# Clustering Metagenome Fragments Using Growing Self Organizing Map

Marlinda Vasty Overbeek, Wisnu Ananta Kusuma, Agus Buono
*Departement of Computer Science*
*Faculty of Mathematics and Natural Science, Bogor Agricultural University*
Email : marlinda_vasty@yahoo.com, ananta@ipb.ac.id, pudesha@yahoo.co.id

***Abstract*— The microorganism samples taken directly from environment are not easy to assemble because they contains mixtures of microorganism. If sample complexity is very high and comes from highly diverse environment, the difficulty of assembling DNA sequences is increasing since the interspecies chimeras can happen. To avoid this problem, in this research, we proposed binning based on composition using unsupervised learning. We employed trinucleotide and tetranucleotide frequency as features and GSOM algorithm as clustering method. GSOM was implemented to map features into high dimension feature space. We tested our method using small microbial community dataset. The quality of cluster was evaluated based on the following parameters : topographic error, quantization error, and error percentage. The evaluation results show that the best cluster can be obtained using GSOM and tetranucleotide.**

## I. Introduction

Metagenomics is a study of analyzing high complexity of microbial community which allows culture – independent [1], [2]. As we know, only 1% of microorganism can be cultured by standard cultivation techniques. The rest should be taken directly from the environment, named as metagenome sample. This kind of sample contains mixtures of microorganisms. This characteristic makes assembling process becomes more difficult because it will yield more interspecies chimeras [5].

To solve the problem, we used binning process before or after assembling metagenome fragments. Binning is a techniques to classify or cluster organism based on taxonomy [5], [6].

There is two binning approach, the first approach is binning based on homology such as BLAST [7], [8] and MEGAN [9]. The second one is composition based approach. The composition approach applied unsupervised learning and supervised learning as a method and oligonucleotide as an input in the features spaces. There are many application developed based on this approach. Some applications that employed unsupervised learning are TETRA [10], Self Organizing Clustering [11], Self Organizing Map [12], and Growing Self Organizing Map [1], [13]. The ones that used supervised learning are PhyloPythia [14], Naive Bayessian Classification [15], and Phymm [16].

One of researches used GSOM combined with oligonucleotide to explore the genome signatures. Clear species-specific separation of sequence was obtained in the $\geq 8$ kbp fragments test. The fragments were derived from 30 species, which is separated into 3 dataset, 10 species per set [1].

In this research, we employed binning based on composition with unsupervised learning. We proposed 1 kbp DNA sequence derived from 18 species. We reads the fragments uniformly. The previous research [1] used long fragments (8 kbp). Using short length (1 kbp) gave a poor performance [5], [17]. In this research, we will overcome the limitation of using short fragment. The purpose of this research is to know the performance of GSOM in clustering the metagenome fragments with short fragment (1 kbp fragment lenght).

## II. MATERIAL AND METHODS

### *Growing Self Organizin Map (GSOM)*

GSOM consists of 3 main phase (Figure 1), which were initialization phase, growing phase and smoothing phase [18], [19].

### *Initialization phase*

In this phase, the algorithm initialize four starting nodes. Four starting nodes which were randomly selected from the input dataset. The initialization nodes were shown in Figure 2.

Next, the global parameter, Growth Threshold (GT) was calculated for the given dataset according to the user requirement. The GT value is defined as :

$$GT = -D \times \ln(SF)$$

Where D is dataset dimension, SF is Spread Factor. SF value was determined by user and SF took the values between zero and one.

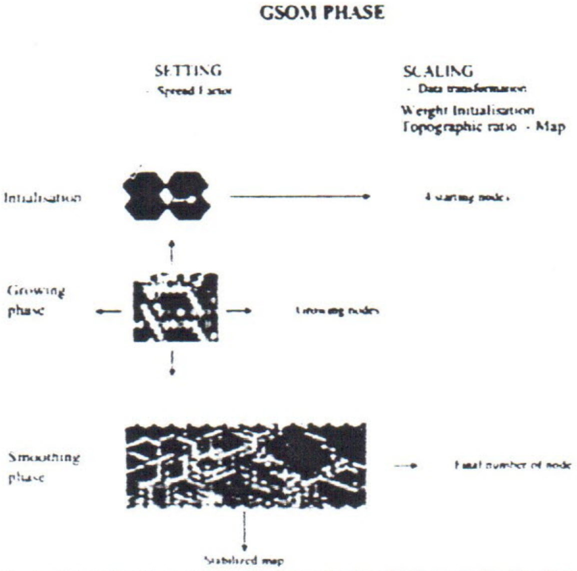


Fig 1. GSOM phase. Contains three phase which are initialisation phase (initialisation 4 starting nodes), growing phase (nodes is growing) as a important phase, and smoothing phase (final number of nodes)

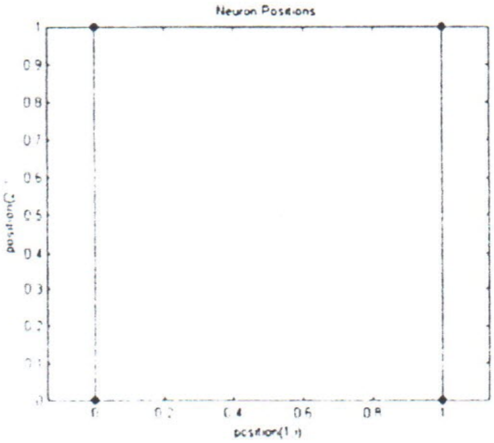


Fig 2. Initialization starting node. The values is between one and zero, randomly

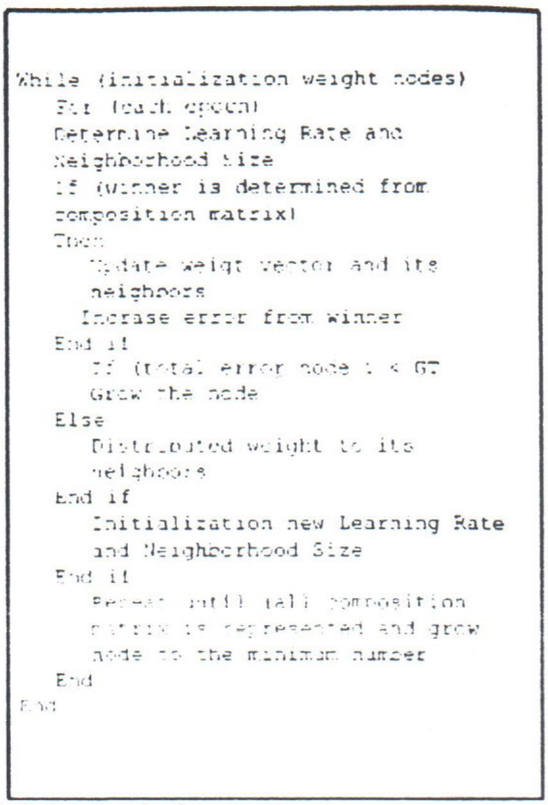
*Growing phase*

Growing phase is the most important phase in GSOM method, beacause in this phase the map would be set as dynamic to overcome the limitation of static map structure of SOM. Below is the pseudocode of growing phase in performing metagenome fragments clustering shown in Figure 3.

Error values is the distance between the input and winner node. The growth process depended on the growth threshold. When a node is not in the boundary of the network, it cannot grow new neighbors due its position.

*Smoothing phase*

In the final phase, the learning rate and neighborhood parameter would be decreased. This parameter changed in every iteration. When the minimum level is reached, the value would be close to zero.



```
While (initialization weight nodes)
   For (each epoch)
   Determine Learning Rate and
   Neighborhood size
   If (winner is determined from
   composition matrix)
   Then
      Update weigt vector and its
      neighbors
     Incrase error from winner
   End if
      If (total error node ) < GT
      Grow the node
   Else
      Distributed weight to its
      neighbors
   End if
      Initialization new Learning Rate
      and Neighborhood Size
   End if
      Repeat until (all composition
      matrix is represented and grow
      node to the minimum number
   End
End
```

Fig.3. Pseudo code of growing phase

Compared with the previous work by Chan *et al* [1], we used 1 kbp fragments lenght instead of 8 kbp. Using short fragments increased the complexity of clustered microorganism and made the project more difficult, since it always caused fragments to overlap and made them being mistaken in fragments assembling.

Because of that, in this research we transformed the features extraction result to 0 – 1 values. The data transformation was used to reduce the data variation and helped to increase the level of truth.

## III. RESULT

The proposed binning method was tested on simulated metagenome fragment generated by MetaSim [20]. The simulated dataset of microbes DNA sequence was randomly sampled from NCBI database [21]. In this research, we randomly took 18 microbes: 9 microbes for data training and 9 microbes for data testing with 1 kbp fragment length and then clustered into 3 different phylum: *Proteobacteria*, *Bacteroidetes*, and *Chlamydiae*; respectively. Each set of the genome sequence was separated into two orders of oligonucleotide frequencies (trinucleotide and tetranucleotide frequency). We set the Spread Factor

0.6 for trinucleotide frequency and 0.8 for tetranucleotide frequency.

The simulated dataset was extracted by *k-mer* frequencies method to get the specific oligonucleotide frequency and to put it in the composition matrix. After extracting, each of the dataset was scaled to obtain a dataset between zero and one. After scaling, we separated the dataset into data training and data testing (Table I and Table II). We used data training to obtain the trained model and used data testing to evaluate the GSOM algorithm. The flowchart of GSOM algorithm for clustering metagenome fragments are shown in Figure 4.

TABLE I
DATA TRAINING

| # | Microbes |
|---|---|
| 1 | Acidithiobacillus ferivorans SS3 chromosome |
| 2 | Bunchnera aphidicola (Cinara tujufilina) chromosome |
| 3 | Burkholderia glumae BGR1 chromosome 1 |
| 4 | Blattabacterium sp (Blaberus giganteus) chromosome |
| 5 | Flavobacterium branchiophilum FL-15 |
| 6 | Prevotella denticola F0289 chromosome |
| 7 | Chlamydia muridarum Nigg |
| 8 | Chlamydophila felis Fe/C-59 |
| 9 | Simkania negevensis Z chromosome chromosome gsn 131 |

TABLE II
DATA TESTING

| # | Microbes |
|---|---|
| 1 | Brevundimonas subvibriodes ATCC 15264 chromosome |
| 2 | Brucella canis ATCC chromosome I |
| 3 | Rhizobium etli CFN 42 |
| 4 | Bacteroides fragilis 638R |
| 5 | Prevotella melaninogenica ATCC 25845 chromosome I |
| 6 | Prevotella ruminicola 23 chromosome |
| 7 | Chlamydophila pneumoniae AR39 |
| 8 | Parachlamydia acanthaanmoebae UV-7 chromosome |
| 9 | Waddlia chondrophila WSU 86-1044 chromosome |

To evaluate the clustering performance, we used topology preservation (topographic error), mapping precision (quantization error) and error percentage [22], [23]. We also used time parameter to calculate the efficiency.

Quantization error is a common error measurement that measure the average distance between each data vector and its Best Matching Unit (BMU). Definition of BMU is a a randomly sampled vectors tha count the nearest distance between vectors [22]. Shortly, quantization error measure the mapping precission between input vector $\overline{xi}$ and nearest weight vector $m_{\overline{xi}}$.

$$qe = \frac{1}{N}\sum_{N}\|\overline{xi} - m_{\overline{xi}}\|$$

The expected map is obtained when the value of qe reached the minimum value.

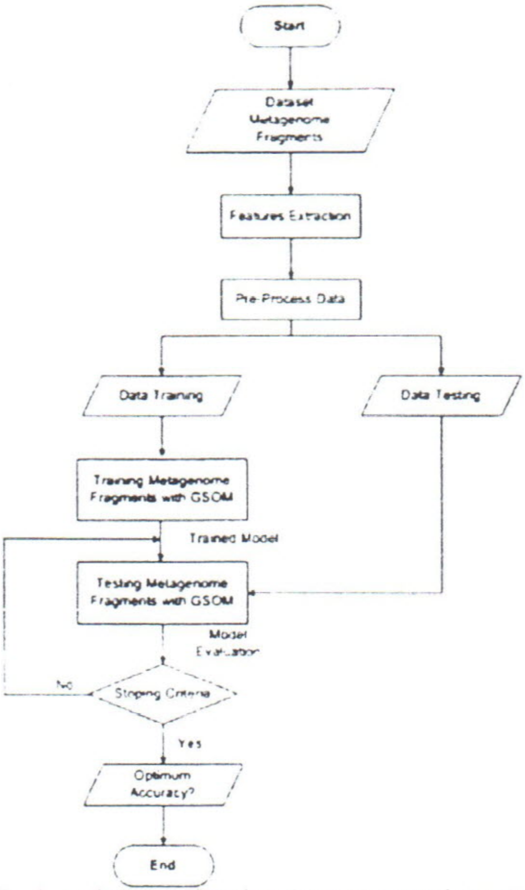


Fig. 4 Analysis procedure clustering metagenome fragments using GSOM algorithm

TABLE III
METAGENOME FRAGMENTS ANALYSIS VALUE

| # | Trinucleotide | Tetranucleotide |
|---|---|---|
| Topographic error | 0.067 | 0.066 |
| Quantization error | 1.304 | 0.742 |
| Error percentage | 18.74% | 18.48% |
| Time (sec) | 600 | 2880 |

To measure the topology preservation, we use topographic error. The topographic error considered the map structure and explained the correspondence between input data. This error measures the proportion of all data vectors for which first and second BMU are not adjacent vectors [22].

$$te = \frac{1}{N}\sum_{i=1}^{N} u(\overline{xi})$$

Where $u(\overline{xi}) = 1$ if $\overline{xi}$ data vector first and second BMUs are adjacent and 0, otherwise.

The error percentage used in this research was calculated based on the result of misclassification of the metagenome fragments data.

From the analysis result (Table III), we can see that both trinucleotide and tetranucleotide gave a good results. The error percentages were almost the same

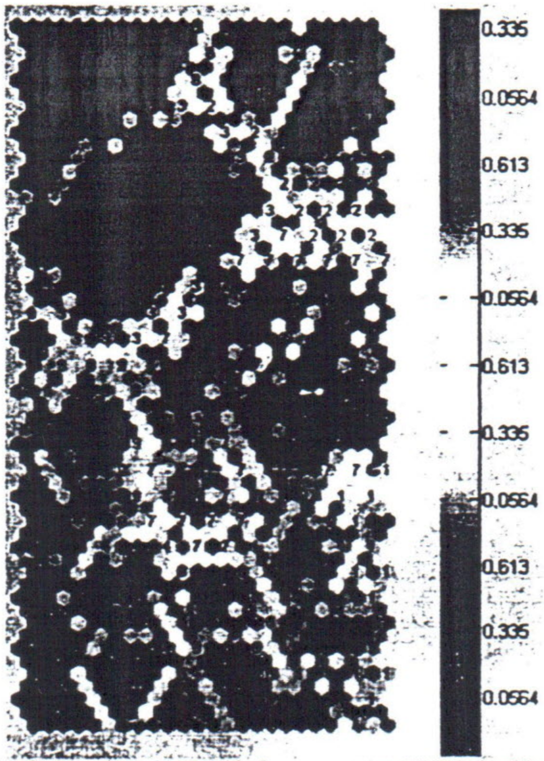


Fig. 5. Mapping trinucleotide frequency using GSOM algorithm. We use 0.6 SF value to control the neuron growth. Map stop growing in 28 × 13 map size

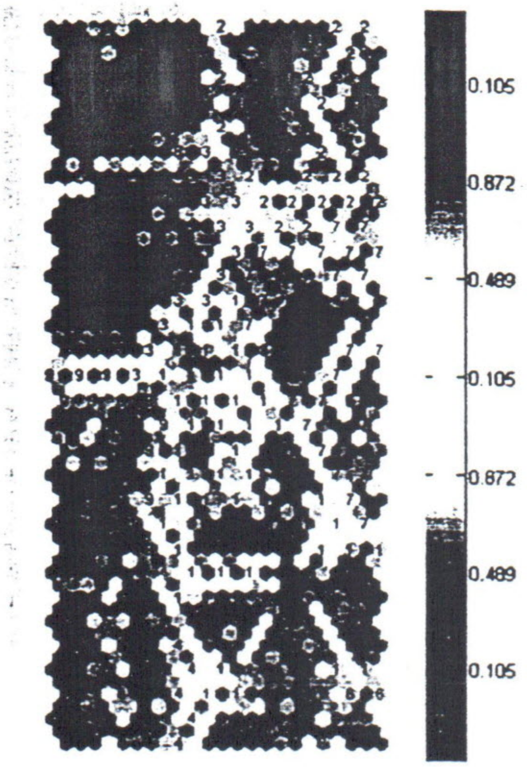


Fig. 6. Mapping tetranucleotide frequency using GSOM algorithm. We use 0.8 SF value to control the neuron growth. Map stop growing in 30 × 12 map size

with the values of 18. 74% and 18.48 % for using trinucleotide and tetranucleotide, respectively. This tendency was also shown by topology preservation. Both of oligonucleotide frequencies gave the topographic error of 0.067 for trinucleotide frequency and 0.066 for tetranucleotide frequency.

However both topograhic error result were enough to prove that every BMUs in the map grid was not adjacent vectors. It showed that both maps gave a quite good map preserve to clustering a metagenome fragments.

Moreover by analyzing their quantization error, we can conclude that tetranucleotide gave better cluster than trinucleotide. Tetranucleotide frequency gave a mapping precision result of 0.742, better than that of trinucleotide frequency which is 1.304. It means that clusters constructed using tetranucleotide frequency feature are more dense. We also showed the mapping results using trinucleotide frequency (Figure 5) and using tetranucleotide frequency (Figure 6).

## IV. CONCLUSION

Our method, combining GSOM and oligonucleotide can show a good performance in clustering metagenome fragments in phylum level with short fragment (1 kbp). The results showed that the performance of clustering using tetranucleotide was better than using trinucleotide. The error percentage result of using tetra-nucleotide is 18.48% and the quantization error was 0.472 less than using trinucleotide. Moreover, the topographic error is small, which means the neuron in map grid is not adjacent one and another. Based on these results we can conclude that GSOM algorithm is suitable for mapping the metagenome fragments with 1 kbp fragment length and has the opportunity to be implemented on large microbial community dataset.

## REFERENCES

[1] C. K. Chan, A. L. Hsu, S. Tang, and S. K. Halgamuge, "Using Growing Self-Organising Maps to Improve the Binning Process in Environmental Whole-Genome Shotgun Sequencing," *Journal of Biomedicine and Biotechnology*, vol. 2008, 2008.

[2] M. A. O. Malley, "Metagenomics," *2012*, 2012. [Online]. Available: http://www.maureenomalley.org/publications.html

[3] S. Harayama, Y. Kasai, and A. Hara, "Microbial communities in oil-contaminated seawater," *Current Opinion in Biotechnology*, vol. 15, pp. 205–214, 2004.

[4] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-tillson, C. Pfannkoch, Y. Rogers, and H. O. Smith, "Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science*, vol. 66, no. 2004, pp. 66–74, 2008.

[5] A. Meyerdierks and F. O. Glockner, "Metagenomic Analysis," *Handbook of Marine Genomics*, vol. 1, pp. 33–71, 2010.

[6] W. A. Kusuma, "Combined Approaches for Improving the Performance of de novo DNA Sequence Assembly and Metagenomic Classification of Short Fragments from Next Generation Sequencer," Tokyo Institute of Technology, 2012.

[7] H. Wu, "PCA - based Linear Combination of Oligonucleotide Frequencies for Metagenomic DNA

Fragment Binning." *IEEE Symposium on CIBCB*. vol. 8, pp. 46–53, 2008.

[8] H. Zheng and H. Wu, "A Novel LDA and PCA-based Hierarchical Scheme for Metagenomic Fragment Binning." *IEEE Symposium on CIBCB*, 2009.

[9] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data MEGAN analysis of metagenomic data." *Genome Research*, vol. 17, pp. 1–11, 2007.

[10] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences.," *BMC bioinformatics*, vol. 5, p. 163, Oct. 2004.

[11] K. Amano, H. Ichikawa, H. Nakamura, H. Numa, K. Fukami-Kobayashi, Y. Nagamura, and N. Onodera, "Self-organizing Clustering: Non-hierarchical Clustering for Large Scale DNA Sequence Data," *IPSJ Digital Courier*, vol. 3, no. 2, pp. 193–197, 2007.

[12] T. Abe, S. Kanaya, M. Kinouchi, Y. Ichiba, T. Kozuki, and T. Ikemura, "Informatics for Unveiling Hidden Genome Signatures," *Genome Research*, vol. 179, no. 4, pp. 693–702, 2003.

[13] A. L. Hsu and S. K. Halgamuge, "Enhancement of Topology Preservation and Hierarchical Dynamic Self- Organising Maps for Data Visualisation."

[14] A. C. McHardy, H. G. Martin, P. Hugenholtz, and I. Rigoutsos, "Accurate Phylogenetic Classification of Variable-Lenght DNA Fragment," *Nat Methods*, vol. 4, no. 1, pp. 63–72, 2010.

[15] G. Rosen, E. Garbarine, D. Caseiro, R. Polikar, and B. Sokhansanj, "Metagenome Fragment Classification using N-mer Frequency Profiles," *Advances in Bioinformatics*, 2008.

[16] A. Brady and S. L. Salzberg, "Phymm and PhymmBL : Metagenomic Phylogenetic Classification with Interpolated Markov Models," *Nat Methods*, vol. 6, no. 9, pp. 673–676, 2010.

[17] S. Prabhakara and R. Acharya, "Unsupervised Two-Way Clustering of Metagenomic Sequences," *Journal of Biomedicine and Biotechnology*, vol. 2012, 2012.

[18] G. Zhu, "The Growing Self-organizing Map for Clustering Algorithms in Programming Codes," *IEEE International Conference on AICI*, vol. 3, pp. 178–182, 2010.

[19] D. D. Silva, D. Alahakoon, and S. Dharmage, "Cluster Analysis using the GSOM : Patterns in Epidemiology," in *IEEE International Conference on ICIAF*, 2007, pp. 63–69.

[20] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson, "Metasim - Sequencing Simulator for Genomic Sequence," *PLoS ONE*, vol. 3, no. 10, pp. 1–34, 2008.

[21] S. Federhen, "The NCBI Taxonomy database," *Nucleic Acids Research*, vol. 40, no. December 2011, pp. 136–143, 2012.

[22] E. A. Uriarte and F. D. Martin, "Topology Preservation in SOM," *International Journal of Applied Mathematics and Computer Sciences*, pp. 19–22, 2005.

[23] J. Vesanto, M. Sulkava, and J. Hollm, "On the Decomposition of the Self-Organizing Map Distortion Measure Decomposition of the SOM distortion measure."