MicroOhm JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING

MicroOhm Journal of Electrical and Computer Engineering
Vol. 1, No. 1; July 2025
E-ISSN: XXXX-XXXX

Application of K-Means Clustering Statistical Model in DNA Base Frequency Distribution Analysis

Luthfie Budie[1], Syaputra Ervian[2]

[1]Universitas Islam Negeri Sumatera Utara; LuthfieBudieAndika@gmail.com
[2]Universitas Sinar Husni; Syaputraervian@gmail.com

ABSTRACT

For effective analyses, reliable statistical models are required the rapid advances in genomics have generated increasingly complex and high-dimensional datasets, requiring the use of reliable statistical models for effective analysis in genomics have generated complex and large data. Statistical models play a critical role in supporting various genomic analyses, including evolutionary studies, Genome-Wide Association Studies (GWAS), transcriptomics, and gene regulatory network reconstruction. Researchers can identify genetic variants associated with diseases using through these models, researchers are able to identify disease-associated genetic variants, analyze gene expression patterns, and predict phenotypes from genetic data. To address challenges such as noise, high dimensionality, and multiple testing in genomic data, machine learning techniques particularly classification, prediction, and clustering are increasingly used. One such technique is the K-Means clustering algorithm. K-Means clusters genomic data based on statistical similarities in features derived from DNA sequences. Such clustering methods enhance our understanding of underlying genetic mechanisms and support their application in clinical genomics. This study specifically investigates the application of the K-Means clustering model for analyzing DNA base frequency distributions as a case of statistical modeling in genomics.

Keywords : *Bioinformatics, Statistics, Genomics;*

Corresponding Author:
Author Name: Luthfie Budie
Affiliation: Universitas Islam Negeri Sumatera Utara
Email: LuthfieBudieAndika@gmail.com

1. INTRODUCTION

Next-generation sequencing (NGS) technologies have dramatically transformed genomic research, enabling the rapid collection of vast and complex biological datasets. While these advances open new frontiers in genetics, biomedicine, and evolutionary studies, they also introduce analytical challenges due to high dimensionality and inherent data complexity (Saudale, 2020). To uncover meaningful biological insights from such data e.g., gene expression profiles and DNA sequences robust analytical approaches are needed (Siswantining et al., 2022), where statistical modeling plays a central role in managing data variability and structure (Adiyana et al., 2022).

Statistical models are instrumental in a wide range of genomic applications, including identifying disease-associated genes, analyzing population-level variation, and predicting phenotypes from genotypes (Jajang et al., 2022). Traditional approaches such as logistic regression and mixed linear models are commonly applied to study genotype-phenotype associations (B & H, 2020). Additionally, unsupervised learning methods such as clustering and principal component analysis (PCA) are widely used to uncover latent structures in gene expression data (Muningsih et al., 2020).

Although prior research has successfully utilized statistical models such as support vector machines (SVM) and principal component analysis (PCA) to classify and explore genetic data (Putri et al., 2020; Chandra, 2022), these techniques typically rely on labeled data. In practice, genomic datasets often lack predefined categories or annotations, highlighting the need for unsupervised learning approaches that can uncover intrinsic data structures without supervision.

Unsupervised learning methods such as K-Means clustering are particularly suited for analyzing unlabeled genomic data. DNA sequences can be numerically represented using the relative frequencies of nitrogenous bases (A, T, C, and G), enabling clustering based on statistical similarity. This approach facilitates the identification of biologically relevant groups and patterns in high-throughput datasets (Suryanto, 2024). However, it is important to acknowledge the limitations of K-Means, such as sensitivity to initialization and assumptions of spherical cluster shapes.

Despite methodological advances, issues such as high dimensionality, missing data, and sampling bias persist in genomic data analysis (Alhadi, 2020). This study aims to apply the K-Means clustering algorithm to analyze the frequency distribution of DNA bases, with the objective of uncovering inherent patterns and subgroup structures within genomic sequences. The findings are expected to contribute to a deeper understanding of genomic organization and support potential applications in disease gene discovery and personalized medicine.

2. LITERATURE REVIEW

Numerous studies emphasize the critical role of statistical models in analyzing complex genomic data, which often involves thousands to millions of variables and exhibits high dimensionality. Statistical techniques such as probabilistic modeling, regression analysis, and principal component analysis (PCA) are widely adopted to extract latent biological patterns from such data (Sinaga et al., 2020; Saudale, 2020).

K-Means clustering, a widely used unsupervised learning algorithm, enables the grouping of DNA sequences or gene expression profiles based on similarities in statistical distribution, without requiring labeled data. It is particularly effective for identifying latent structures within genomic datasets, such as genes sharing similar functions or expression patterns potentially linked to disease phenotypes (Muningsih et al., 2020).

Statistical modeling plays a vital role in genome-wide association studies (GWAS), where it is used to identify associations between genetic variants and phenotypic traits. Techniques such as linear mixed models and Bayesian frameworks are commonly employed to manage population structure and reduce noise, thereby improving the accuracy and robustness of GWAS findings (Fadli et al., 2020).

Statistical modeling is also essential in gene expression analysis, where data are typically obtained from microarray experiments or RNA sequencing (Agustina et al., 2020). In parallel, K-Means clustering can be applied to DNA sequence data by converting each sequence into a numerical representation based on base frequency distributions. This approach is well suited for exploratory analysis of unlabeled genomic data, enabling the identification of latent biological groupings (Suryanto, 2024). The following section outlines key statistical concepts and their applications in genomic data analysis:

2.1. Regression Modelling and Classification of Genetic Data

Regression modelling is used to analyse the relationship between genetic data and phenotypic outcomes. For example, logistic regression is used to predict the probability of a condition, such as a disease, based on the genetic variables present. This technique calculates the probability of the condition using the formula:

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}}$$

Explanation:
f(x) = function
e = Euler number
-x = means the negative of the value of x.

This technique is widely used in genomic research to find relationships between genes and diseases, as well as to make predictions based on individual or group genetic data (Ayuningtyas et al., 2020).

2.2. Principal Component Analysis (PCA)

Genomic data often has very large dimensions, so dimension reduction methods are required. PCA identifies data structure and reduces data complexity without losing important information. In gene expression analysis, it is particularly beneficial for finding patterns that may not be visible in very large datasets (Chandra, 2022; Suryanto, 2024). The following is a flowchart of the PCA principal component analysis process in genomic data processing.

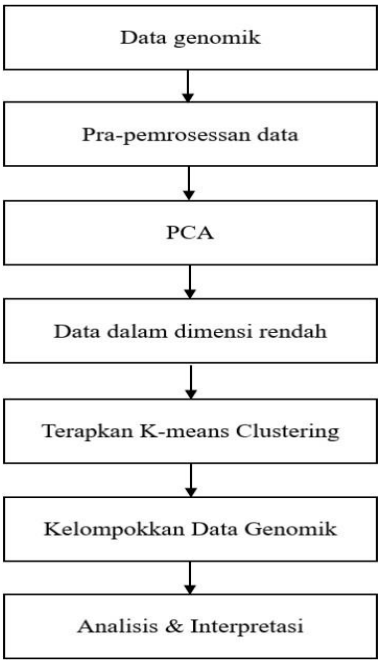

Figure 1. Block Diagram of PCA Principal Component Analysis

After the data has been dimension reduced using PCA, the next step is to cluster the data using the K- Means Clustering method. The process of applying K-Means after dimension reduction is illustrated below.

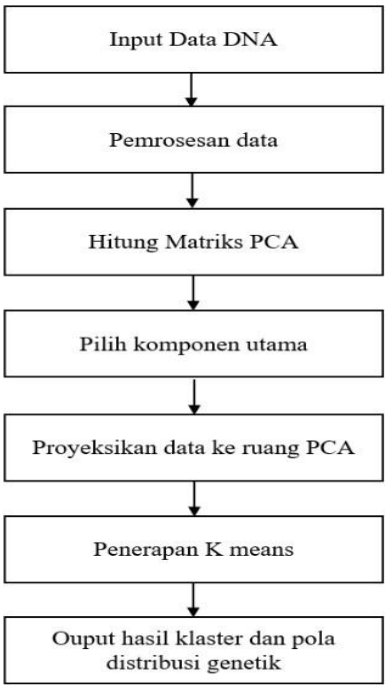

Figure 2. Block Diagram of K-Means Application Process

To reduce the size of complex genomic data, the preceding process diagram shows the Principal Component Analysis (PCA) step. The PCA process identifies the principal components by calculating the covariance matrix of the preprocessed input DNA data. Next, it extracts the eigenvalues and eigenvectors. To project the data to a low-dimensional space, the component with the highest variance is selected. Advanced analyses, such as data clustering using the K-Means method, are facilitated by these PCA results, which result in a simpler yet still informative representation of the data (Toraismaya et al., 2020).

2.3. Linear and Non-Linear Models in DNA base frequency distribution analysis

Linear models are used in DNA base frequency distribution analysis to describe the simple relationship between base frequency and other biological variables. However, non-linear models are important for describe more complex relationships because DNA base data usually has complex and non-linear patterns. For example, linear models can be used to quantify the effect of base frequency variation on genetic traits, while non-linear models such as decision trees and Support Vector Machines (SVM) can aid in the classification of complex base distribution patterns. In addition, DNA base frequency data can be clustered using clustering methods such as K-Means based on pattern similarity; this aids the understanding of structure and variation in genomic data (Ayuningtyas et al., 2020; Suryanto, 2024).

2.4. Clustering Method

Clustering is an unsupervised learning technique commonly used in genomic data analysis to group gene expression data or DNA base frequency distributions based on similarity patterns. The goal is to find natural structures or groups in the data without the need for prior class labelling. Genomics uses clustering to group gene expression data that has similar patterns. To find clusters in data, techniques such as hierarchical clustering or K- means are often used (Chandra, 2022). This can help in gene function analysis or disease classification (Chandra, 2022).

2.5. K-Means Clustering Method

K-Means clustering method is one of the most popular unsupervised learning methods and is widely used in genomic data analysis, such as DNA base frequency distribution research. The K-Means process starts with a randomly selected initial centroid. Then, based on geometric distance, each data point is assigned to the nearest centroid. Next, the centroid position is updated by calculating the average of the points belonging to the cluster. Once the centroid and cluster positions have not changed significantly, this iteration is repeated. Using K-Means, DNA base frequency distribution data can be clustered based on pattern similarity, which helps in finding significant genetic features and simplifies the interpretation of complex genomic data (Chandra, 2022). The following are block diagram of K-Means process in genomic data processing:

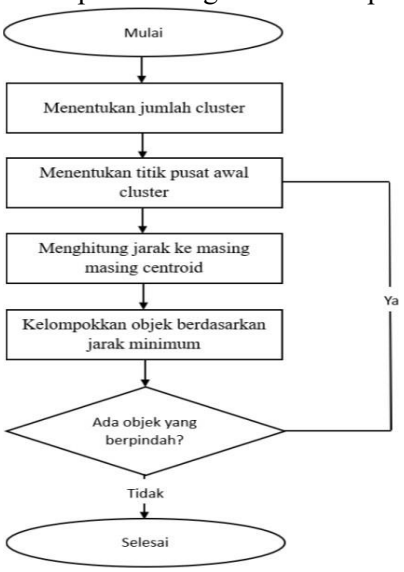

Figure 3. Block Diagram of K-Means Process

2.6. Hypothesis Testing in Genomics

In applying K-Means Clustering to DNA base frequency distribution data, cluster validation is an important step to ensure that the groups formed truly reflect meaningful biological patterns. With this approach, the clustering results are not only visually visible, but also supported by statistical measures that strengthen the scientific interpretation of the DNA base distribution within each group (Suryanto, 2024).

3. METHOD

This study adopts a quantitative methodology involving the application of statistical models to genomic data, particularly focusing on computing base frequencies and evaluating probabilistic distributions across DNA sequences (Ambarwati, 2020). The analysis includes an evaluation of base frequency distributions using descriptive and inferential statistics to identify patterns that may indicate genomic structure or biological function.

3.1. Analysis Method

The analysis of base distribution in DNA sequences was conducted through a structured series of statistical steps, beginning with data preprocessing. DNA sequence data, represented as character strings containing nucleotide bases (A, T, C, and G), were provided as input for analysis.

After data preparation, the absolute frequency of each nucleotide base (A, T, C, G) within the sequence was computed. This step quantifies the occurrence of each base, providing a basis for further statistical analysis of distributional patterns.

Subsequently, the relative frequency (empirical probability) of each base was calculated by dividing its absolute count by the total number of nucleotides. This normalization allows for comparison across sequences of varying lengths and provides insight into base usage patterns.

To visualize the distribution patterns, base frequencies were plotted using histograms generated with Matplotlib, a Python-based data visualization library. These visual representations aid in identifying whether the base occurrences are uniformly distributed or exhibit skewness.

To support interpretation, descriptive statistics including the mean and standard deviation of base frequencies were computed. These metrics quantify the variability among base occurrences and help assess whether the distribution is balanced or biased across the sequence.

3.2. Data Used

The data used in this study is a DNA sequence consisting of a series of nitrogenous bases: A (Adenine), T (Thymine), C (Cytosine), and G (Guanine) (Kusuma & Adrianus, 2020). DNA sequence data can be obtained from Kaggle, which provides public genomic data from various species including humans.

The DNA sequence in this dataset comes from the DNA Sequence Dataset with the file human.txt which is part of the human chromosome stored in text format. Here is an example of the DNA sequence used:

ATGCCCCAACTAAATACTACCGTATGGCCCACCATAATTACCCCCATACTCCTTACACTATTCCTCATCA
CCCAACTAAAAATATTAAACACAAACTACCACCTACCTCCCTCACCAAAGCCCATAAAAATAAAAAATTA
TAACAAACCCTGAGAACCAAAATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCACAATCCTAG

This dataset is used as the basis for analysing DNA base frequency distribution and clustering using the K- Means Clustering method.

3.3. Analysis Steps

The DNA base frequency distribution analysis steps are carried out systematically. First, it starts from Providing Data by providing several DNA sequences in strings consisting of A (Adenine), T (Thymine), C (Cytosine), and G (Guanine) characters. After the data has been provided, the Base frequency calculation is carried out using the count() method in python to calculate the number of occurrences of each base in the entire DNA sequence (Umam et al., 2020).

Next is to calculate the base probability, this probability is obtained by dividing the frequency of each base by the total length of the DNA sequence. To provide a visual understanding of the

frequency distribution of the data, a histogram is used to describe the frequency distribution of each base, so that the distribution pattern can be seen directly (Saputral et al., 2020).

Finally, basic statistical values are calculated, which can give an idea of how far the variance of frequency distribution can occur and how even the distribution of bases in the analysed DNA sequence is (Saputral et al., 2020).

3.4. Evaluation and interpretation of results

Histogram is one of the most useful data visualisation techniques for DNA base frequency distribution analysis. Histograms provide a visual representation of the base frequency distribution. With histogram, we can see if there is a striking imbalance between one base and another (Saputri, 2020).

Histograms are essential for the data exploration stage, especially when working with complex and large biological data. This is because the use of histograms in DNA frequency distribution analysis helps with initial understanding before entering advanced statistical stages such as variance analysis or clustering (Ningrum et al., 2021).

4. RESULTS AND DISCUSSION

The first process in genomic data analysis is to calculate the frequency and likelihood of each base in the analysed DNA sequence. This process is done with the aim of gaining an understanding of how key bases appear in a given sequence.

4.1 Process Flow of K-Means Clustering Analysis on DNA data

In the first stage, researchers will explain the flow of the analysis process using the K-Means Clustering method to cluster DNA base frequency data. The following is a block diagram for the K-Means analysis process flow:

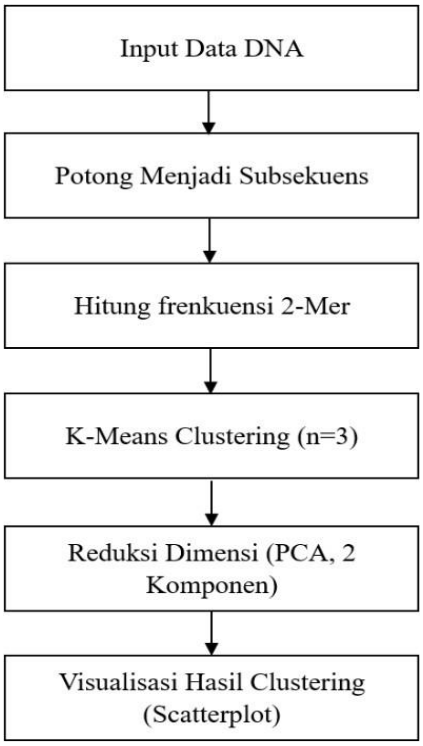

Figure 4. Block Diagram of K-Means Analysis Process

First is to read the DNA file and then cut it into subsequences next Clustering the DNA data using K-Means and PCA is applied to reduce the dimensions of the K-Mer feature results (without cluster columns) and after completion it will display the visualisation of the clustering results in 2D with the PCA results.

4.2 Results of Frequency and Probability Analysis of Bases

At this stage, we calculated the frequency of base occurrence of each analysed DNA sequence, which consisted of the letters A, T, G, C, and N. The results are as follows:

Table 1. Frequency of Each Base

Basa	Frenquency
A (adenine)	1.402.159
T (Thymine)	1.222.500
G (Guanine)	1.453.104
C (Cytosine)	1.456.269
N (unknown)	530

The table above shows that base C has the highest frequency (1,456,269), followed by base G (1,453,104), then bases A (1,402,159) and T (1,222,500) and finally N (530). Frequency differences between these bases indicate an imbalance in the distribution of their occurrence in the DNA sequences studied.

After calculating the frequencies, we also calculated the relative likelihood (probability) of each base to the total number of bases in the DNA sequence. This probability is obtained by dividing the frequency of each base by the total number of bases, which is 5,534,562. The probability value reflects how likely a base is to appear in the entire DNA sequence analysed, the results are as follows:

Table 2. Probability of Each Base

Basa	Probability
A (adenine)	0.2533
T (Thymine)	0.2209
G (Guanine)	0.2626
C (Cytosine)	0.2631
N (unknown)	0.0001

The probability table above shows that base C has the highest probability (0.2631), and slightly more than base G (0.2626) then bases A (0.2533) and T (0.2209) and finally N (0.0001). These probabilities indicate the propensity of each base to appear in the analysed DNA sequence, with base C having a slightly higher probability than base G.

4.3 K-Means Clustering Algorithm

K-Means clustering is an invisible learning method used to group a set of data into a number of clusters based on similar characteristics. DNA base frequency uses the K-means algorithm to cluster the base frequency positions of certain DNA bases based on the distribution features obtained. The following are the centroid results for each iteration of each cluster during the clustering process until convergence:

Table 3. Iteration 1

Cluster	A	T	C	G
C1	120.3	110.2	130.5	118.4
C2	980.4	900.7	970.8	890.6
C3	1450.1	1200.4	1350.2	1250.7

Table 4. Iteration 2

Cluster	A	T	C	G
C1	160.1	148.9	205.0	172.4
C2	326.3	284.1	329.3	323.1
C3	1406.5	1224.7	1391.0	1286.9

Table 5. Iteration 3

Cluster	A	T	C	G
C1	159.7	147.9	204.6	173.0
C2	326.3	284.1	329.3	323.1
C3	1406.5	1224.7	1391.0	1286.9

After the 2nd iteration, the centroids C2 and C3 are stable, meaning that they do not change, so the process is stopped with 3.iteration.

4.4 Base Frequency Distribution (Visualisation)

The results of the frequency analysis were also visualised in the form of a bar graph to facilitate interpretation of the base frequency distribution. This graph shows the frequency distribution of each base in the DNA sequence more clearly. A histogram of the base frequency distribution follows:

Figure 5. Visualisation of Clustering Results

In DNA sequences, the distribution of base frequencies tends to be unbalanced, as shown in the graph above. The frequency of base A is higher than bases C and G, which helps understand the imbalance in the sequence.

4.5 Base Frequency Distribution (Visualisation)

At this stage, a basic analysis of the frequency of occurrence of DNA bases was carried out. The data analysed included total occurrences, mean, minimum value, maximum value and standard deviation.

Table 6. Basic Statistics of DNA Base Frequencies

Basa	Total	Average	Minimum	Maximum	Standard Deviation
A	1.402.159	320,13	2	5.460	353,01
T	1.222.500	279,11	1	5.407	301,46
C	1.456.269	332,48	0	3.693	334,50
G	1.453.104	331,76	2	4.361	330,05

The table above shows that base C has the highest number of base occurrences (1,456,269) and base G the highest number of base occurrences (1,453,104). In contrast, base T has the lowest number of base occurrences, which is 1,222,500. The highest values in the frequency of base occurrences per unit of data are found in bases C (332.48) and G (331.76), indicating a relatively high and even distribution. The minimum and maximum values show considerable variation in frequency between positions, with the highest value in base A at 5,460.

The high standard deviation of the four bases (around 300) shows a significant spread in frequency, indicating significant variation in the distribution of occurrences of these DNA bases.

5. CONCLUSION

This study aimed to apply the K-means clustering algorithm to analyse DNA base frequency distributions with the goal of identifying patterns relevant to genetic variation. By leveraging unsupervised clustering, the method enabled the detection of latent structures within genomic data, independent of predefined labels. The analysis revealed that imbalances in base distribution patterns may reflect underlying genetic variability, offering a potential foundation for exploratory genomic studies. Although preliminary, these results suggest the feasibility of employing such statistical models in supporting applications such as biomarker discovery and genomic stratification in biomedical research.

REFRENCES

Adiyana, I., Sumertajaya, I. M., & Afendi, F. M. (2022). Application of fuzzy C-means and weighted scoring methods for mapping blankspot villages in Pemalang Regency. *Indonesian Journal of Statistics and Its Applications, 6*(1), 77–89. https://doi.org/10.29244/ijsa.v6i1p77-89

Agustina, D., Putri, E., Fauzi, F., Alawiyah, S. N., & Wasono, R. (2020). Application of support vector machine (SVM) method for classification of microarray gene expression data. In *Proceedings of Edusainstech Seminar* (pp. 284–289).

Alhadi, B. (2020). *Development of genomic biomarkers with a data science approach for diabetes and cancer disease analysis* [Master's thesis, University of Indonesia].

Ambarwati, E. (2020). *Introduction to quantitative genetics*. UGM Press.

Ayuningtyas, D., Sartono, B., & Afendi, M. (2020). Application of genetic algorithm for selection of logistic regression variables. *Indonesian Journal of Statistics and Its Applications, 9*(1).

B, Y., & Song, H. (2020). Logistic regression for genomic studies. *Genomic Applications Journal, 4*(2), 45–57.

Chandra, R. A. (2022). Application of statistical techniques in the analysis of genetic variation. *Journal of Biology and Statistics, 4*(2), 101–115.

Fadli, A., & Kusuma, W. A. (2020). Association of single nucleotide polymorphism and type 2 diabetes mellitus disease phenotype using gradient boosting. *IPB University Repository*. https://repository.ipb.ac.id/handle/123456789/104131

Jajang, J., Pratikno, B., & Mashuri, M. (2022). Modelling dengue fever by using conditional autoregressive Bessag-York-Mollie. *Indonesian Journal of Statistics and Its Applications, 6*(1), 101–113. https://doi.org/10.29244/ijsa.v6i1p101-113

Kusuma, W. A., & Adrianus, A. (2020). Constructing bidirected overlap graph for DNA sequence assembly. *Journal of Information Technology and Computer Science (JTIIK), 7*(2), 407–416. https://doi.org/10.25126/jtiik.202072070

Muningsih, E., Hasan, N., & Sulistyo, G. B. (2020). Application of principal component analysis (PCA) method for clustering data on foreign tourist visits to Indonesia. *Bianglala Informatika, 8*(2). Retrieved from http://www.bps.go.id

Ningrum, M. D., Rizal, A., & Nurhayati, S. (2021). Nucleotide frequency visualisation for genetic pattern identification in bioinformatics. *Journal of Computer Technology and Systems, 9*(2), 145–152.

Putri, E., Fauzi, F., Alawiyah, S. N., & Wasono, R. (2020). Application of support vector machine (SVM) methods for classification of microarray gene expression data. In *Proceedings of Edusainstech Seminar*.

Saputral, A., & Sari, D. (2020). Application of genomic data analysis techniques using Python. *XYZ University Informatics Journal, 5*(2), 123–135. https://doi.org/10.1234/jinfor.v5i2.5678

Saputri, A. (2020). *Analysis and visualisation of DNA multiple sequence alignment using dynamic programming Needleman-Wunsch and neighbor-joining tree* [Undergraduate thesis].

Saudale, F. Z. (2020). Biochemistry in the era of genomic big data: Challenges, applications and innovation opportunities. *Chemistry Notes, 2*, 21–43.

Sinaga, V. T. R. A., & Rahmawati, R. (2020). Comparison of principal component regression with partial least squares regression on human development index of East Java Province. *Gaussian Journal, 8*(4), 496–505.

Siswantining, T., Vivaldi, K. G., Sarwinda, D., Soemartojo, S. M., Mattasari, I., & Al-Ash, H. (2022). Implementation of ensemble self-organising maps for missing values imputation. *Indonesian Journal of Statistics and Its Applications, 6*(1), 1–12. https://doi.org/10.29244/ijsa.v6i1p1-12

Suryanto, S. (2024). Statistical models in genomics: Applications and analysis techniques. *Statistics Indonesia, 7*(1), 65–78.

Toraismaya, A., Sasongko, L. R., & Rondonuwu, F. S. (2020). Principal component and K-means cluster analysis for spectrum data of black tea grades for assessment alternative quality. *Journal of Fundamental Mathematics and Applications (JFMA), 3*(2), 148–157. https://doi.org/10.14710/jfma.v3i2.8663

Umam, K., & Sagara, R. (2020). Use of N-mers frequency in DNA sequence analysis. *Jambura Journal of Mathematics, 2*(2), 73–86. https://doi.org/10.34312/jjom.v2i2.4320