JSI : Jurnal Sistem Informasi *(E-Journal), VOL.12, NO.2, Oktober 2020*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

Clustering Fragmen Metagenom Menggunakan Metode Growing Self Organizing Map (GSOM) (Studi Kasus Dinas Lingkungan Hidup Kota Jayapura)

Nur Ain Banyal*[1], Surianti [2]
[1,2]STMIK Umel Mandiri Jayapura
e-mail:: **nur.ain.banyal@gmail.com [1], surianti12p@gmail.com [2]**

Abstrak

Metagenome merupakan mikroorganisme yang diambil secara langsung dari alam. Proses sequenc inggenom dari metagenome mengakibatkan bercampurnya berbagai organisme. Metode pengumpulan data yang digunakan adalah Observasi,menggunakan teknik terjun langsung ke lapangan pada Dinas Lingkungan Hidup Kota Jayapura. Penelitian ini menggunakan data fragmen metagenom dari 300 mikrob. Teknik pengambilan data fragmen metagenom yang digunakan adalah cluster sampling. Lokasi penelitian bertempat di Dinas Lingkungan Hidup Kota Jayapura. Tujuan penelitian ini adalah untk menganalisis efektifitas dan efisiensi metode Growing Self Organizing Map dalam pengelompokan mikrob yang berskala besar dengan panjang fragmen yang pendek berdasarkan frekuensi oligonukleotida. Untuk ekstraksi fitur, digunakan k-mer frequency dan spaced. Digunakan fragmen yang pendek karena pada penelitan terdahulu, panjang fragmen yang digunakan adalah fragmen yang panjang (≥ 8 kbp), sehingga pada penelitian ini hendak mengatasi kelemahan dari penggunaan fragmen pendek dalam pengelompokan fragmen metagenom. Hasil dari pengelompokan fragmen metagenom tersebut akan di uji efektifitas dan efisiensinya.

Kata kunci: *Clustering, Growing Self Organizing Map, Metagenom.*

Abstract

Metagenome is a microorganism that is taken directly from nature. The sequencen process of metagenome results in the mixing of various organisms. The data collection method used was Observation, using a direct plunge technique to the Jayapura City Environmental Department. This study uses data from metagenome fragments of 300 microbes. The technique of collecting metagenome fragment data used is cluster sampling. The research location is at the Jayapura City Environmental Agency. The purpose of this study is to analyze the effectiveness and efficiency of the Growing Self Organizing Map method in large-scale microbial grouping with short fragment lengths based on oligonucleotide frequencies. For feature extraction, k-mer frequency and spaced are used. Short fragments are used because in previous studies, the length of the fragment used was a long fragment (≥ 8 kbp), so in this study we want to overcome the drawbacks of using short fragments in the grouping of metagenome fragments. The results of the grouping of metagenome fragments will be tested for effectiveness and efficiency.

Keywords*: Clustering, Growing Self Organizing Map, Metagenom.*

1. PENDAHULUAN

Penelitian tentang analisis metagenom dalam lingkup bioinformatika terus berkembang. Secara umum, analisis materi genetik dilakukan dengan cara membudidayakannya di laboratorium, kemudian di-sequencing 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],[2],[3]Salah satu contoh dari kesulitan

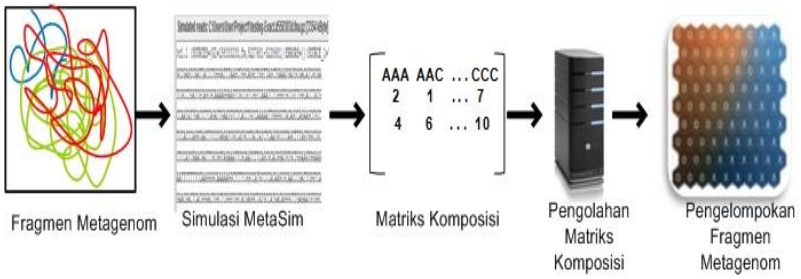
untuk isolasi lagsung dari lingkungan adalah proyek laut Sargasso. Low-abundance adalah rendahnya representasi relatif keanekaragaman mikrob dalam sampel lingkungan sehingga masih banyak mikrob yang belum dikenali dan dimanfaatkan. Low-abundance pada fragmen metagenom yang berukuran besar sering menimbulkan kendala dalam perakitan genom dan menyebabkan mikrob sulit dikelompokan secara filogenetik. Kesalahan dalam perakitan fragmen metagenom disebut dengan interspesies chimeras, Sebagai solusi masalah, binning digunakan untuk mengelompokan mikrob berdasarkan tingkatan taksonomi.[4],[5].Ada dua pendekatan binning, yaitu berdasarkan homologi dan berdasarkan komposisi. Binning berdasarkan homologi akan melakukan pencarian penjajaran sekuens dengan membandingkan fragmen metagenom dengan basis data yang digunakan, yaitu National Centre for Biotechnology Information (NCBI) dan hasilnya akan disimpulkan pada tiap level taksonomi.[6],[7]. Hal tersebut menyebabkan pendekatan dengan homologi membutuhkan banyak waktu dalam proses pengelompokan.Contoh metode yang menggunakan pendekatan homologi adalah BLAST.Pada saat dilakukan perakitan fragmen-fragmen ini, akan menghasilkan chimeric contigs gabungan fragmen yang berasal dari organisme berbeda. Dari beberapa pendekatan binning berdasarkan komposisi dengan unsupervised learning,metode GSOM memberikan hasil terbaik dalam pemetaan fragmen metagenom. Karena itu, pada penelitian tentang pengelompokan fragmen metagenom ini akan menggunakan metode GSOM. [8],[9]. 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. Penelitian fragmen metagenom pada penelitian ini menggunakan komunitas spesies yang cukup besar, yaitu 300 spesies dan data spesies tersebut diambil dari basis data NCBI.[10],[11]. Panjang fragmen yang digunakan adalah 1 kbp dengan frekuensi oligonukleotida trinukleotida dan tetranukleotida. Digunakan fragmen yang pendek karena pada penelitan terdahulu, panjang fragmen yang digunakan adalah fragmen yang panjang (≥ 8 kbp), sehingga 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 di uji efektifitas dan efisiensinya.[12],[13].

2. METODE PENELITIAN

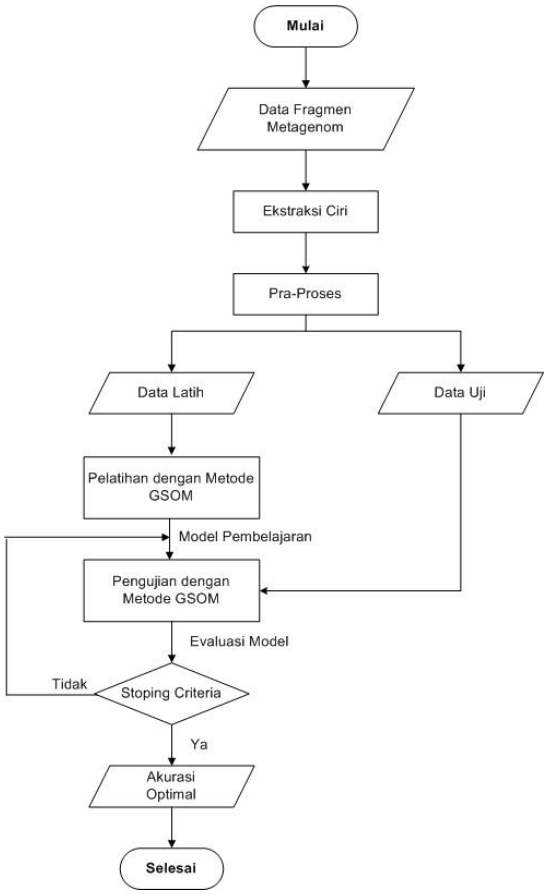
Penelitian ini metode pengumpulan data yang digunakan adalah Observasi. Observasi adalah teknik terjun langsung ke lapangan pada Dinas Lingkungan Hidup Kota Jayapura. Penelitian ini menggunakan data fragmen metagenom dari 300 mikrob dan kemudian dikelompokan berdasarkan tingkat taksonomi filum. Teknik 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. Lokasi penelitian bertempat di Dinas Lingkungan Hidup Kota Jayapura. Waktu penelitian dilakukan selama 6 bulan, dimulai dari bulan Agustus hingga Desember tahun 2019.[14]. Ilustrasi pemetaan fragmen metagenom, dapat dilihat pada gambar .1.

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.12, NO.2, Oktober 2020*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id



Gambar 1 Skema penelitian pengelompokan fragmen metagenom

Pengelompokan fragmen metagenom terdiri dari 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 pengelopokan untuk mengetahui efektifitas dan efisiensi pemetaan dengan *GSOM*. Tahap yang dilakukan untuk pengelompokan fragmen metagenome, [15]. dapat dilihat pada gambar 2.



Gambar. 2. Metodelogi Penelitian

Data yang digunakan adalah super kingdom bacteria dan merupakan hasil simulasi 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. dengan 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. Ekstraksi ciri adalah pembacaan *frekuensi oligonukleotida* (*trinukleotida dan tetranukleotida*) dengan k-mer dan juga ekstraksi menggunakan *spaced k-mer* yang memperhatikan kondisi *don't care* pada perhitungan *frekuensi oligonukleotida*. Ekstraksi ciri akan menampilkan pola kemunculan k pada suatu waktu dalam suatu sekuens. Pra proses data dilakukan Untuk mencegah adanya hasil implementasi yang bias, maka pengelompokan *fragmen metagenom* didahului dengan [16]. normalisasi data hasil ekstraksi fitur. Normalisasi data adalah salah satu bagian dari data transformasi, yaitu teknik mengubah data menjadi nilai yang lebih mudah untuk dipahami Tujuan lebih khusus dari normalisasi data adalah mendapatkan bobot yang sama dari semua atribut data dan tidak bervariasi atau hasil dari pembobotan tersebut tidak terdapat atribut yang lebih prior atau dianggap lebih utama dari pada yang lain. Untuk penelitian ini, normalisasi data yang digunakan adalah decimal scaling. Data *fragmen metagenom* akan diubah bobotnya menjadi data yang memiliki rentang [0, 1] menggunakan *transformasi linear* sederhana. Jumlah data adalah 200 mikrob untuk data latih dengan total jumlah fragmen yang digunakan adalah 200 000 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*.

3. HASIL DAN ANALISA

3.1. Basis Data *Fragmen Metagenom*

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 pada tabel 1 berikut :

Tabel 1 Jumlah data mikrob

Data	Jumlah
Data Latih	200
Data Uji	100

3.2. Ekstraksi Ciri dengan *Spaced k-mer*

Selain menggunakan *k-mer frequency* untuk ekstraksi ciri, digunakan *spaced k-mer*. Ekstraksi dengan *spaced k-mer* lebih ekonomis dilihat dari sisi penerimaan informasi (*information retrieval*) karena ekstraksi ini menggunakan kondisi *don't care* sehingga waktu yang dibutuhkan tidak terlalu lama tapi sudah mendapatkan informasi tentang komposisi dari fragmen metagenom dengan lebih terperinci. Data fragmen metagenom dihitung hampir sama dengan menggunakan *k-mer frequency*, tapi ekstraksi ini memperhatikan *don't care* yang mempunyai pola $111\ 1*11\ 1**11$, dengan $*$ adalah kondisi *don't care*. Sehingga dari perhitungan didapat dimensi fitur adalah sebanyak 192.

JSI : Jurnal Sistem Informasi *(E-Journal), VOL.12, NO.2, Oktober 2020*
ISSN Print : 2085-1588
ISSN Online :2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index
email: jsifasilkom@unsri.ac.id

Ukuran matriks komposisi dengan ekstraksi *spaced k-mer* pada data latih adalah $200\,000 \times 192$ dan $100\,000 \times 192$ untuk data uji.

3.3. Pembagian Data Latih dan Data Uji

Hasil praproses matriks komposisi dibagi menjadi data latih dan data uji dengan jumlah mikrob masing-masing 200 untuk data latih dan 100 untuk data uji. Beberapa mikrob yang digunakan sebagai data latih dan data uji masing-masing ditunjukkan pada tabel 2.

Tabel 2 Pembagian mikrob data latih dan data uji

Data Latih		Data Uji	
No	Mikrob	No	Mikrob
1	Acetobacterium woodi DSM 1030 chromosome	1	Acaryochloris marina MBIC11017 chromosome
2	Acidaminococcus fermentans DSM 20731 chromosome	2	Acetobacter pasterianus IFO 3283-01
3	Acidithiobacillus ferrivorans SS3 chromosome	3	Acholeplasma laidlawii PG-8A chromosome
4	Acidovorax sp.JS42 chromosome	4	Acidimicrobium ferroxidans DSM 10331 chromosome
5	Acinetobacter sp.ADP1 chromosome	5	Actinobacillus pleuropneumoniae serovar 3 str. JL03 chromosome
……..		……..	
200	Zymomonas mobilis subsp.pomaceae ATCC 29192 chromosome		Weissella korensis KACC 15510 chromosome

Data *fragmen metagenom*, masing-masing data latih dan data uji akan di bangkitkan sebanyak 200 000 untuk data latih dan 100 000 untuk data uji. Banyaknya pembangkitan data dari tiap kelompok filum dihitung secara otomatis ketika data disimulasi oleh MetaSim untuk setiap mikrob. Hasil perhitungan pembangkitan data latih dan data uji ditampilkan pada tabel 3 dan tabel 4.

Tabel 3 Pembangkitan Data Latih

No	Filum	Jumlah pembacaan
1	ACTINOBACTERIA	22 335
2	AQUIFICAE	2208
3	BACTEROIDETES	28 450
4	CHLOROBI	5102
5	CHLAMYDIAE	9330
6	VERRUCOMICROBIA	4679
7	CHLOROFLEXI	13 760
8	CYANOBACTERIA	16 376
9	DEINOCOCCUS-THERMUS	7606
10	ACIDOBACTERIA	10 781
11	FIRMICUTES	17 559
12	FUSOBACTERIA	3400
13	GEMMATIMONADETES	1484
14	NITROSPIRAE	2831
15	PLANCTOMYCETES	10 830
16	PROTEOBACTERIA	18 984
17	SPIROCHAETES	8702
18	SYNERGISTETES	1922
19	TENERICUTES	11 651
20	THERMOTOGAE	2010

Tabel 4 Pembangkitan Data Uji

No	Filum	Jumlah pembacaan
1	ACTINOBACTERIA	5452
2	AQUIFICAE	2144
3	BACTEROIDETES	5330
4	CHLOROBI	3950
5	CHLAMYDIAE	3764
6	VERRUCOMICROBIA	3716
7	CHLOROFLEXI	8652
8	CYANOBACTERIA	5685
9	DEINOCOCCUS-THERMUS	3873
10	ACIDOBACTERIA	10 199
11	FIRMICUTES	7648
12	FUSOBACTERIA	3281
13	GEMMATIMONADETES	1398
14	NITROSPIRAE	2751
15	PLANCTOMYCETES	9168
16	PROTEOBACTERIA	12 518
17	SPIROCHAETES	5829
18	SYNERGISTETES	1846

19	TENERICUTES	999
20	THERMOTOGAE	1797

3.4. Pengelompokan *Fragmen Metagenom* dengan *GSOM*

Frekuensi oligonukleotida adalah frekuensi kemunculan pasangan basa pada fragmen metagenom, dan pada penelitian ini muncul sebanyak trinukleotida, tetranukleotida, dan menggunakan frekuensi *spaced k-mer* yang memperhatikan kondisi *don't care*. Dalam penelitian ini, kemunculan frekuensi trinukleotida pada fragmen metagenom adalah sebanyak 64 fitur, frekuensi tetranukleotida sebanyak 256 fitur, dan frekuensi *spaced k-mer* sebanyak 192 fitur.

4. KESIMPULAN

Metagenom adalah penelitian tentang bagaimana menganalisis mikrob berskala besar dan memperbolehkan adanya pengkulturan secara langsung. Pengelompokan fragmen metagenom secara langsung bisa berakibat fatal karena bisa menyebabkan terjadinya interspesies chimeras atau kesalahan dalam perakitan fragmen metagenom. Studi metagenom merupakan langkah penting pada pengelompokan taksonomi. Metagenome merupakan mikroorganisme yang diambil secara langsung dari alam. Proses sequenc inggenom dari metagenome mengakibatkan bercampurnya berbagai organisme. Hal ini menyebabkan kesulitan pada proses perakitan DNA. Oleh karena itu, dibutuhkan proses pemilahan yang disebut binning. Pengelompokan fragmen metagenom pada lingkungan juga pada umumnya menggunakan supervised learning, sedangkan supervised learning merupakan pembelajaran yang menggunakan contoh dan bergantung pada ketersediaan data latih. Selain itu, pengelompokan juga menggunakan panjang fragmen yang panjang, yaitu ≥8 kbp dan berkomunitas kecil atau kurang dari 100 mikrob. Tujuan penelitian ini adalah untk menganalisis efektifitas dan efisiensi metode *Growing Self Organizing Map* dalam pengelompokan *mikrob* yang berskala besar dengan panjang fragmen yang pendek berdasarkan frekuensi oligonukleotida. *Frekuensi oligonukleotida* yang digunakan adalah trinukleotida, tetranukleotida, dan juga kombinasi frekuensi yang memperhatikan kondisi don't care, yaitu *spaced k-mer*. Untuk ekstraksi fitur, digunakan *k-mer frequency* dan *spaced*. Berdasarkan uji kombinasi parameter menggunakan *frekuensi oligonukleotida*, kombinasi terbaik antara *Learning Rate* dan *Neighborhood Size* untuk *frekuensi trinukleotida* adalah 0.1 untuk *Learning Rate*, 1 untuk Neighborhood Size dengan perhitungan quantization error adalah 0.531, 0.101 untuk topographic error, dan 16.84% untuk persentase error. Kombinasi terbaik tetranukleotida adalah 0.75 untuk Learning Rate dan 1 untuk *Neighborhood Size*, dengan memberikan nilai *error* 0.886 untuk *quantization error*, 0.09 untuk topographic error, dan 15.43% untuk persentase *error*. Untuk spaced k-mer, kombinasi terbaik adalah 0.5 untuk Learning Rate dan 1 untuk *Neighborhood Size* dengan *quantization error* adalah 0.665, 0.06 untuk topographic error dan 13.07% untuk
persentase *error*. Perhitungan kombinasi untuk ketiga *frekuensi oligonukleotida* menggunakan map size dan dan training lenght yang sama, yaitu [10 10] dan 10 *epochs k-mer frequency*. Dari hasil kombinasi parameter, *frekuensi spaced k-mer* menjadi frekuensi terbaik untuk pengelompokan fragmen metagenom dengan *metode Growing Self Organizing Map.*

REFERENSI

[1] Marlinda Vasty Overbeek. 2013. Pengelompokan Fragmen Metagenom Dengan Metode *Growing Self Organizing Map* [Tesis]. Sekolah Pascasarjana Institut Pertanian Bogor
[2] Arini. 2013. Metagenom Fragmen Binning Using Support Vector Machine (SVM) Method [skripsi]. Bogor-Indonesia : Institut Pertanian Bogor.Supriyanto. 2011. Sistem Konsultasi Online Agribisnis Cabai. [Tesis]. Bogor: Sekolah Pascasarjana IPB.
[3] Choi JH, Cho HG. 2002. Analysis of Common k-mers for Whole Genome Sequence Using SSB-Tree. Gemome Information. 13 : 30-41
[4] De Silva D, Alahakoon D, Dharmage S. 2007. Cluster Analysis using the GSOM : Patterns in Epidemiology. Proc. IEEE International Conference on ICIAF. 5(7):63 – 69. doi : 10.1109/ICIAFS.2007.4544781
[5] Ellyana, F. 2014. Klasifikasi Fragmen Metagenom Menggunakan Fitur Spaced N-Mers dan K-Nearest Neighbor [skripsi]. Bogor(ID): Institut Pertanian Bogor. Federhen S. 2012. The NCBI Taxonomy Database. Nucleic Acids Research. 40: 136-143. doi : 10.1093/nar/gkr1178T
[6] Harayama S, Kasai Y, Hara A. 2004. Microbial Communities in Oil-contaminated Seawater. Current Opinion in Biotechnology. 15:205-214Barendregt W, Bekker MM, Bouwhuis DG, Baauw E. 2014.Identifying Usability and Fun problem in a Computer Game During First Use and After Some Practice. Human Computer Interaction Studies Journal :830-846.
[7] 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.
[8] O'Malley M. 2012. Metagenomics. SpringerDitjenbun] Direktorat Jenderal Perkebunan. 2012. Memilih Bibit Kelapa Sawit yang Baik dan Benar. Jakarta(ID): Ditjenbun.
[9] Rahmawati V.2013. Comparison of Feature Extraction Methods Spaced K-Mers and K-mers in Fragmen Metagenom Classification using Naive Bayes Classifier [skripsi]. BogorIndonesia: Institut Pertanian Bogor.
[10] Richter DC, Ott F, Auch AF, Schmid R, Hudson DH. 2008. MetaSim-Sequencing Simulator for Genomics and Metagenomics. PLoS ONE. 3(10). doi:10.1371/journal.pone.0003373
[11] Sheaffer RL, Mendenhall W, Ott RL. 1990. Elementary Survey Sampling. 4th ed. Boston (US) : PWS – KENT Publishing Company
[12] Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W et al. 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science. 304 : 66 – 74. doi : 10.1126/science.1093857
[13] Vesanto J, Himberg J, Alhoniemi E, Parhankangas J. 2000.SOM Toolbox for Matlab 5. Helsinski University of Technology
[14] Wu H. 2008. PCA – based Linear Combinations of Oligonucleotide Frequencies for Metagenomic DNA Fragment Binning. IEEE Symposium on CIBCB. 8: 46-53
[15] Wu X, Lee W, Tseng C. 2006. ESTmapper : Efficiently Aligning Sequence DNAs to Genomes. 19th IEEE International Paralel and Distributed Processing Symposium.

[16] Zhu G, Zhu X. 2010. The Growing Self-Organizing Map for Clustering Algorithm in Programming Codes. IEEE International Conference on AICI. 3:178-182. doi : 10.1109/AICI.2010.276.