TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 13, No. 3, March 2015, pp. 537 ~ 545
DOI: 10.11591/telkomnika.v13i3.7115

Clustering Fragments Metagenome Using Self-Organizing Map

Yunita Fauzia Achmad*[1], Wisnu Ananta Kusuma[2], Heru Sukoco[3]
[1,2,3] Department of Computer Science, Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University, Bogor 16680, Indonesia
*Corresponding author, email: yunita.achmad12p@apps.ipb.ac.id[1], w.ananta.kusuma@gmail.com[2],
hsrkom@ipb.ac.id[3]

Abstract

Metagenome is a combination of several microorganisms collected from the environment. In metagenome analysis, it is required binning for grouping metagenome fragment yielded by sequencer. This research used the composition approach for conducting metagenome fragment binning. In this approach, binning could be implemented using unsupervised or supervised learning. We used Self-Organizing Map (SOM) *for conducting binning used on unsupervised learning. We compared two techniques of training in* SOM, *namely sequential training and batch training for finding the best techniques. The results showed that the batch training could obtain 3.8% error valued on the map of [10 15]. This error value is smaller than that of sequential training.*

Keywords: *batch training, binning, metagenome, self-organizing map, sequential*

1. Introduction

Bioinformatics is a discipline which originally arose for of introducing order into the massive data sets produced by the new technologies of molecular biology, such as large-scale DNA sequencing, the measurement of multiple gene expression, and the technology includes proteomics techniques. Bioinformatics integrate a number of related disciplines such as computer science, cybernetics, molecular evolution, genomics and proteomics, biometry and biostatistics, mathematical and computational biology, i.e., [10]. Bioinformatics introduce metagenome is involves sampling of microbial DNA from natural environments rather than relying on traditional, single species cultivation techniques [8].

Research on metagenome have done one of study [14] using about 1Gb of DNA sequences that have successfully sequenced from the Sargasso Sea samples showed the presence of microbial communities are far more diverse than previously thought.

Binning has two approaches, namely homology approach and composition approach. Homology approach is an approach that does alignment of DNA sequences by comparing fragments metagenome and database sequences. The results are concluded in each taxonomic level [2]. This causes the homology approach requires a lot of time in the process of grouping [8]. Example of methods homology is BLAST and Megan.

Composition approach has the advantage as a bypass. Composition approach is an approach that uses the base pairs of feature extraction results as fill for unsupervised learning and supervised learning. Method uses unsupervised learning is SOM and TETRA while, the supervised learning method is PhyloPythia

In this study the algorithm to be used as training and testing in metagenome fragment clustering is a self-organizing map (SOM). SOM successfully applied to high-dimensional data [6]. This study also uses k-mers on feature extraction method k-mer, frequency k-mer used is tetra-nucleotide and penta-nucleotide [2].

2. Research Method

2.1. National Center for Biotechnology Information (NCBI)

The first step takes the data collection. Data used were taken from national center of biotechnology information (NCBI). NCBI is a server that contains database of heath information

Received December 1, 2014; Revised January 18, 2015; Accepted February 2, 2015

and biotechnology. Database continuously update according to the latest discoveries concerning DNA, protein, the compounds active, and taxonomy [5].

2.2. MetaSim

Furthermore, the data that has been collected in the simulation sequencer to obtain fragments. Sequencer Software for used is MetaSim. MetaSim is a simulation software used to generate the data metagenome [11].

2.3. Bacteria Data

Data used is data bacteria are divided into two, namely the data training and data testing. The data training consists of 10 organisms with long reading 10 000 readings, whereas the data testing consist of 9 organisms with long reading 5 000 readings. Length of 5 kbp fragment used and the level of taxonomy genus [7].

2.4. Extraction Feature

The next step of feature extraction, method extraction feature are used k-mer is one of the characteristic extraction method that calculates the pattern of occurrence of k (the length of fragment metagenome) at a time in a sequence [3]. The attendance patterns k in sequences is calculated using the four main bases (A C T and G) raised to a series of base pairs (attendance patterns: 4^k with k ≥) [1]. Illustration of k-mer feature extraction can be seen in Figure 1.

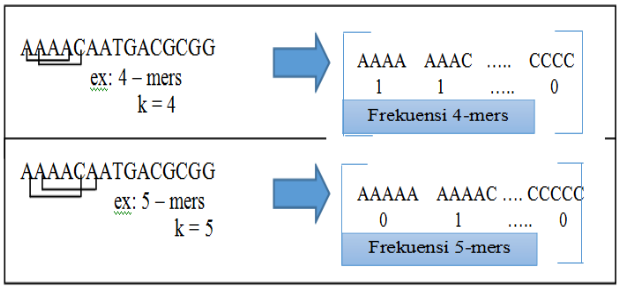


Figure 1. Extraction feature with k-mer

2.5. Normalization

Normalization is the process of scalling the value attribute of the data, so as to achieve the specified range. The purpose of normalization to remove redundancy of data, reduce complexity and simplify the data modification [4]. In this study, the normalization method used is the min max normalization method. Min-max normalization method using standardization of data by putting the data in the range of 0 to 1, the smallest value as 0 and the-gratest value as 1.

2.6. Self-Organizing Map (SOM)

SOM was first offered by Touve Kohonen in 1989 from Ireland. SOM is a neural network method based on competitive and unsupervised learning, because SOM does not have a target. SOM divide the data into several group defined clusters [6]. SOM has to training method, namely sequential training and batch training [12].

2.6.1. Sequential Training

Sequential training also referenced as on-line or stochastic, where the update is made for each sample of the training set [12].

The following sequential training algorithm:

1. Random weight initialization, determines maximum value for radius of neighborhood and learning value α
2. Stop condition if value false, otherwise go to step 3 to 9
3. Input vector x, proceed to step 4 to 6
4. Calculate distance Eculidean with formula:

$$D(j) = \sum_{i=1}^{n}(W_{ij-} X_i)2$$

5. Determine index j for D (j) most minor
6. Update weights for all j neighborhood of all inputs, with formula:

$$W_{ij}(Baru) = W_{ij}(Lama) + \alpha(t) * \left(X_i - W_{ij}(Lama)\right)$$

7. Update learning value α. value obtained with multiplying value learning rate function of the value of learning rate reduction, with formula:

$$\alpha(baru) = 0.6 * \alpha(lama)$$

8. Reducing radius of neighborhood (N) at a specific time
9. Stop condition is met if desired values close to zero are met. If the value α becomes very small, the weights readings will also be very small so that the training process can be stopped.

2.6.2. Batch Training

Batch training, where the update is performed after the presentation of all sample of the training set.

The following batch training algorithm:

1. Initialize weights randomly
2. Input vector x, proceed to step 3 to 7
3. Calculate Eculidean distance with formula (2)
4. Determining the best value (BMU) $u_{c(i)}$ at each vector x (assign each vector model most similar) called "**Voronoi Set**"
5. Updating of each vector and adjacency with formula:

$$m_i(i+1) = \frac{\sum_k u_{ic(k)}(t).x_k}{\sum_k u_{ic(k)}(t)}$$

6. Renew (decrease) radius of neighborhood
7. Stop condition is met, if value the specified epoch have been met.

Value of learning rate is that of the equations explicitly batch training refurbished and are not included at the parameters.

2.7. Evaluation

At this stage of the evaluation, there is the evaluation calculation performed is quantization error, topographic error and percentage error [13]. Quantization error is a measure commonly used at SOM method. This measurement is for measures the average distance between each vector and looking for the Best Matching Unit (BMU). Topographic error is a measure that uses the input samples for determines advanced mapping of the input space to the map grid. Whereas percentage error is used for calculate the mapping error at each grouping.

3. Result and Analysis

3.1. Sequential Training

Sequential experimental results of training are divided into two, namely based learning rate and of neighborhood. Experiments based on the learning rate can be seen in Figure 2 and 3, while the experiment is based on neighborhoods can be seen in Figure 4 and 5.

a) Based Learning Rate (LR)

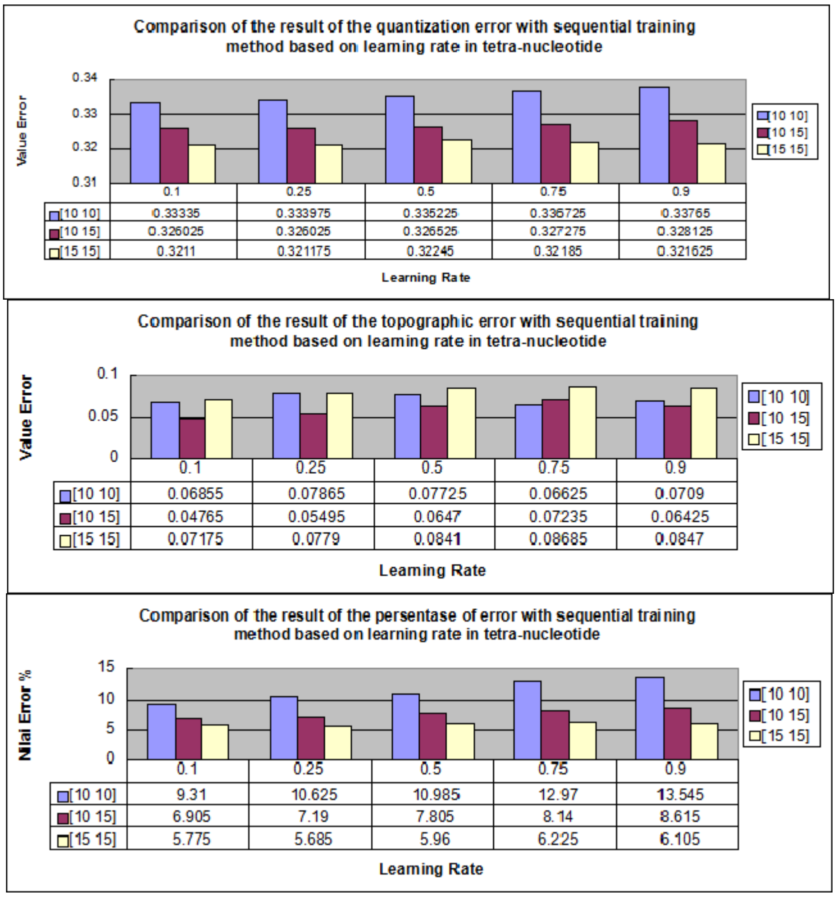
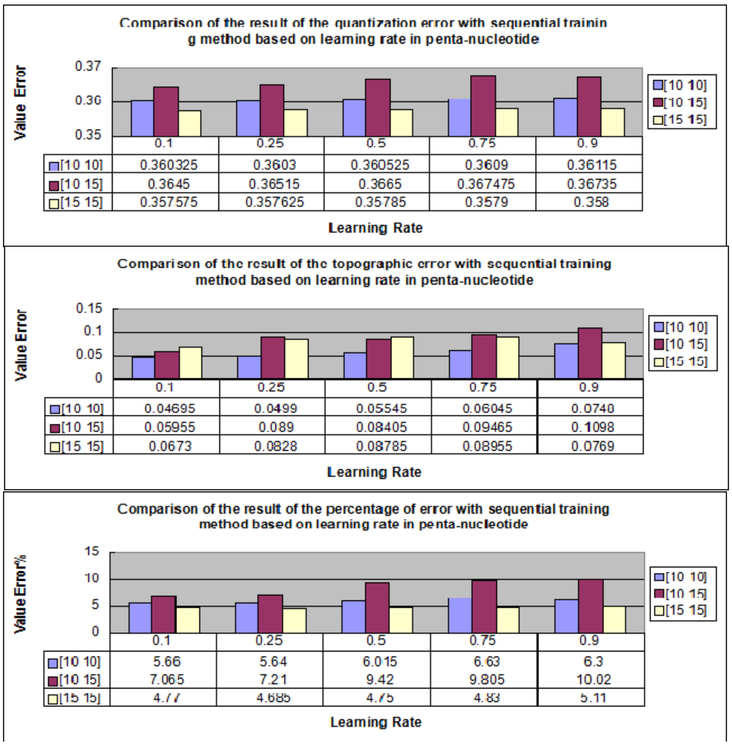


Figure 2. Comparison of the map size tetra-nucleotide based on learning rate with sequential training

Based on Figure 2 the results of calculations using the sequential training methods to tetra-nucleotide frequency based on learning rate with a comparison between the size of the map based on is used with calculation evaluation is used then, calculation quantization error which has the smallest error is contained in the map size [15 15] with a learning rate of 0.1 is 0.3211. Calculation topographic error which has the smallest error is contained in the map size [10 15] with a learning rate of 0.1 is 0.0476. Calculation percentage of error that have the smallest percentage contained in the map size [15 15] with the learning rate of 0.25 is 5.68%.

Based on Figure 3 the results of calculations using the sequential training methods to penta-nucleotide frequency based on learning rate with a comparison between the size of the map based on is used with calculation evaluation is used then, calculation quantization error which has the smallest error is contained in the map size [15 15] with a learning rate of 0.1 is 0.3575. Calculation topographic error which has the smallest error is contained in the map size [10 10] with a learning rate of 0.1 is 0.0469. Calculation percentage of error that have the smallest percentage contained in the map size [15 15] with the learning rate of 0.25 is 4.68%.



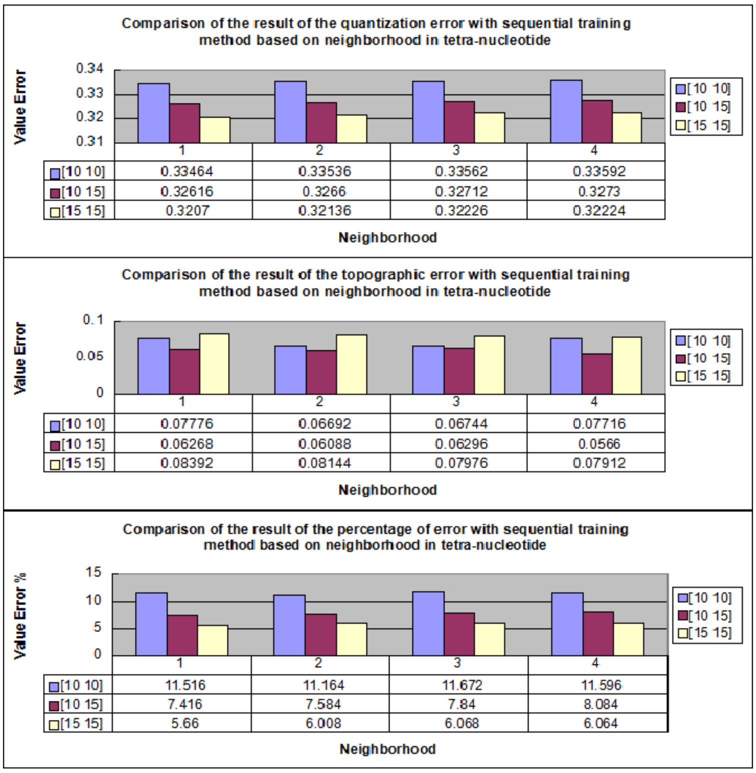
	0.1	0.25	0.5	0.75	0.9
□[10 10]	0.360325	0.3603	0.360525	0.3609	0.36115
■[10 15]	0.3645	0.36515	0.3665	0.367475	0.36735
□[15 15]	0.357575	0.357625	0.35785	0.3579	0.358

	0.1	0.25	0.5	0.75	0.9
□[10 10]	0.04695	0.0499	0.05545	0.06045	0.0740
■[10 15]	0.05955	0.089	0.08405	0.09465	0.1098
□[15 15]	0.0673	0.0828	0.08785	0.08955	0.0769

	0.1	0.25	0.5	0.75	0.9
□[10 10]	5.66	5.64	6.015	6.63	6.3
■[10 15]	7.065	7.21	9.42	9.805	10.02
□[15 15]	4.77	4.685	4.75	4.83	5.11

Figure 3. Comparison of the map size penta-nucleotide based on learning rate with sequential training

b) Based neighborhood



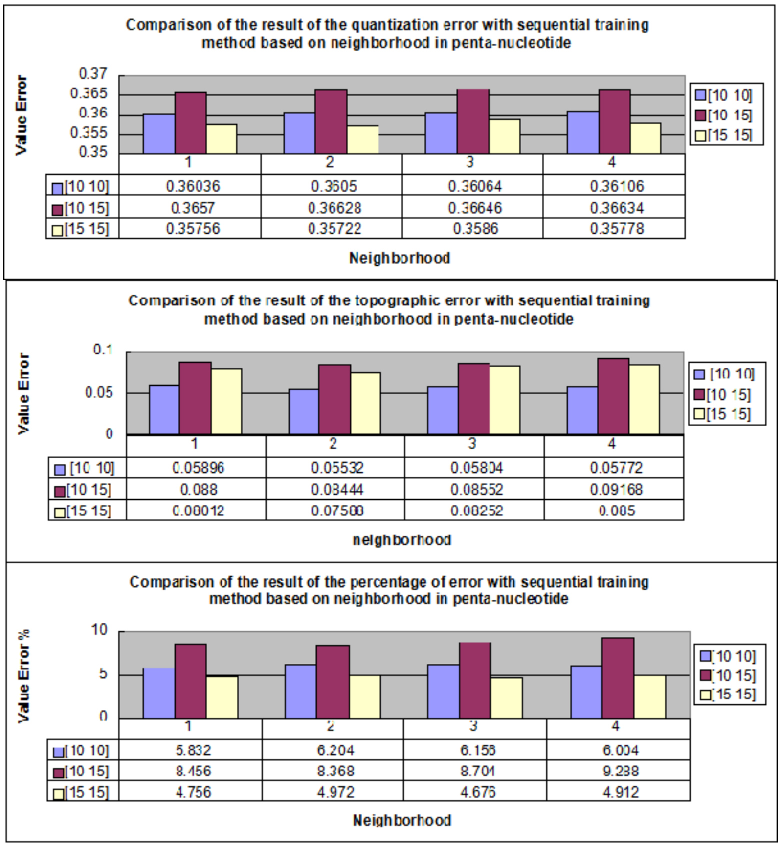
	1	2	3	4
□[10 10]	0.33464	0.33536	0.33562	0.33592
■[10 15]	0.32616	0.3266	0.32712	0.3273
□[15 15]	0.3207	0.32136	0.32226	0.32224

	1	2	3	4
□[10 10]	0.07776	0.06692	0.06744	0.07716
■[10 15]	0.06268	0.06088	0.06296	0.0566
□[15 15]	0.08392	0.08144	0.07976	0.07912

	1	2	3	4
□[10 10]	11.516	11.164	11.672	11.596
■[10 15]	7.416	7.584	7.84	8.084
□[15 15]	5.66	6.008	6.068	6.064

Figure 4. Comparison of the map size tetra-nucleotide based on neighborhood with sequential training

Based on Figure 4 the results of calculations using the sequential training methods to tetra-nucleotide frequency based on neighborhood with a comparison between the size of the map based on is used with calculation evaluation is used then, calculation quantization error which has the smallest error is contained in the map size [15 15] with a neighborhood of 1 is 0.307. Calculation topographic error which has the smallest error is contained in the map size [10 15] with a neighborhood of 4 is 0.0566. Calculation percentage of error that have the smallest percentage contained in the map size [15 15] with the neighborhood of 1 is 5.66%.


Comparison of the result of the quantization error with sequential training method based on neighborhood in penta-nucleotide

	1	2	3	4
[10 10]	0.36036	0.3605	0.36064	0.36106
[10 15]	0.3657	0.36628	0.36646	0.36634
[15 15]	0.35756	0.35722	0.3586	0.35778

Comparison of the result of the topographic error with sequential training method based on neighborhood in penta-nucleotide

	1	2	3	4
[10 10]	0.05896	0.05532	0.05804	0.05772
[10 15]	0.088	0.03444	0.08552	0.09168
[15 15]	0.00012	0.07500	0.00252	0.005

Comparison of the result of the percentage of error with sequential training method based on neighborhood in penta-nucleotide

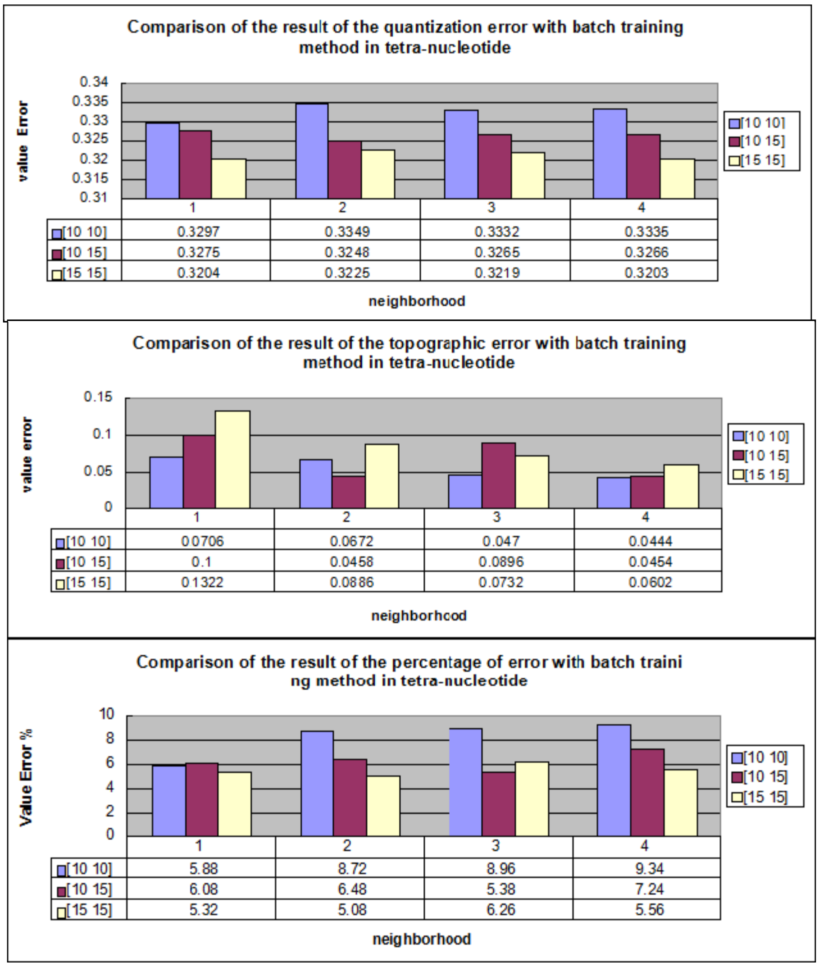
	1	2	3	4
[10 10]	5.832	6.204	6.155	6.004
[10 15]	8.456	8.368	8.701	9.238
[15 15]	4.756	4.972	4.673	4.912

Figure 5. Comparison of the map size penta-nucleotide based on neighborhood with sequential training

Based on Figure 5 the results of calculations using the sequential training methods to penta-nucleotide frequency based on neighborhood with a comparison between the size of the map based on is used with calculation evaluation is used then, calculation quantization error which has the smallest error is contained in the map size [15 15] with a neighborhood of 2 is 0.3572. Calculation topographic error which has the smallest error is contained in the map size [10 10] with a neighborhood of 2 is 0.0553. Calculation percentage of error that have the smallest percentage contained in the map size [15 15] with the neighborhood of 3 is 4.67%.

3.2. Batch Training

Different from the sequential training the batch training experiments tested only by neighborhood not using learning rate. The experimental batch training results can be seen in Figure 5 and 6.



	1	2	3	4
[10 10]	0.3297	0.3349	0.3332	0.3335
[10 15]	0.3275	0.3248	0.3265	0.3266
[15 15]	0.3204	0.3225	0.3219	0.3203

	1	2	3	4
[10 10]	0.0706	0.0672	0.047	0.0444
[10 15]	0.1	0.0458	0.0896	0.0454
[15 15]	0.1322	0.0886	0.0732	0.0602

	1	2	3	4
[10 10]	5.88	8.72	8.96	9.34
[10 15]	6.08	6.48	5.38	7.24
[15 15]	5.32	5.08	6.26	5.56

Figure 6. Comparison of the map size tetra-nucleotide with batch training method

Based on Figure 6 the results of calculations using the batch training methods to tetra-nucleotide frequency a comparison between the size of the map based on is used with calculation evaluation is used then, calculation quantization error which has the smallest error is contained in the map size [15 15] with a neighborhood of 4 is 0.3203. Calculation topographic error which has the smallest error is contained in the map size [15 15] with a neighborhood of 4 is 0.0444. Calculation percentage of error that have the smallest percentage contained in the map size [15 15] with the neighborhood of 2 is 5.08%.

Based on Figure 7 the results of calculations using the batch training methods to penta-nucleotide frequency a comparison between the size of the map based on is used with calculation evaluation is used then, calculation quantization error which has the smallest error is contained in the map size [15 15] with a neighborhood of 1 is 0.3563. Calculation topographic error which has the smallest error is contained in the map size [10 15] with a neighborhood of 2 is 0.0408. Calculation percentage of error that have the smallest percentage contained in the map size [10 15] with the neighborhood of 2 is 3.8%.

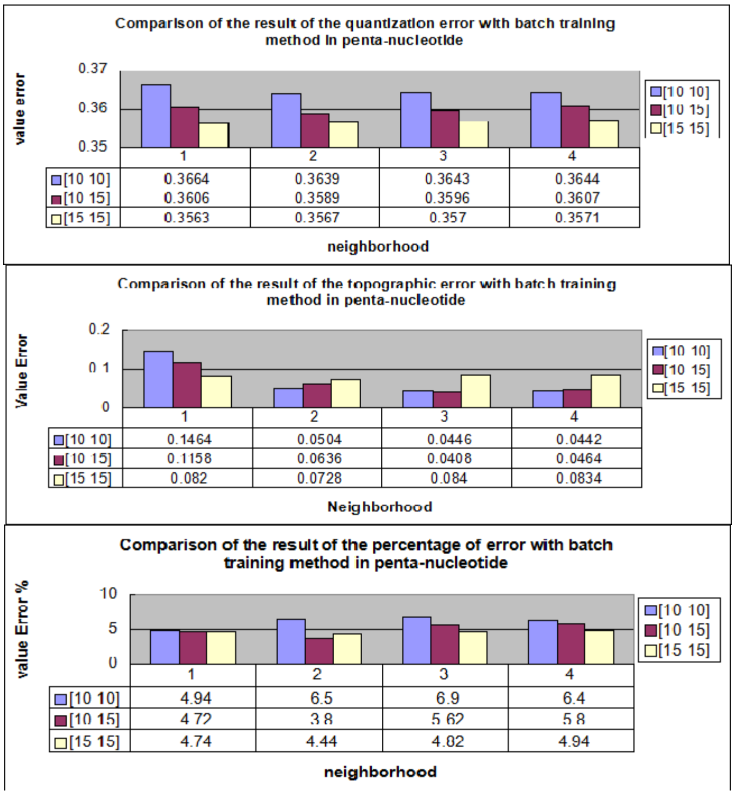


Figure 7. Comparison of the map size penta-nucleotide with batch training method

3.3. The Training Time Comparison between Sequential Training and Batch Training Method

Duration of training in this study is determined by the amount of data and the value of the epoch. Training is another factor influencing the length of training time. The Table 1 can in seen a comparison the training time between sequential training and batch training method.

Table 1. Comparison of the training time

Map size	Sequential training	Batch training
[10 10]	30 minutes	5 minutes
[10 15]	1 hour 10 minutes	25 minutes
[15 15]	2 hour	50 minutes

4. Conclusion

Conclusions that can be drawn in this study that batch training method produces a better grouping and faster than sequential training method. The smallest percentage error obtained penta-nucleotide frequencies contained in the map [10 15] and neighborhood 2 that is equal 3.64%.

References

[1] Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuku T, Ikemura T. Informatics for unveiling hidden genome signatures. *Genome Research*. 2003; 179(4):693-701.doi: 10.1101 / gr.634603.

[2] Chan CK, Hsu AL, Tang SL, Halgamuge SK. 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.

[3] Choi JH, Cho HG. Analysis of Common k-mers for Whole Genome Sequence Using SSB-Tree. *Gemome Information*. 2002; 13: 30-41.

[4] Han J, Kamber M, Pei J. *Data Mining: Concept and Techniques 3nd Edition.* Waltham USA: Morgan Kaufmann Publishers. 2012.

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

[6] Kohonen T. Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*.1982; 43(1): 59-69.

[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. Tokyo (JP); Tokyo Institute of Technology. 2012.

[8] Naser S, Breland A, Harris FC, Nicolescu M. A Fuzzy Classifier to Taxonomically Group DNA Fragments within a Metagenom. University of Nevada Reno. 2002.

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

[10] Polanski A, Kimmel M. *Bioinformatics*. Berlin (DE): Springer. 2007.

[11] Richter DC, OTT F, Auch AF, Schmid R, Huson DH. MetaSim: a sequencing simulator for genomics andmetagenomics. PLoS ONE, 3, e3373. 2009.

[12] Silva B. A study of a hybrid parallel SOM algorithm for large maps in data mining. Master Thesis FCT – UNI. 2008.

[13] Uriarte EA, Martin FD. Topology Preservation in SOM. *International Journal of Applied Mathematics and Computer Sciences.* 2005; 1(1): 19-22.

[14] Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Tillson HB, Pfannkoch C, Rogers YH, Smith HO. 2004. *Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science 304*. doi: 10.1126/Science.1093857.