ISSN 2085-4552

Classification of Metagenome Fragments With Agglomerative Hierarchical Clustering

Alex Kurniadi[1], Marlinda Vasty Overbeek[2]
Informatics Study Program, Universitas Multimedia Nusantara, Tangerang, Indonesia
[1]alex.kurniadi@student.umn.ac.id, [2]marlinda.vasty@umn.ac.id

Accepted 8 July 2021
Approved 12 August 2021

***Abstract*—Unlike genomics which study specifically culturable microorganisms, metagenomics is a field that studies microorganic samples retrieved directly from the environment. Such samples produce widely varying fragments when sequenced, many of which are still unidentified or unknown. Assembly of these fragments in the goals of identifying the species contained among them are thus prone to make said goals more difficult, so it becomes necessary for binning techniques to come in handy while trying to classify these mixed fragments onto certain levels in the phylogenetic tree. This research attempts to implement algorithms and methods such as k-mers to use for feature extraction, linear discriminant analysis (LDA) for dimensionality reduction, and agglomerative hierarchical clustering (AGNES) for taxonomic classification to the genus level. Experimentation is done across different objective measurements, including the length of the observed metagenome fragment that spans from 0.5 Kbp up to 10 Kbp for both the 3-mer and 4-mer contexts (k = 3 and k = 4). The averaged validity scores of the resulting data clusters generated from both the training and test sets, computed with the silhouette index metric, are 0.6945 and 0.0879 for the 3-mer context, along with 0.5219 and 0.1884 for the 4-mer context.**

***Index Terms*—AGNES; k-fold; k-mers; LDA; machine learning; metagenomics**

I. INTRODUCTION

In an effort to analyze the genetic material of nonculturable microorganisms (which cannot be cultured in laboratories), samples were taken directly from the environment. The sample taken may contain fragments of genetic material (genome) from a variety of different species. When the sequencing and assembly procedures are carried out on this mixture of fragments simultaneously, the mismatch between the genomes of one species with another will result in chimeric contigs that lead to the phenomenon of interspecies chimerae, so that the species diversity of the sample cannot be known [1] [2]. The term "contig" itself is taken from the English word "contiguous", and is defined as a strand of genomic fragments (DNA) of a species that are close together, representing a subset of DNA [3]. Chimeric contigs are then defined as a single contig strand composed of genomic fragments from two or more different species [4]. To minimize the chance of these occurrences, it is necessary to apply a binning technique so that each distinctive fragment of the compound can be separated as well as possible from one another.

This binning technique has two approaches, namely homology-based binning and composition-based binning. The homology approach was carried out by aligning the sample metagenome fragment sequences against the sequence data from the NCBI and concluding the results at the taxonomic level. Meanwhile, the compositional approach uses the result of feature extraction in the form of base pairs as input for the learning model [2].

There are two main approaches of learning for machine learning models, namely learning by example (supervised learning) and learning by observation (unsupervised learning). In the context of classification, supervised learning already has categorization information in its learning base, while unsupervised learning only has training data as a learning base. The method used in this study belongs in the realm of unsupervised learning.

In this study, metagenome fragments in the data from NCBI sources will be grouped using the *k*-mers method for feature extraction, linear discriminant analysis (LDA) method for data dimension reduction, and agglomerative hierarchical clustering (AGNES) algorithm for grouping. The *k*-mers method was chosen to be used in this study because this method works by calculating the number of occurrences of short strands (substring; polymer) along *k* letters in one genome strand, which will later highlight characteristic distinctions based on differences in the number of frequencies of each polymer between individual samples [5]. The LDA method was chosen because this method aims to try to explain a dependent variable as output (genus taxonomic level of the studied data) based on the values of the independent variables as input (genome strands belonging to the sample data). Meanwhile, the agglomerative hierarchical clustering method was chosen on the basis of its bottom-up workflow, because metagenomic fragment analysis starts from base pair units which then forms various long strands from each fragment based on the frequency of occurrence of certain base pair combinations.

II. LITERATURE STUDY

A. Metagenomics

Metagenomics is a branch of science that studies genetic material in samples taken directly from the environment [2]. Unlike genomics, where the analysis is carried out only on certain isolated (or previously cultured) organisms, metagenomic analysis is carried out directly on a group of microorganism communities without the need for first culturing efforts. This has brought interesting insights into the ecological systems of various habitats [6]. The emergence of this research field was triggered by the development of sequencing technology, with Roche's 454 pyrosequencing technique as an example [7].

B. Machine Learning

Machine learning is a field of science that studies computer algorithms that develop independently through experience [11]. Models built for machine learning work by taking input in the form of training data as learning material to analyze new, foreign data (test data). The learning approaches that are commonly applied in designing machine learning models consist of supervised learning and unsupervised learning. Supervised learning refers to a learning process in which the training data contains relevant input and output information, while in unsupervised learning, the training data only contains input information, so the model must draw its own conclusions (outputs) based on the learned input [8].

C. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a dimensionality reduction method that selects and extracts a set of the most discriminatory features from the data for multi-class classification [9]. This method is regarded as one of the most useful and popular methods with various applications, belonging in the realm of clustering algorithms [10]. The main difference between LDA and another dimensionality reduction method that is also frequently used, principal component analysis (PCA), lies in the main focus of each method. LDA and PCA aim to find the component with the highest variance in the data, but LDA also prioritizes the level of separability between data classes [11].

The LDA method can be summarized into seven steps as follows [11].

1) Standardize the initial *d*-dimensional data, where *d* represents the amount of features present in the data.

2) Compute the *d*-dimensional mean vectors m_i for each class in the data.

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x_m \tag{1}$$

$$m_i = \begin{bmatrix} \mu_{i(\text{feature } 1)} \\ \mu_{i(\text{feature } 2)} \\ \vdots \\ \mu_{i(\text{feature } n)} \end{bmatrix}^T, i \in \{1, 2, \dots, c\} \tag{2}$$

3) Compute the between-class scatter matrix S_B and the within-class scatter matrix S_W.

$$S_W = \sum_{i=1}^{c} S_i \tag{3}$$

$$S_i = \sum_{x \in D_i} (x - m_i)(x - m_i)^T \tag{4}$$

$$S_B = \sum_{i=1}^{c} n_i \, (m_i - m)(m_i - m)^T \tag{5}$$

4) Compute the eigenvectors along with the respective eigenvalues from the matrix $S_W^{-1}S_B$.

5) Sort the computed eigenvalues in descending order.

6) Take *k* eigenvectors with the highest eigenvalues to form the $d \times k$-dimensional transformation matrix *W*, where each eigenvector acts as one column.

7) Use the transformation matrix *W* to transform the initial *d*-dimensional data matrix *X* into a new feature matrix of dimension *k*.

Feature data that has been extracted is first normalized with min-max scaling method into the range [0, 1]. This normalization is applied to each feature with the formula below [11]:

$$x_{normal} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{6}$$

For each feature, the normalized value x_{normal} is obtained by calculating the weight of the individual values of *x* against the range of values in that feature, namely the difference between the largest and smallest values in said feature, denoted as x_{max} and x_{min}.

After computing the within-class and between-class distribution matrices, the eigenvectors and eigenvalues can be obtained by solving the matrix $S_W^{-1}S_B$. With the largest eigenvalues obtained, the corresponding eigenvectors are then combined into columns for the transformation matrix *W*. The transformation process is then carried out by multiplying the matrix *W* as shown in (7), where *X* is the initial data matrix and *X'* the new, reduced feature matrix:

$$X' = XW \tag{7}$$

D. Hierarchical Clustering

Hierarchical clustering is a method in statistics for performing clustering. This method works by first

measuring the level of inequality between two clusters of data before combining the two clusters into a new cluster (agglomerative) or splitting each cluster into two new clusters (divisive). The advantages of this method include ease of understanding and application and the absence of an obligation to first know the number of clusters desired [12]. The difference between agglomerative and divisive methods lies in the flow of work, where agglomerative hierarchical clustering forms large clusters of data by combining individual samples or a collection of small clusters (bottom-up), while divisive hierarchical clustering forms small clusters of data by splitting up clusters of data. larger [12]. The hierarchical clustering method that will be used in this study is the agglomerative method.

The level of inequality is measured based on the distance metric used between two data clusters, and the merging or splitting of the two clusters is carried out based on the linkage criterion used. There are a number of metrics and criteria commonly used in hierarchical clustering research, including the Euclidean distance metric (8) and the single-linkage linkage criterion (abbreviated as SLINK; (9)) [13]:

$$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2} \quad (8)$$

$$\min\{d(a,b) \mid a \in A, b \in B\} \quad (9)$$

E. Silhouette Index

The performance measurement of models using clustering algorithms is different from the measurements on classification algorithms in general. In evaluating clustering performance, the measured qualities are the closeness between samples in the same cluster, the separation of one sample from similar samples that are part of different clusters, and the separation between clusters [14]. To measure the validity of the results of such an algorithm, there are several different benchmark systems that can be used, and one such system used in this study is the silhouette index.

Silhouette index works by comparing the average distance of an individual sample *i* to all other samples in the same cluster, C_i, with the closest distance of the sample to all other clusters, C_k. The formula for the silhouette index is as follows.

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, if\ |C_i| > 1 \quad (10)$$

The variable *a(i)* represents the average distance of an individual sample *i* to every other sample *j* in the C_i group, *b(i)* represents the distance of a sample *i* to the nearest other cluster C_k, and *s(i)* states the silhouette index value for the sample *i*. In the case where a C_i cluster only has *i* as its sole sample, then *s(i)* will be zero. The formulas for two sample distances *a(i)* and *b(i)* can be seen in (11) and (12).

$$a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \neq j} d(i,j) \quad (11)$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \quad (12)$$

III. METHODOLOGY

A. Preparing the Genome Data

Genomic data for each observed microorganism species is obtained from the NCBI website and processed using the MetaSim tool to generate metagenome data in FASTA format files. The data set used in this study consists of eighty species from ten genera. Research will be carried out on different fragment lengths, ranging from 0.5 Kbp (kilo base pair; 10^3 bp), 1 Kbp, 5 Kbp, up to 10 Kbp. The total weighting of all species in the final fragment is applied to 10,000.

B. Feature Extraction

The resultant FASTA data is first preprocessed before the data can be used by the machine learning model. The preprocessing procedure here begins with feature extraction using the *k*-mers method, where each strand of fragments is observed to note how often each combination of base pairs A, T, G, and C, as a key component determining the characteristics of microorganisms, appears on the thread. The length of base pair combinations that will be observed here are 3-mers such as AAA, AAC, AAG, up to TTC, TTG, TTT (4^3 = 64 combinations), and 4-mers such as AAAA, AAAC, up to TTTG, TTTT (4^4 = 256 combinations), as shown in Fig. 1. This distinction between combinations is important, because DNA is the genetic material in almost all living things, including microorganisms and humans, where different combinations will yield different physiological characteristics [16].

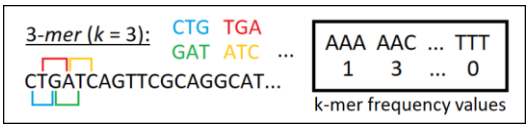
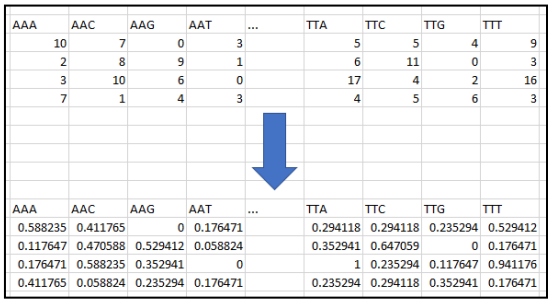


Fig. 1. Illustration of 3-mer frequency counting

C. Data Normalization

The extracted features data is then normalized using the min-max scaling method. In min-max normalization, all feature values are rescaled into a new range of values while maintaining the weight of each value. In this study, the range of values used is [0, 1], as can be seen in Fig. 2.



AAA	AAC	AAG	AAT	...	TTA	TTC	TTG	TTT
10	7	0	3		5	5	4	9
2	8	9	1		6	11	0	3
3	10	6	0		17	4	2	16
7	1	4	3		4	5	6	3

AAA	AAC	AAG	AAT	...	TTA	TTC	TTG	TTT
0.588235	0.411765	0	0.176471		0.294118	0.294118	0.235294	0.529412
0.117647	0.470588	0.529412	0.058824		0.352941	0.647059	0	0.176471
0.176471	0.588235	0.352941	0		1	0.235294	0.117647	0.941176
0.411765	0.058824	0.235294	0.176471		0.235294	0.294118	0.352941	0.176471

Figure 2. Illustration of min-max scaling

D. *Dimensionality Reduction*

After normalization is complete, dimensionality reduction is carried out on the data set before it is ready to be used in machine learning models. Dimensionality reduction aims to eliminate redundant characteristics by transforming the characteristics from the data matrix with larger dimensions into smaller dimensions [16]. The reduction algorithm used in this study is Linear Discriminant Analysis (LDA). In this study, the LDA algorithm is run by following the steps described in [11], except for data standardization. This is because the data to be reduced has already been normalized, so the LDA procedure here can be continued without having to standardize the data again.

E. *Clustering*

Analysis is then carried out on the training data set first to see and determine the most optimal conditions. The optimal conditions are to be used as the basis for the analysis of the test data set in order to group the data set into genera. The algorithm used for this testing phase is agglomerative hierarchical clustering.

To measure the validity of each data cluster resulting from the model grouping, there are several benchmark systems that can be used, including the silhouette index used in this study. Silhouette index of each sample is calculated based on the average distance of the sample to all other samples in the same cluster and its distance to all other clusters by taking the cluster closest to the sample.

IV. DISCUSSION

A. *Metagenome Data*

For the sample data, metagenome DNA fragment strands processed by MetaSim were prepared from 80 species of microorganisms belonging to 10 different genera. The total weighting of all species for the metagenome fragments is 10,000.

B. *Feature Extraction*

The preprocessing stage begins by performing feature extraction in the form of base pair combinations from the metagenome fragment data. After the feature extraction is successful, a new set of numerical data is obtained which can later be used by the machine learning model. The number of data rows in the sample table corresponds to the results of the previous MetaSim processing, while the number of data columns adjusts to the observed *k*-mer value, which is 4^k columns.

C. *Normalization*

The sample data is first normalized using the min-max scaler from the preprocessing module scikit-learn [14]. The values in each column are calculated and converted to the relative weights of each value against all other values in the same column. The range of values used as a reference for normalization is [0, 1], which means that the lowest value in the column becomes 0, while the highest value in the same column becomes 1.

D. *Dimensionality Reduction*

After normalizing the data, dimensionality reduction is performed to obtain a new, smaller matrix containing the projection of the sample data set matrix while maintaining the integrity of information between data classes. This stage begins by calculating the mean vector for each existing class, resulting in a collection of *d*-dimensional vectors where *d* is the number of features.

With the mean vectors, the within-class and between-class distribution matrices S_W and S_B are computed. These two distribution matrices are then used to obtain a series of eigenvectors and eigenvalues as indicators of the integrity of information between classes in the data. Eigenvalues that are not close to zero are taken to create a transformation matrix *W*, which will later be used to transform the $n \times d$-dimensional sample data set into an $n \times k$-dimensional matrix, where k does not exceed the number of features (d) nor the number of classes subtracted by one (c - 1).

E. *Data Splitting*

The sample data set is divided into training data and test data, with the proportion of test data being 20% of the initial set.

F. *Clustering*

The reduced data set is then grouped into clusters by the model, using agglomerative hierarchical clustering (AGNES) algorithm based on information from the training data. Because this model aims to group the test data into genera that have been presented previously, the number of clusters adjusts to the number of existing genera, which is 10 clusters. The distance metric used is Euclidean, and the relationship criterion used is single-linkage, where the shortest distance between two data points from two different clusters is taken as the distance between the two clusters.

G. *Evaluation*

Evaluation of the machine learning model is carried out using the silhouette index assessment method, in which each individual sample processed by the model

is assessed based on the sample's proximity to its own cluster with the closest other cluster [14]. The final value of the silhouette index of the learning model is obtained by averaging values of the silhouette index for the entire data set. The range of values in this scoring method is [-1, 1], with -1 as the worst value and 1 as the best value. A value that tends to be negative indicates that the samples are grouped into the wrong clusters, whereas a value close to 0 (zero) indicates that the clusters tend to overlap with each other [14].

The process of evaluating the model's performance in grouping sample data spans across a series of observational contexts on both training (80% sample size) and test (20% sample size) data sets. These contexts consist of the variations in total length of metagenome fragments (0.5 Kbp to 10 Kbp) and the polymer length (3-mer and 4-mer), as shown in Tables I and II below.

TABLE I. TRAINING DATA SILHOUETTE INDEXES

Total fragment length	***k*-mer**	
	***k* = 3**	***k* = 4**
0.5 Kbp	0.6022	0.0162
1 Kbp	0.6285	0.1318
5 Kbp	0.7443	0.0403
10 Kbp	0.8029	0.1634
Average	**0.6945**	**0.0879**

TABLE II. TEST DATA SILHOUETTE INDEXES

Total fragment length	***k*-mer**	
	***k* = 3**	***k* = 4**
0.5 Kbp	0.4397	0.1083
1 Kbp	0.4678	0.0572
5 Kbp	0.5444	0.2160
10 Kbp	0.6356	0.3724
Average	**0.5219**	**0.1885**

From the four average values above, it is then known that the data grouping quality of the model in the 3-mer context is better than that in the 4-mer context (0.6945 > 0.0879; 0.5219 > 0.1885). In the 3-mer context, the silhouette index value being closer to 1 indicates that the data has been grouped fairly well and is quite separate between each data cluster. Meanwhile, in the 4-mer context, the silhouette index being closer to 0 indicates that after data is grouped by the model, there are many data clusters that overlap with each other, so that the separation between data clusters becomes unclear.

V. CONCLUSION

In this study, the metagenomic fragment data set was first preprocessed with LDA as the dimensionality reduction method and *k*-mers as the feature extraction method. The preprocessed data set was then grouped with agglomerative hierarchical clustering algorithm and the resulting clusters were evaluated with the silhouette index metric.

The silhouette index values, a measure of the validity of data grouping by the model, ranged from 0.6022 ~ 0.8029 for the training set and 0.4397 ~ 0.6356 for the test set in the 3-mer context. In the 4-mer context, the silhouette index values ranged from 0.0162 ~ 0.1634 for the training set and 0.0572 ~ 0.3724 for the test set. This means that in the 3-mer context, the resulting data have been clustered quite well and the clusters are quite separate between each other. However, in the 4-mer context, there are many clusters that overlap with each other, causing the silhouette index value to come close to zero.

REFERENCES

[1] Overbeek, M. V., Kusuma, W. A. and Buono, A. (2013) 'Clustering metagenome fragments using growing self organizing map', *2013 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2013*, pp. 285–289. doi: 10.1109/ICACSIS.2013.6761590.

[2] Simangunsong, V. F. R. (2015) 'Klasifikasi Fragmen Metagenom Menggunakan Principal Component Analysis Dan K-Nearest Neighbor', pp. 1–34.

[3] Gregory, S. G. (2005) 'Contig Assembly', *Encyclopedia of Life Sciences*, pp. 1–4. doi: 10.1038/npg.els.0005365.

[4] Scholz, M. *et al.* (2020) 'Large scale genome reconstructions illuminate Wolbachia evolution', *Nature Communications*, 11(1). doi: 10.1038/s41467-020-19016-0.

[5] Rosen, G. *et al.* (2008) 'Metagenome Fragment Classification Using -Mer Frequency Profiles', *Advances in Bioinformatics*, 2008, pp. 1–12. doi: 10.1155/2008/205969.

[6] Richter, D. C. *et al.* (2008) 'MetaSim - A sequencing simulator for genomics and metagenomics', *PLoS ONE*, 3(10). doi: 10.1371/journal.pone.0003373.

[7] Margulies, M. *et al.* (2005) 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature*, 437(7057), pp. 376–380. doi: 10.1038/nature03959.

[8] Bishop, C. M. (2006) Pattern Recognition and Machine Learning, The Ecstatic and the Archaic: An Analytical Psychological Inquiry. Springer.

[9] Yan, C. *et al.* (2020) 'Self-Weighted Robust LDA for Multiclass Classification with Edge Classes'. Available at: http://arxiv.org/abs/2009.12362.

[10] Liu, J. *et al.* (2020) 'Capped norm linear discriminant analysis and its applications', (1), pp. 1–11. Available at: http://arxiv.org/abs/2011.02147.

[11] Raschka, S. and Mirjalili, V. (2019) *Python Machine Learning*. Third. Birmingham: Packt Publishing Ltd.

[12] Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. 99th edn. Hoboken: John Wiley & Sons, Inc.

[13] SAS Institute Inc. (2009) *SAS/STAT(R) 9.2 User's Guide, Second Edition*. 2nd edn. Edited by A. Jones and E. Huddleston. Cary, NC: SAS Institute Inc.

[14] Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: http://scikit-learn.sourceforge.net. (Accessed: 3 June 2021).

[15] *What is DNA?: MedlinePlus Genetics* (2021). Available at: https://medlineplus.gov/genetics/understanding/basics/dna/ (Accessed: 28 June 2021).

[16] Tharwat, A. *et al.* (2017) 'Linear discriminant analysis: A detailed tutorial', *AI Communications*, 30(2), pp. 169–190. doi: 10.3233/AIC-170729.