of microbial communities". *Annual Review of Genetics*. 38:525–552. DOI:10.1146/annurev.genet.38. 072902.091216

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)