ISSN : 1411-1624
e-ISSN: 2621-8089

CLASSIFICATION FRAGMEN METAGENOM MENGGUNAKAN *PRINCIPAL COMPONENT ANALYSIS NEIGHBOR*

Surianti
Dosen STMIK Umel Mandiri
Jalan Raya Abepura Depan Perpustakaan Daerah Kotaraja
Sur-el: surianti12p@gmail.com

Abstract: *Metagenomics is a study of metagenom analysis which its genetic materials is obtained directly from environmental samples. The process of metagenome sequencing produce fragments from mixture organisms. Thus, assembling fragments directly will generate chimeric contigs. Furthermore, a bining process is required to classify these fragments into a particular taxonomic level. In this study, the classification of metagenome fragment were extracted using n-mers, reduced its dimension using principal component analysis and classified using knearest neighbor. The experiments were conducted from in the various fragment length from 0.5 Kbp to 10 Kbp. The best results were obtained using KNN with k=7 and implementing 4-mers frequency. The accuracies of classifying known organisms obtained using PCA 95% were ranged from 91.6% to 99.9%. Moreover, the accuracies were slightly decreased when classifying unknown organisms, from 89.64% to 99.32%.*

Keywords: *Classification, Fragments metagenom, n-mers*

Abstrak: *Metagenomika adalah ilmu yang mempelajari tentang analisis metagenom yang materi genetiknya diperoleh langsung dari sampel lingkungan. Ketika mengsekuens sampel metagenom ini maka akan dihasilkan fragmen-fragmen. Pada saat fragmen-fragmen tersebut dirakit akan dihasilkan chimeric contigs atau gabungan fragmen dari berbagai organisme. Selanjutnya diperlukan proses binning yang bertujuan untuk mengklasifikasikan fragmen-fragmen tersebut ke dalam tingkat taksonomi tertentu. Pada penelitian ini peneliti melakukan klasifikasi fragmen metagenom yang diekstrasi menggunakan n-mers kemudian direduksi dimensinya menggunakan principal component analysis dan diklasifikasi menggunakan knearest neighbor. Nilai k yang terbaik pada KNN adalah 7. Nilai n tertinggi pada n-mers adalah 4. Akurasi pada organisme dikenal dari fold terbaik dengan menggunakan PCA 95% untuk panjang fragmen 0.5 Kbp sampai 10 Kbp berkisar antara 91.6% sampai 99,9%. Untuk organisme tidak dikenal dengan PCA 95% tingkat akurasi berkisar antara 89.64% sampai 99.32%.*

Kata kunci*: fragmen metagenom, Growing Self Organizing Map, Pengelompokan*

1. PENDAHULUAN

Penelitian tentang analisis metagenom dalam lingkup bioinformatika terus berkembang. Secara umum, analisis materi genetik dilakukan dengan cara membudidayakannya di laboratorium, kemudian di-s*equencing* dan dilakukan perakitan. Proses ini dilakukan untuk menghasilkan urutan rantai DNA yang berisi informasi genetik suatu organisme. Akan tetapi, dari banyak mikroorganisme hanya 1% yang dapat dikulturkan. Sisanya harus mengambil sampel langsung dari lingkungan. Ilmu yang mempelajari tentang analisis metagenom dan materi genetiknya diperoleh langsung dari sampel lingkungan disebut metagenomika [1]. Sampel ini ketika di-*sequencing* akan menghasilkan fragmen-fragmen. Fragmen-fragmen yang berasal dari berbagai organisme. Pada saat dilakukan perakitan fragmen-fragmen ini, akan menghasilkan *chimeric contigs* gabungan fragmen yang berasal dari organisme

ISSN : 1411-1624
e-ISSN: 2621-8089

berbeda. Untuk itu diperlukan proses *binning* yang bertujuan untuk mengklasifikasikan fragmen-fragmen tersebut ke dalam tingkat taksonomi tertentu. *Low-abundance* pada fragmen metagenom yang berukuran besar sering menimbulkan kendala dalam perakitan genom dan menyebabkan mikrob sulit dikelompokan secara filogenetik [2]. Kesalahan dalam perakitan fragmen metagenom disebut interspecies chimeras [3]. Untuk menyelesaikan permasalahan tersebut, binning digunakan untuk mengelompokan mikrob berdasarkan tingkatan taksonomi. Ada dua pendekatan binning, yaitu berdasarkan homologi dan berdasarkan komposisi. Binning berdasarkan homologi melakukan pencarian penjajaran sekuens dengan membandingkan fragmen metagenom dengan basis data sekuens antara lain *National Centre for Biotechnology Information* (NCBI) dan hasilnya akan disimpulkan pada tiap level taksonomi. Hal tersebut menyebabkan pendekatan dengan homologi membutuhkan banyak waktu dalam proses pengelompokan.

Pendekatan kedua adalah pendekatan berdasarkan komposisi. Pendekatan ini menggunakan pasangan basa hasil ekstraksi fitur sebagai masukkan untuk pembelajaran dengan contoh (*supervised*) atau pembelajaran dengan observasi (*unsupervised*). Tidak seperti pendekatan secara homologi, pendekatan secara komposisi tidak perlu membandingkan dan menyimpulkan setiap hasil pencarian pada tiap level taksonomi sehingga waktu yang diperlukan untuk pengelompokan lebih cepat dibandingkan dengan pendekatan secara homologi. Sebagian besar proses binning masih menggunakan pembelajaran dengan contoh (*supervised learning*). Pembelajaran dengan contoh bergantung pada ketersediaan data latih padahal data latih yang tersedia tidak cukup merepresentasikan keragaman mikrob [4]. Pembelajaran dengan observasi (*unsupervised learning*) memberikan solusi terhadap keterbatasan data latih yang tersedia karena *unsupervised learning* akan menyusun data fragmen metagenom secara lebih terstruktur sebelum perbandingan sekuens dilakukan. Dengan demikian fragmen metagenom akan lebih cepat dan lebih kuat (*robust*) untuk dirakit [5]. Hasil yang didapatkan adalah pada pengelompokan mikrob dengan empat frekuensi oligonukleotida (di-, tri-, tetra-, dan pentanukleotida) pada tiga dataset mikrob, pengelompokan menggunakan frekuensi dinukleotida tidak terlalu memberikan hasil yang baik sehingga disarankan utuk menggunakan frekuensi oligonukleotida yang lebih tinggi. Oleh karena itu, penelitian ini melakukan klasifikasi fragmen metagenom menggunakan n-mers sebagai ekstrasi ciri, kemudian dilakukan pereduksian dimensi menggunakan *principal component analysis* dan diklasifikasikan menggunakan algoritme k-nearest neighbor. Akurasi yang diperoleh akan dibandingkan dengan penelitian [6], dan [7].

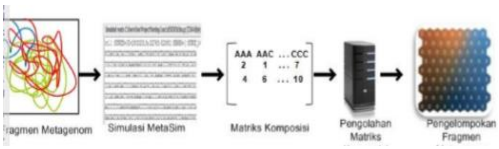
Penelitian fragmen metagenom menggunakan *unsupervised learning* umumnya hanya menggunakan komunitas yang kecil. Sedangkan untuk ekstraksi ciri, pengelompokan fragmen metagenom masih menggunakan k-mer dan belum memperhatikan kondisi *don't care.* Ekstraksi ciri dengan memperhatikan kondisi *don't care* disebut dengan *spaced k-mer* [7]. *Spaced k-mer* menyediakan vektor berdimensi lebih kecil yang berisi informasi yang lebih kaya dan berguna dibandingkan dengan vektor masukan hasil ekstraksi fitur menggunakan *k-mer* [8]. Pada penelitian ini

ISSN : 1411-1624
e-ISSN: 2621-8089

digunakan komunitas spesies yang cukup besar, yaitu 300 spesies dan data spesies tersebut diambil dari basis data NCBI. Panjang fragmen yang digunakan adalah 1 kbp dengan frekuensi *oligonukleotida trinukleotida* dan *tetranukleotida.* Alasan digunakan fragmen yang pendek karena pada penelitan terdahulu, panjang fragmen yang digunakan adalah fragmen yang panjang (≥ 8 kbp). Pada penelitian ini hendak mengatasi kelemahan dari penggunaan fragmen pendek dalam pengelompokan fragmen metagenom. Selain itu, penelitian ini menggunakan kondisi *don't care* untuk menghitung hasil matriks komposisi. Hasil dari pengelompokan fragmen metagenom tersebut akan diuji efektifitas dan efisiensinya.

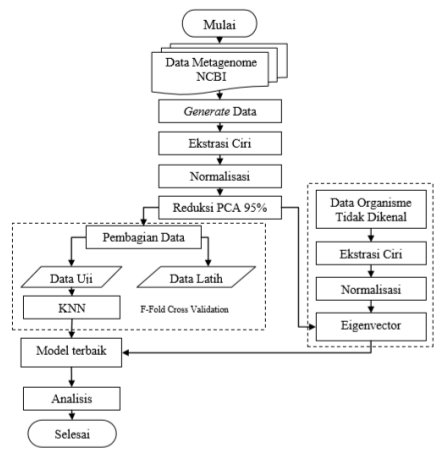
2. METODOLOGI PENELITIAN

Penelitian ini menggunakan data fragmen metagenom dari 300 mikrob dan kemudian dikelompokan berdasarkan tingkat taksonomi filum. Teknik pengambilan data fragmen metagenom yang digunakan adalah *cluster sampling*. Teknik *cluster sampling* adalah teknik yang menggunakan sampel yang memiliki jumlah item yang banyak pada suatu kelompok atau koleksi dan merupakan teknik yang sederhana serta rendah biaya [9]. Sesuai dengan tujuan penelitian ini, metode GSOM digunakan untuk pengelompokan fragmen metagenom. Data awal akan disimulasi menggunakan MetaSim [10] dan menghasilkan sekuens DNA. Hasil simulasi ini yang akan digunakan pada pengekstraksian ciri sehingga didapat matriks komposisinya. Selanjutnya fragmen metagenom akan dikelompokan menjadi 20 kelompok yang berbeda berdasarkan kesamaan dari pemetaan yang dihasilkan. Ilustrasi pemetaan fragmen metagenom, ditunjukkan pada Gambar 1.



Gambar 1. Skema penelitian *metagenome*

Pengelompokan fragmen metagenom terdiri atas beberapa tahap, yaitu data akan diekstraksi ciri untuk mendapatkan matriks komposisi, praproses data, dan dikelompokan dengan metode GSOM untuk mendapatkan model pembelajaran. Hasil pembelajaran dengan metode GSOM mampu memetakan data fragmen metagenom berdasarkan tingkat taksonomi filum. Tahap akhir adalah evaluasi terhadap hasil pengelompokan untuk mengetahui efektifitas dan efisiensi pemetaan dengan GSOM. Tahap yang dilakukan untuk pengelompokan fragmen metagenom digambarkan pada Gambar 2.

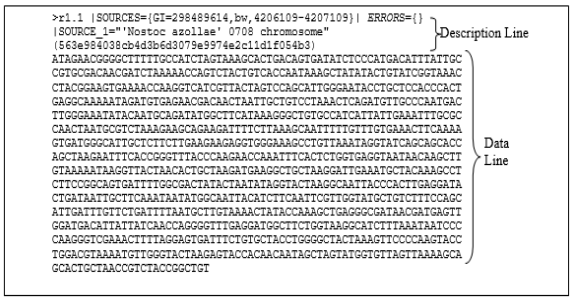


Gambar 2. Prosedur penelitian

Data yang digunakan adalah *super kingdom bacteria* dan merupakan hasil simulasi

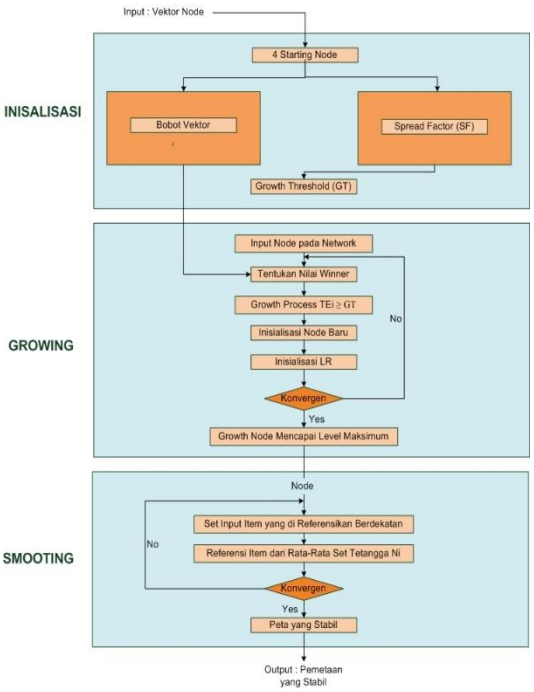
ISSN : 1411-1624
e-ISSN: 2621-8089

sampel metagenomik yang diambil dari basis data NCBI. Pengelompokan fragmen metegenom didasarkan pada tingkat taksonomi filum, yaitu sebanyak dua puluh filum dan untuk simulasi fragmen metagenom digunakan simulator MetaSim [10] panjang fragmen seragam, yaitu 1 kbp. Data yang digunakan berformat FNA (*FASTA Nucleic Acid*). Total mikrob yang digunakan adalah 300 mikrob yang nantinya akan dikelompokan pada 20 filum yang berbeda. Contoh data hasil simulasi dengan MetaSim dapat dilihat pada gambar 3.



Gambar 3. Contoh data hasil simulasi

Jumlah data adalah 200 mikrob untuk data latih dengan total jumlah fragmen yang digunakan adalah 200000 fragmen. Sedangkan untuk data uji digunakan 100 mikrob dengan total jumlah fragmen sebanyak 100 000 fragmen. Perkiraan fragmen per mikrob adalah sebanyak 1000 fragmen. Frekuensi oligonukleotida yang digunakan juga beragam untuk masing-masing dataset, yaitu trinukleotida, tetranukleotida, dan juga menggunakan *spaced k-mer*. Pengelompokan fragmen metagenom dilakukan dengan GSOM. Arsitektur metode GSOM terdiri dari beberapa fase, yaitu fase inisialisasi, fase *growing*, dan fase *smoothing*. Untuk melakukan pengelompokan data, awalnya dilakukan inisialisasi bobot vektor (biasanya di inisialisiasi empat node), nilai *Growth Threshold* (GT) yang digunakan sebagai batasan dari topologi peta berdasarkan nilai penyebaran atau *Spread Factor* (SF) dan dimensi dataset *D* (dimensi *D* adalah pembacaan fragmen metagenom × frekuensi oligonukleotida). *GT* jika diformulasikan adalah sebagai berikut $GT = -D \times \ln(SF)$. Selain itu dilakukan pembobotan vektor dari tiap pembacaan fragmen metagenom dan pembacaan pada penelitian ini dilakukan sebanyak 100 000 fragmen untuk data uji dan 200 000 fragmen untuk data latih. Gambar 4 menampilkan blok diagram pengelompokan dengan metode GSOM.

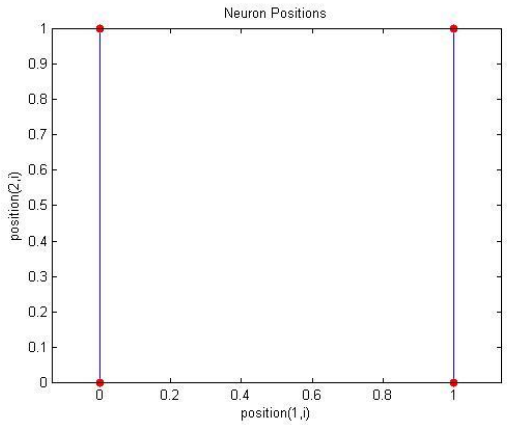


Gambar 4. Blok diagram pengelompokan GSOM

Perhitungan GT digunakan untuk menentukan dan mendapatkan hasil topologi peta yang ideal. Untuk mendapatkan hasil peta yang ideal, maka harus ditentukan penyebaran dari titik-titik neuron. Pengontrolan ini ditentukan oleh nilai *Spread Factor* (SF). Nilai SF pada penelitian ini digunakan berbeda pada

ISSN : 1411-1624
e-ISSN: 2621-8089

tiap frekuensi, yaitu 0.6 untuk frekuensi trinukleotida dan *spaced k-mer*,dan 0.8 untuk frekuensi tetranukleotida.

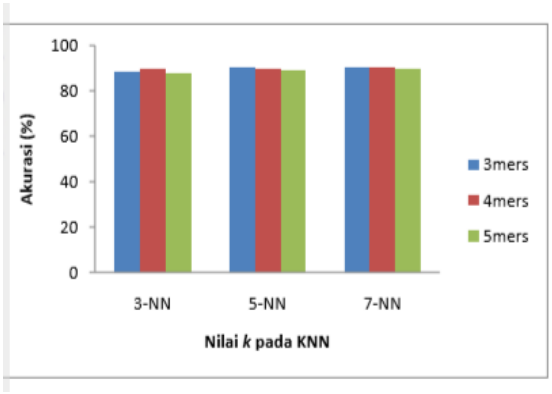


Gambar 5. Inisialisasi starting node

3. HASIL DAN PEMBAHASAN

Pengelompokan fragmen metagenom dikembangkan dengan bahasa pemrograman C++ dan Matlab 7.11.0 (R2010b). Fragmen metagenom akan dikelompokan dalam 20 kategori, yaitu 20 filum berdasarkan *NCBI Taxonomy Browser*. Penelitian ini menggunakan 300 mikrob. Mikrob dikelompokan berdasarkan tingkat taksonomi filum. Data yang digunakan diunduh pada basis data NCBI. Setelah diunduh, data tersebut disimulasikan menggunakan MetaSim. Hasil simulasi akan diekstraksi dan menghasilkan matriks komposisi yang digunakan sebagai model pembelajaran. Jumlah data fragmen metagenom yang digunakan yaitu data latih terdiri dari 200 dan data uji terdiri dari 100 data mikrob.

Fragmen metagenom hasil simulasi MetaSim akan diekstraksi dengan *k-mer frequency*. Ekstraksi dengan k-mer akan membentuk matriks komposisi sesuai dengan berapa banyak data yang dibangkitkan dan frekuensi oligonukleotida yang digunakan. Frekuensi fragmen metagenom yang diekstraksi dengan *k-mer frequency* adalah trinukleotida dan tetranukleotida. Banyak data yang dibangkitkan adalah 200 000 untuk data latih dan 100 000 untuk data uji. Fitur yang digunakan adalah sebanyak 64 untuk trinukleotida, dan 256 untuk tetranukleotida. Sehingga didapat perhitungan untuk tiap frekuensi oligonukleotida akan diperoleh matriks komposisi dengan ukuran $200\,000 \times 64$, $200\,000 \times 256$, $100\,000 \times 64$, dan $100\,000 \times 256$ masing-masing untuk data latih dan data uji. Dari hasil akurasi untuk organisme dikenal, setiap nilai fold tertinggi dari beragam nilai n-mers dan KNN akan digunakan untuk pengujian organisme tidak dikenal. Hasil akurasi selengkapnya untuk panjang fragmen 1 Kbp dan 5 Kbp.

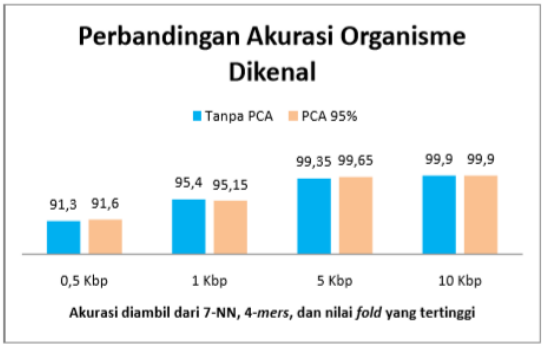


Gambar 6. Akurasi terhadap nilai *K* dan *N*

Perbandingan akurasi organisme menggunakan PCA dan tanpa PCA untuk organisme dikenal dapat dilihat pada Gambar 11. Secara umum, hasil akurasi yang diperoleh menggunakan PCA dan tanpa PCA tidak jauh berbeda. Untuk panjang fragmen 0,5 Kbp dan 5 Kbp akurasi PCA lebih tinggi. Tetapi untuk

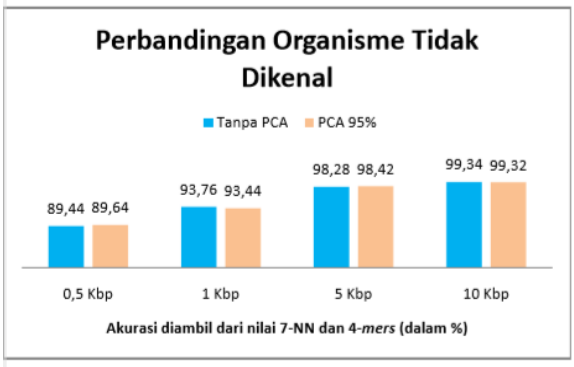
ISSN : 1411-1624
e-ISSN: 2621-8089

panjang 1 Kbp lebih tinggi akurasi tanpa PCA. Untuk panjang 10Kbp akurasinya sama. Sehingga dapat ditarik kesimpulan, walaupun dimensi matriks sudah direduksi tetapi akurasi menggunakan PCA dan tanpa PCA tidak berbeda jauh.



Gambar 7. Perbandingan akurasi organisme dikenal

Perbandingan akurasi organisme menggunakan PCA dan tanpa PCA untuk organisme tidak dikenal dapat dilihat pada Gambar 12. Hasil yang diperoleh tidak berbeda jauh dengan organisme dikenal. Untuk panjang fragmen 0,5 Kbp dan 5 Kbp akurasi PCA lebih tinggi. Tetapi untuk panjang 1 Kbp dan 10 Kbp lebih tinggi akurasi tanpa PCA. Sehingga dapat disimpulkan, walaupun dimensi matriks sudah direduksi tetapi akurasi menggunakan PCA dan tanpa PCA tidak berbeda jauh.



Gambar 8. Perbandingan akurasi organisme tidak dikenal

4. KESIMPULAN

Pada penelitian ini dilakukan klasifikasi fragmen metagenom menggunakan metode K-Nearest Neighbor dan direduksi dimensi menggunakan Principal Component Analysis. Untuk nilai k yang terbaik pada KNN adalah 7-NN. Untuk nilai n tertinggi pada n-mers adalah 4-mers. Akurasi pada organisme dikenal dari fold terbaik dengan menggunakan PCA 95% untuk panjang fragmen 0.5 Kbp sampai10 Kbp berkisar antara 91.6% sampai 99,9%.

Tanpa PCA diperoleh akurasi berkisar antara 91.3% sampai 99.9%. Untuk organisme tidak dikenal dengan PCA 95% akurasi yang diperoleh berkisar antara 89.64% sampai 99.32%. Sedangkan tanpa PCA akurasi yang diperoleh berkisar antara 89.44% sampai 99.34%. Selain itu, waktu komputasi dengan menggunakan PCA mengalami penurunan walaupun panjang fragmen semakin meningkat. Selisih waktu komputasi setelah direduksi mencapai 88,109 detik pada 5-mers dengan panjang 10 Kbp. Hasil akurasi yang diperoleh seluruhnya cukup baik, baik menggunakan PCA dan tanpa PCA. PCA mampu menghasilkan akurasi yang tidak berbeda jauh dengan tanpa PCA, selain itu waktu komputasi juga dapat direduksi. Setelah dibandingkan dengan penelitian terkait Kusuma 2014, dapat dilihat bahwa akurasi yang diperoleh pada penelitian ini lebih tinggi dari penelitian sebelumnya.

ISSN : 1411-1624
e-ISSN: 2621-8089

DAFTAR PUSTAKA

[1] Wu H. PCA–based Linear Combinations of Oligonucleotide Frequencies for Metagenomic DNA Fragment Binning. *IEEE Symposium on CIBCB*. 8: 46-53. 2013.

[2] Chan CK, Hsu AL, Tang SL, Halgamuge SK. 2012. Using Growing Self-Organizing Maps to Prove the Binning Process in Environmental Whole-Genome Shotgun Equencing. *Journal of Biomedicine and Biotechnology*. 2013.

[3] Meyerdierks A, Glockner FO. Metagenome Analysis. *Advances in Marine Genomics*. 1: 33 – 71. 2014.

[4] Prabhakara S, Acharya R. Unsupervised Two-Way Clustering of Metagenomic Sequence. *Journal of Biomedicine and Biotechnology.* 2012.

[5] Nasser S, Brelan A, Harris FC, Nicolescu M. A Fuzzy Classifier to Taxonomically Group DNA Fragments within A Metagenome. *Proc. Annual Meeting of the NAFIPS 08*. 8: 1-6. 2014.

[6] Ellyana F. Klasifikasi Fragmen Metagenom Menggunakan Fitur Spaced NMers dan K-Nearest Neighbor [skripsi]. Bogor (ID): Institut Pertanian Bogor. 2014.

[7] Kusuma WA. *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. 2014.

[8] Kusuma Y. Metagenome fragment binning based on characterization vector. *International Conference on Bioinformatics and Biomedical Technology* (*ICBBT*); Mar 25–27. 2012.

[9] Sheaffer RL, Mendenhall W, Ott RL. Elementary Survey Sampling. *4th ed. Boston (US): PWS – KENT Publishing Company*. pp. 116-119. 2012.

[10] Richter DC, Ott F, Auch AF, Schmid R, Hudson DH. MetaSim-Sequencing Simulator for Genomics and Metagenomics. *PLoS ONE*. 3(10). 2012.