Available online at https://li02.tci-thaijo.org/index.php/ssstj

Suan Sunandha Science and Technology Journal
©2024 Faculty of Science and Technology, Suan Sunandha Rajabhat University

The Efficacy of Clustering Algorithms for Young '*Nam-Hom*' Coconut Gene Expression Data in Unveiling the Specific Genes Determining the Flavor: A Comparative Analysis of K-means and Fuzzy C-means

Kairung Hengpraprohm[1], Supoj Hengpraprohm[1*], Kriengkrai Meethaworn[2]

[1]Programm in Data Science,
[2]Programm in Agricultural innovation and management,
Faculty of Science and Technology, Nakhon Pathom Rajabhat University. 85 Malaiman Rd., Mueang, Nakhon Pathom, 73000, Thailand.
***Corresponding author e-mail: supojn@yahoo.com**

Received: *31 March 2024* / **Revised:** *20 May 2024* / **Accepted:** *12 July 2024*

Abstract

This study explores the application of K-means and Fuzzy C-means clustering techniques to analyze gene expression data related to the flavor of young '*Nam-Hom*' coconuts. By comparing these clustering methods, the research aims to identify gene clusters that significantly influence the aromatic and off-flavor profiles of young '*Nam-Hom*' coconuts stored at different temperatures (4°C and 25°C). Specifically, our findings highlight clusters involved in lipid metabolism and cold stress response which are crucial for developing desirable and undesirable flavors, such as LOX1 and ADH2 genes. The study advances our understanding of coconut genetics demonstrates the utility of clustering techniques in agricultural genomics, offering valuable pathways for future genetic enhancement and storage optimization strategies aimed at improving coconut aroma.

Keywords: Young '*Nam-Hom*' coconut, Gene expression data, Clustering algorithms

1. Introduction

Young '*Nam-Hom*' coconut (*Cocos nucifera* L.) is a crop of paramount economic importance and a symbol of agricultural heritage in tropical countries including, Thailand. It is a globally popular fruit because of its refreshing, high nutritional value, and tender-soft jelly-like endosperm (kernel) (Yong, Ge, Ng, & Tan, 2009). Moreover, it plays a significant role in the international market, and Thailand is the leading exporter of young '*Nam-Hom*' coconut. The main markets are the United States and China. In 2023, the exported value was more than 640,810 tons, 430 million USD (Office of Agricultural Economics, 2024). Despite the global challenges posed by the COVID-19 pandemic in 2020, the export of young '*Nam-Hom*' coconuts from Thailand experienced a notable increase of 30%. The demand for coconuts, both in fresh forms and for processing into various products, remains high. Specifically, a significant domestic and international demand for '*Nam-Hom*' coconut varieties contributes to an export value exceeding one billion baht annually, with an upward trend observed.

However, the thriving export market for Thai young '*Nam-Hom*' coconuts encounters several production and preservation challenges, including cracking of fruits (Meethaworn, 2021; Pakcharoen, Meethaworn, & Mohpraman, 2012), browning reaction after peeled (Mohpraman & Siriphanich, 2012), aromaloss (Siriphanich et al., 2011), and the development of off-flavor during long-term storage at low temperatures (Meethaworn, Imsabai, Zhang, Chen, & Siriphanich, 2022; Meethaworn, Luckanatinwong, Zhang, Chen, & Siriphanich, 2019). These issues, particularly off-flavor-flavor during cold storage, necessary for the minimum four-week shipping period to international destinations, represent a critical challenge. Coconuts are stored at temperatures between 2 and 4 degrees Celsius to maintain freshness during transport. However, this can lead to a phenomenon known as chilling injury, resulting in the emergence of rancid oil-like odors in the coconut kernel and coconut water of the young '*Nam-Hom*' coconuts.

Previous studies have indicated that only a subset of genes expressed at different temperatures are implicated in the developed off-flavor-flavors, suggesting a complex genetic basis for this trait. Therefore, this research aims to explore the gene expressions (traits) influencing the flavor in young '*Nam-Hom*' coconuts, employing machine learning techniques such as data clustering with K-means and Fuzzy C-means algorithms. By identifying informative features (genes) through clustering similar gene expression data, this study intends to reveal the groups of genes associated with the development of flavor in young '*Nam-Hom*' coconuts. The goal is to use this information to improve or limit the occurrence of off-flavor-flavor during storage, thereby enhancing the quality of coconuts for export.

2. Backgrounds

2.1 K-means clustering

K-means clustering (Sinaga and Yang, 2020) is an essential unsupervised learning technique utilized across various disciplines, including bioinformatics, for effectively organizing large datasets into meaningful clusters. This algorithm partitions a dataset into k distinct clusters based on the similarity of data points, making it particularly useful for identifying patterns and groupings in complex data sets such as gene expression profiles.

Workflow Summary:

(1) *Initialization*: Begin by selecting k initial centroids randomly or through a more deliberate method, where k is the desired number of clusters.

(2) *Assignment*: Assign each data point in the dataset to the nearest centroid, typically determined by the shortest Euclidean distance, thereby forming preliminary clusters.

(3) *Update*: Recalculate the position of each centroid to be the mean of the data points assigned to its cluster, effectively moving it to the cluster center.

(4) *Iteration*: Repeat the assignment and update steps iteratively until the centroids stabilize (no longer move significantly), signaling the algorithm has converged on a solution.

K-means aims to minimize within-cluster variance, the sum of squared distances between data points and their respective cluster centroid, achieving compact, homogenous clusters.

2.2 Fuzzy C-Means clustering

Fuzzy C-Means (FCM) clustering (Bezdek et al., 1984) extends beyond the traditional hardly partitioning methods, like K-means, by allowing data points to belong to multiple clusters with varying degrees of membership. This soft clustering technique is valuable, particularly in fields: such as bioinformatics, where the biological phenomena under study often exhibit inherent ambiguities and overlapping characteristics.

Workflow Summary:

(1) *Initialization*: Select the number of clusters, k, and initialize the cluster centers randomly or based on a heuristic. Unlike K-means, FCM starts with an assumption about the cluster centers that will be refined.

(2) *Membership Assignment*: Assign each data point a degree of membership for each cluster. This degree ranges from 0 (no membership) to 1 (full membership), based on the distance to the cluster centers, allowing data points to be partially in more than one cluster.

(3) *Centroid Update*: Update the cluster centers to reflect the calculated memberships, with each center being the weighted mean of all points, where weights are the membership degrees. This step accounts for the fuzziness of cluster boundaries by considering the degree to which points belong to clusters.

(4) *Iteration*: Iterate the membership assignment and centroid update steps until the cluster centers stabilize within a predetermined tolerance or until a maximum number of iterations is reached. Convergence is achieved when changes in the degrees of membership between two consecutive iterations are negligible.

The primary goal of FCM is to minimize an objective function that represents the distance between data points and cluster centers, weighed by the membership degrees. This minimization leads to the formation of clusters that best capture the underlying structure in the data, considering the fuzziness of natural groupings.

2.3 The Silhouette Score

The Silhouette Score (Shahapure & Nicholas, 2020) is a metric used to assess the quality of clusters created by a clustering algorithm. Introduced by Peter J. Rousseeuw in 1987, this measure evaluates how similar an object is to its cluster compared to the other clusters. The value of the Silhouette Score ranges from -1 to 1, where a high score indicates that the object is well-matched to its cluster and poorly matched to neighboring clusters, thus providing a clear indication of the appropriateness of the cluster assignments.

Calculation of the Silhouette Score: The Silhouette Score for each sample in the dataset is calculated using the formula $s=(b-a)/max(a,b)$, where:

a is the average distance from the sample to all other points in the same cluster.

b is the smallest average distance from the sample to points in a different cluster, minimized over clusters.

Interpretation:

A score approaching 1, signifies that the sample is far from the neighboring clusters.

A score approaching 0, indicates that the sample is on or very close to the decision boundary between two neighboring clusters.

A score approaching -1, means the sample has been assigned to the wrong cluster.

In the context of gene expression data analysis, the Silhouette Score facilitates the evaluation of clustering algorithms like K-means and Fuzzy C-means, offering a quantitative measure to compare the performance of different models. By utilizing this metric, researchers can more confidently discern the underlying patterns in gene expression, leading to more informed conclusions about genetic influences on traits of interest, such as the flavor of coconuts.

2.4 Related research

Alagukumar, S. and Lawrance, R. conducted a comparative study on the efficiency of two association rule mining techniques, FP-Growth and Apriori, applied to different microarray datasets for analyzing gene expression levels. Their results showcased FP-Growth's superior performance in terms of computational speed, underscoring the potential of sophisticated data mining techniques in unraveling complex gene expression patterns (Alagukumar & Lawrance, 2015).

Saensuk et al. studied a unique Thai dwarf green coconut variety known for its distinct "pandan-like" aroma, attributed to the compound 2-acetyl-1-pyrroline (2AP). Their transcriptomic analysis sought to identify the genes responsible for 2AP biosynthesis, providing insights into the genetic basis of flavor profiles in coconuts (Saensuk et al., 2016).

Zhu et al. presented a new approach to data clustering named subspace clustering guided unsupervised feature selection (SCUFS). This technique focuses on selecting representative data features that maintain the integrity of data subgroups, demonstrating that SCUFS can outperform traditional unsupervised feature selection methods in clustering efficacy (Zhu, Zhu, Hu, Zhang, & Zuo., 2017).

Meethaworn et al. explored the effects of low-temperature storage on the developed off-flavor in young coconuts, attributing the phenomenon to the lipoxygenase (LOX) pathway. Their research, which varied storage conditions, indicated that lower temperatures exacerbate the production of undesirable odors, pointing to specific genes involved in lipid oxidation (Meethaworn et al., 2019).

Hengpraprohm et al. investigated the identification of molecular markers indicating the developmental stages of ovarian maturation in black tiger shrimp using microarray data. By applying data mining techniques, they have shown that they can find the genetic information crucial for reproductive system functions, highlighting the application of genomic analyses in improving aquaculture breeding practices (Hengpraprohm, Jungjit, Hengpraprohm, & Thammasiri, 2019).

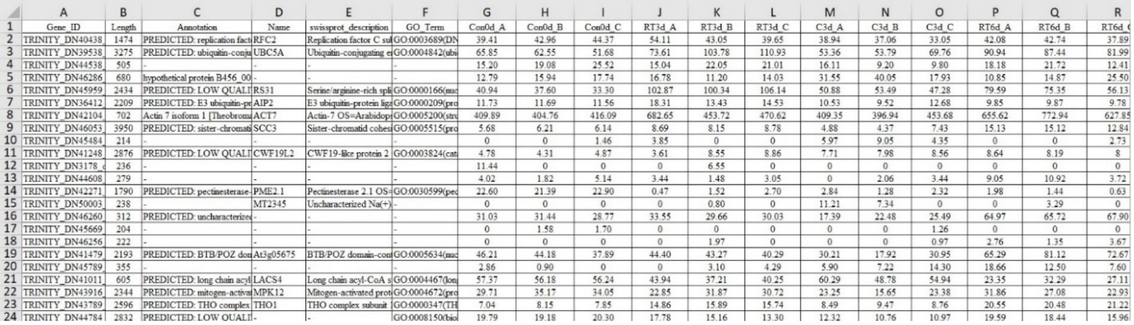
These highlighted studies demonstrate a broad spectrum of approaches and technologies applied in genomics and bioinformatics, ranging from association rule mining to advanced clustering and feature selection techniques. The collective insights from these research efforts deepen our understanding of genetic mechanisms behind important traits offer practical avenues for agricultural innovation and food production enhancement.

3. Materials and Methods

This study employed K-means and Fuzzy C-means clustering techniques to investigate the gene expression groups related to the flavor of young ‘*Nam-Hom*’ coconuts. The methodology is outlined in the following steps:

3.1 Data collection

This step involved collecting and selecting data to ensure its suitability for model development. The research utilized secondary data representing gene expression levels associated with scent development in young coconuts. The dataset used in the experiment was from Meethaworn et al. (2019). This data was collected from coconuts stored at two different temperatures, 4°C and 25°C, for 15 days. The gene expression data, in the form of RNA sequencing technique, was converted into numerical values indicating the expression levels of genes, which were then organized into a Microsoft Excel format as illustrated in Figure 1.



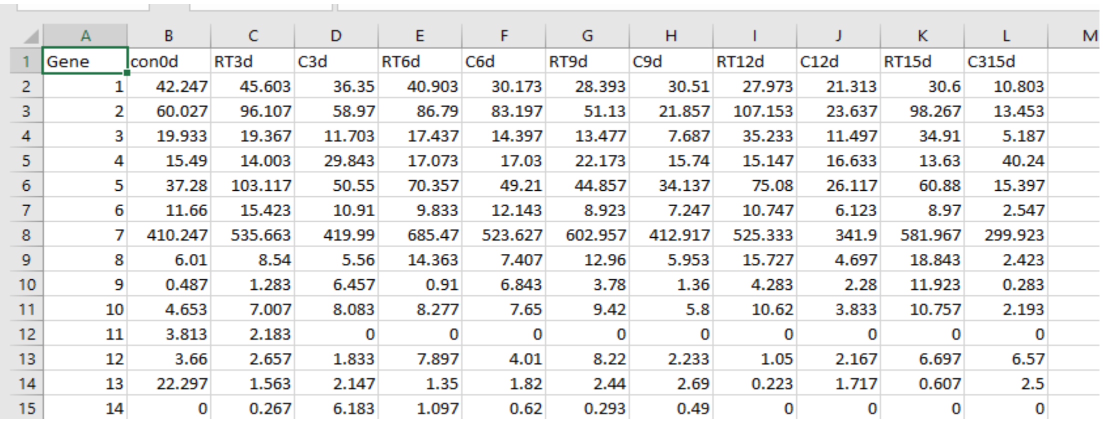
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Gene_ID	Length	Annotation	Name	swissprot_description	GO_Term	Con0d_A	Con0d_B	Con0d_C	RT3d_A	RT3d_B	RT3d_C	C3d_A	C3d_B	C3d_C	RT6d_A	RT6d_B	RT6d_(
2	TRINITY_DN40438_	1474	PREDICTED: replication fact	RFC2	Replication factor C sul	GO:0003689(DN	39.41	42.96	44.37	54.11	43.05	39.65	38.94	37.06	33.05	42.08	42.74	37.89
3	TRINITY_DN39538_	3275	PREDICTED: ubiquitin-conju	UBC5A	Ubiquitin-conjugating e	GO:0004842(ubi	65.85	62.55	51.68	73.61	103.78	110.93	53.36	53.79	69.76	90.94	87.44	81.99
4	TRINITY_DN44538_	505	-	-	-	-	15.20	19.08	25.52	15.04	22.05	21.01	16.11	9.20	9.80	18.18	21.72	12.41
5	TRINITY_DN46286_	680	hypothetical protein B456_00	-	-	-	12.79	15.94	17.74	16.78	11.20	14.03	31.55	40.05	17.93	10.85	14.87	25.50
6	TRINITY_DN45959_	2434	PREDICTED: LOW QUALI	RS31	Serine/arginine-rich spli	GO:0000166(nuc	40.94	37.60	33.30	102.87	100.34	106.14	50.88	53.49	47.28	79.59	75.35	56.13
7	TRINITY_DN36412_	2209	PREDICTED: E3 ubiquitin-pr	AIP2	E3 ubiquitin-protein liga	GO:0000209(pro	11.73	11.69	11.56	18.31	13.43	14.53	10.53	9.52	12.68	9.85	9.87	9.78
8	TRINITY_DN42104_	702	Actin 7 isoform 1 [Theobroma	ACT7	Actin-7 OS=Arabidops	GO:0005200(stru	409.89	404.76	416.09	682.65	453.72	470.62	409.35	396.94	453.68	655.62	772.94	627.85
9	TRINITY_DN46053_	3950	PREDICTED: sister-chromati	SCC3	Sister-chromatid cohesi	GO:0005515(pro	5.68	6.21	6.14	8.69	8.15	8.78	4.88	4.37	7.43	15.13	15.12	12.84
10	TRINITY_DN45484_	214	-	-	-	-	0	0	1.46	3.85	0	0	5.97	9.05	4.35	0	0	2.73
11	TRINITY_DN41248_	2876	PREDICTED: LOW QUALI	CWF19L2	CWF19-like protein 2	GO:0003824(cat	4.78	4.31	4.87	3.61	8.55	8.86	7.71	7.98	8.56	8.64	8.19	8
12	TRINITY_DN3178_c	236	-	-	-	-	11.44	0	0	0	6.55	0	0	0	0	0	0	0
13	TRINITY_DN44608_	279	-	-	-	-	4.02	1.82	5.14	3.44	1.48	3.05	0	2.06	3.44	9.05	10.92	3.72
14	TRINITY_DN42271_	1790	PREDICTED: pectinesterase-	PME2.1	Pectinesterase 2.1 OS=	GO:0030599(pec	22.60	21.39	22.90	0.47	1.52	2.70	2.84	1.28	2.32	1.98	1.44	0.63
15	TRINITY_DN50003_	238	-	MT2345	Uncharacterized Na(+)	-	0	0	0	0	0.80	0	11.21	7.34	0	0	3.29	0
16	TRINITY_DN46260_	312	PREDICTED: uncharacterize	-	-	-	31.03	31.44	28.77	33.55	29.66	30.03	17.39	22.48	25.49	64.97	65.72	67.90
17	TRINITY_DN45669_	204	-	-	-	-	0	1.58	1.70	0	0	0	0	0	1.26	0	0	0
18	TRINITY_DN46256_	222	-	-	-	-	0	0	0	0	1.97	0	0	0	0.97	2.76	1.35	3.67
19	TRINITY_DN41479_	2193	PREDICTED: BTB/POZ don	At3g05675	BTB/POZ domain-con	GO:0005634(nuc	46.21	44.18	37.89	44.40	43.27	40.29	30.21	17.92	30.95	65.29	81.12	72.67
20	TRINITY_DN45789_	355	-	-	-	-	2.86	0.90	0	0	3.10	4.29	5.90	7.22	14.30	18.66	12.50	7.60
21	TRINITY_DN41011_	605	PREDICTED: long chain acyl	LACS4	Long chain acyl-CoA s	GO:0004467(lon	57.37	56.18	56.24	43.94	37.21	40.25	60.29	48.78	54.94	23.35	32.29	27.11
22	TRINITY_DN43916_	2344	PREDICTED: mitogen-activat	MPK12	Mitogen-activated prot	GO:0004672(pro	29.71	35.17	34.05	22.85	31.87	30.72	23.25	15.65	23.38	31.86	27.08	22.93
23	TRINITY_DN43789_	2596	PREDICTED: THO complex	THO1	THO complex subunit 1	GO:0000347(TH	7.04	8.15	7.85	14.86	15.89	15.74	8.49	9.47	8.76	20.55	20.48	21.22
24	TRINITY_DN44784_	2832	PREDICTED: LOW QUALI	-	-	GO:0008150(bio	19.79	19.18	20.30	17.78	15.16	13.30	12.32	10.76	10.97	19.59	18.44	15.96

Figure 1. Table of gene expression data of young ‘*Nam-Hom*’ coconut stored at 25°C and 4°C for 15 days using RNA sequencing technique.

3.2 Data preparation

In this phase, the collected data underwent preparation to make it suitable for analysis using Python. It involved formatting the data into a .csv file format, as demonstrated in Figure 2. The structured data consisted of rows and columns where each row represented one of the total 15,000 genes, and the columns included 12 specific data points, detailed as follows:

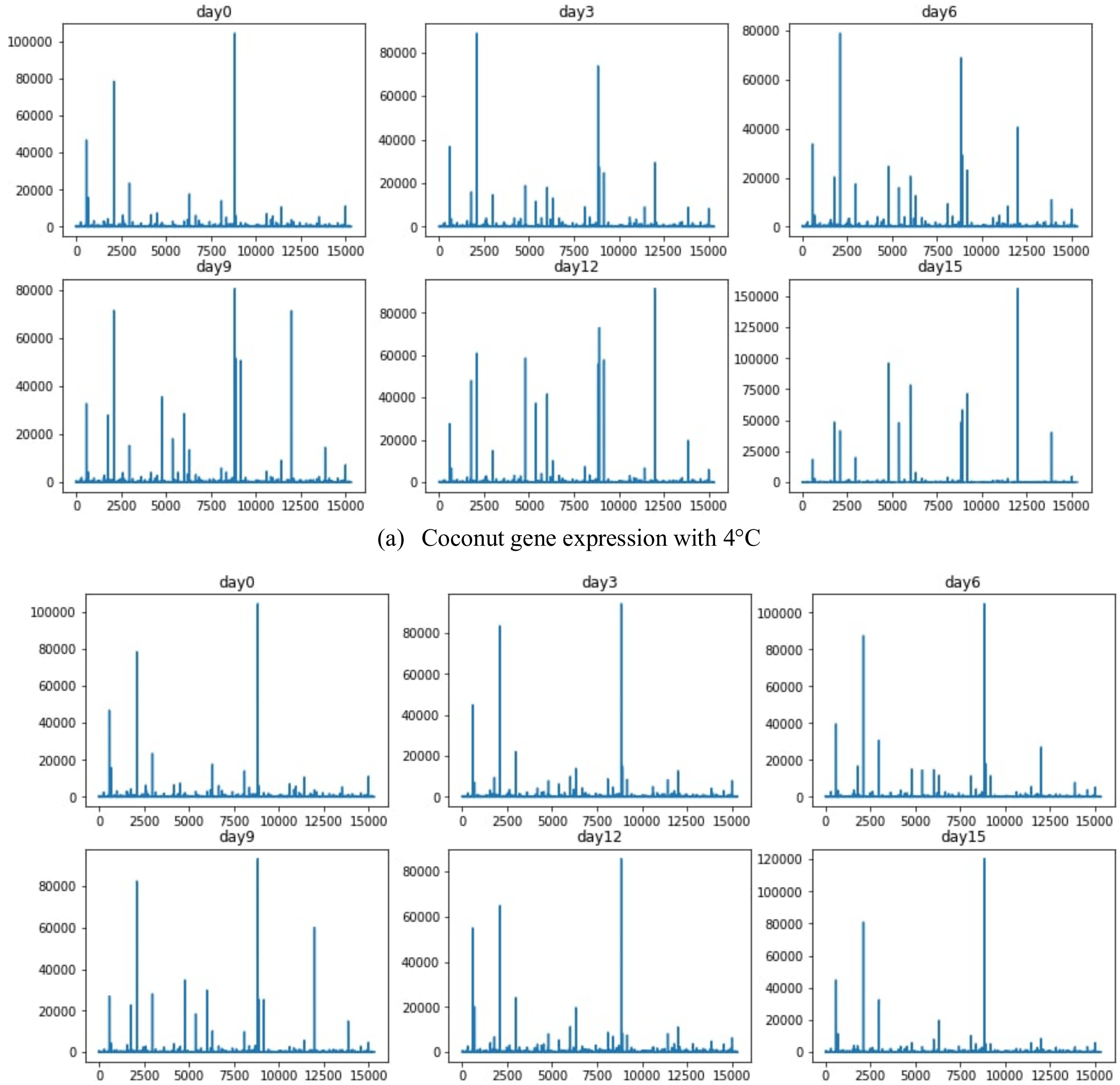
Gene: The gene identifier.
con0d: Gene expression data of young '*Nam-Hom*' at the beginning of storage.
RT3d: Gene expression data of young '*Nam-Hom*' stored at 25°C for 3 days.
C3d: Gene expression data of young '*Nam-Hom*' stored at 4°C for 3 days.
RT6d: Gene expression data of young '*Nam-Hom*' stored at 25°C for 6 days.
C6d: Gene expression data of young '*Nam-Hom*' stored at 4°C for 6 days.
RT9d: Gene expression data of young '*Nam-Hom*' stored at 25°C for 9 days.
C9d: Gene expression data of young '*Nam-Hom*' stored at 4°C for 9 days.
RT12d: Gene expression data of young '*Nam-Hom*' stored at 25°C for 12 days.
C12d: Gene expression data of young '*Nam-Hom*' stored at 4°C for 12 days.
RT15d: Gene expression data of young '*Nam-Hom*' stored at 25°C for 15 days.
C15d: Gene expression data of young '*Nam-Hom*' stored at 4°C for 15 days.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Gene	con0d	RT3d	C3d	RT6d	C6d	RT9d	C9d	RT12d	C12d	RT15d	C315d	
2	1	42.247	45.603	36.35	40.903	30.173	28.393	30.51	27.973	21.313	30.6	10.803	
3	2	60.027	96.107	58.97	86.79	83.197	51.13	21.857	107.153	23.637	98.267	13.453	
4	3	19.933	19.367	11.703	17.437	14.397	13.477	7.687	35.233	11.497	34.91	5.187	
5	4	15.49	14.003	29.843	17.073	17.03	22.173	15.74	15.147	16.633	13.63	40.24	
6	5	37.28	103.117	50.55	70.357	49.21	44.857	34.137	75.08	26.117	60.88	15.397	
7	6	11.66	15.423	10.91	9.833	12.143	8.923	7.247	10.747	6.123	8.97	2.547	
8	7	410.247	535.663	419.99	685.47	523.627	602.957	412.917	525.333	341.9	581.967	299.923	
9	8	6.01	8.54	5.56	14.363	7.407	12.96	5.953	15.727	4.697	18.843	2.423	
10	9	0.487	1.283	6.457	0.91	6.843	3.78	1.36	4.283	2.28	11.923	0.283	
11	10	4.653	7.007	8.083	8.277	7.65	9.42	5.8	10.62	3.833	10.757	2.193	
12	11	3.813	2.183	0	0	0	0	0	0	0	0	0	
13	12	3.66	2.657	1.833	7.897	4.01	8.22	2.233	1.05	2.167	6.697	6.57	
14	13	22.297	1.563	2.147	1.35	1.82	2.44	2.69	0.223	1.717	0.607	2.5	
15	14	0	0.267	6.183	1.097	0.62	0.293	0.49	0	0	0	0	

Figure 2. The prepared gene expression data for the experiment.

This organized data allowed for the comparison of gene expression at different temperatures and time point of storage, aiming to identify variations and patterns in gene expression related to the flavor of young '*Nam-Hom*' coconuts. Figure 3 illustrates the comparative gene expression data of young '*Nam-Hom*' coconuts stored at different temperatures over 15 days. Sub-figure 3(a) presents the comparison data of all 15,000 genes' expression profiles at 4 degrees Celsius across various days. Conversely, Subfigure 3(b) details the gene expression profiles at 25 degrees Celsius.



(a) Coconut gene expression with 4°C

(b) Coconut gene expression with 25°C

Figure 3. Comparative gene expression data of young '*Nam-Hom*' coconuts stored at 4°C and 25°C for 15 days.

By structuring the data in this manner, it became feasible to apply clustering techniques to analyze the gene expressions systematically. This preparation was crucial for the subsequent model construction phase, where the data's structure and organization played a significant role in the effectiveness of the clustering algorithms.

3.3 Model construction

The study aimed to identify significant gene characteristics influencing the scent of coconuts through data mining principles. The K-means and Fuzzy C-means clustering techniques were applied to group the data and compare the performance of these models. The similarity between data clusters and the consistency of model data were evaluated using the Silhouette Score, which ranges from -1 to 1. A score closer to 1 indicates well-suited data grouping, whereas a score near -1 suggests inappropriate clustering. The conceptual framework is depicted in Figure 4.

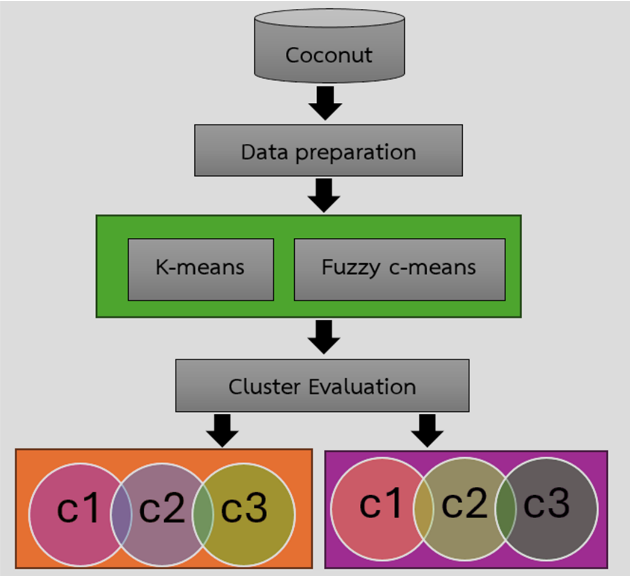


Figure 4. The conceptual framework for gene characteristic selection.

This research began with preparing the young '*Nam-Hom*' coconut data for clustering significant genes, divided into two sets based on storage temperatures at 25°C and 4°C. The K-means and Fuzzy C-means techniques were then used for data clustering. The efficiency of both methods was assessed to determine the best-performing model based on the Silhouette Score. The chosen model's data was then utilized to identify or describe genes impacting the developed flavor in young '*Nam-Hom*' coconuts during storage at both temperatures. The model development was conducted using Python programming, facilitating a thorough comparison and selection process for the genes associated with young '*Nam-Hom*' coconut flavor traits.

4. Results

The comparative analysis of gene clustering related to flavor development in young coconuts highlighted the efficacy of K-means and Fuzzy C-means clustering techniques. According to Table 1, the optimal number of clusters (K) was determined to be 3 based on the analysis depicted in Figure 5, which presents the distortion score elbow chart. The 'elbow' point in the chart is applied to determine the optimal number of clusters. At k=3, there is a noticeable elbow in the chart, indicating a significant reduction in the distortion score with a lesser rate of decrease beyond this point. This suggests that increasing the number of clusters beyond three leads to diminishing returns of clustering compactness and separation improvements. Therefore, k=3 is considered the optimal choice for achieving the most meaningful and distinct clustering of gene expression data without unnecessary complexity.

The K-means clustering showed superior performance achieving a Silhouette Score of 0.996, slightly higher than the Fuzzy C-means, which scored 0.995. This slight difference underscores the precision of both clustering methods in segregating gene expression data effectively.

Table 1. The Silhouette score comparison of K-Means and Fuzzy C-means clustering.

Method	K	Silhouette Score
K-means	3	0.996*
Fuzzy-C-means	3	0.995

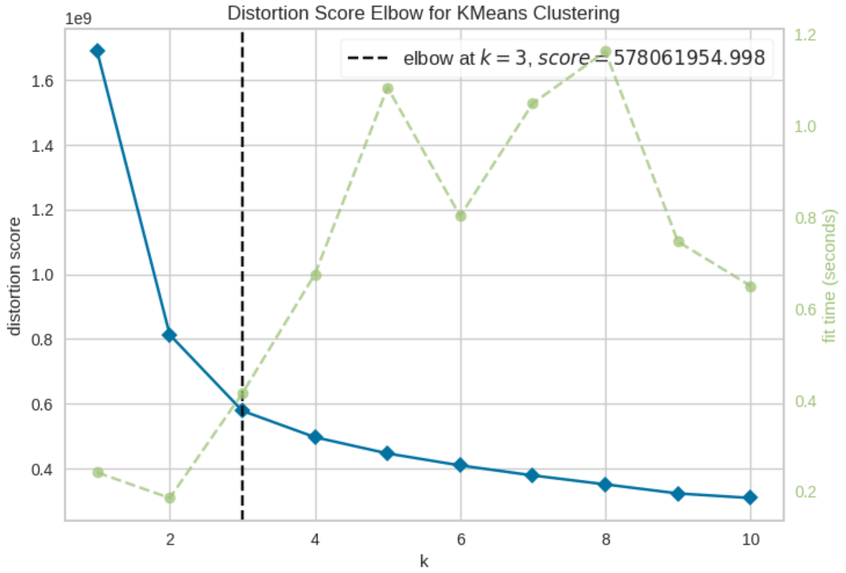


Figure 5. The distortion score elbow for K-Means clustering.

K-means Clustering Findings:

Cluster 1 primarily associated genes with lipid degradation processes and stress response signaling mechanisms during storage. Interestingly, this cluster did not encompass genes directly linked to the generated characteristic '*Nam-Hom*' coconut flavor or off-flavor.

Cluster 2 was identified to contain genes involved with both lipid degradation and transformation processes, indicative of energy utilization for cell nourishment, especially notable during seed germination. Gene expression in this cluster was predominantly observed at 25°C, suggesting that storage at this temperature facilitates energy-related metabolic activities more than at 4°C.

Cluster 3 stood out by including genes related to lipid degradation and the related genes to both '*Nam-Hom*' aroma and off-flavor in young '*Nam-Hom* coconuts. This cluster was the only one encompassing genes like Badh2, which controls the synthesis of 2-AP (Saensuk et al., 2016), a compound linked to the aroma of '*Nam-Hom*' coconuts. The expression of Badh2 was notably higher at 25°C, indicating that lower storage temperatures could diminish the coconut's aromatic quality. Moreover, genes associated with off-flavor development, such as LOX1 and ADH2, showed higher expression levels at 4°C, aligning with the observation that off-flavors are more developed at lower storage temperatures (Meethaworn et al., 2019; Meethaworn et al., 2022).

Furthermore, the exclusive presence of lipase genes in Cluster 3, involved in breaking down free fatty acids from triglycerides, signifies their role in the subsequent oxidation by the LOX pathway, leading to off-flavor generation. The expression of these genes was evident at both 4°C and 25°C, underscoring their potential involvement in scent modulation irrespective of the storage temperature.

The K-means clustering findings delineate three distinct clusters with unique gene expressions linked to the different lipid metabolism aspects and flavor profile modulation. Noted, Cluster 3 includes genes critical for the characteristic synthesis of '*Nam-Hom*' aromas and off-flavors, showing differential expression patterns across the two temperatures studied. Moreover, analysis of the results obtained from the Fuzzy C-means clustering corroborates these findings, displaying similar cluster characteristics and supporting the consistency of the gene grouping trends. This alignment between the two clustering methods underscores the robustness of our clustering approach in capturing the essential aspects of gene expression that influence coconut flavor and quality.

5. Discussion

The results illuminate the complex interplay of genes involved in lipid metabolism and their impact on the aroma compound profile of young '*Nam-Hom*' coconuts during storage. The superior clustering performance of K-means, as evidenced by the higher Silhouette Score, facilitated a clearer delineation of gene groups and their respective functions. The identified specific clusters associated with flavor-related genes provide a foundation for further exploration into genetic engineering and storage practices, aimed at enhancing the aromatic qualities of young ‘*Nam-Hom*’ coconuts while mitigating off-flavors.

This study's findings contribute to the expanded understanding of post-harvest young ‘*Nam-Hom*’ coconut flavor dynamics, emphasizing the importance of storage conditions on gene expression and flavor development. Future research could investigate the regulatory mechanisms governing these flavor-related genes, potentially uncovering new avenues for improving young ‘*Nam-Hom*’ coconut fragrance through genetic manipulation or optimized storage strategies.

6. Conclusions

This research marks a significant step towards understanding the complexities of gene expression related to the flavor of young coconuts, with a focus on the application and comparative efficacy of K-means and Fuzzy C-means clustering techniques. By meticulously analyzing gene expression data under varying storage temperatures, the study unveils the intricacies of young *‘Nam-Hom’* coconut aroma development but also showcases the strengths and nuances of two widely used clustering algorithms in bioinformatics.

Our analysis demonstrated that K-means clustering slightly outperformed Fuzzy C-means, as indicated by the higher Silhouette Score of 0.996. This comparative approach highlighted the subtle differences in clustering efficacy and underscored the suitable K-means for this specific application in genetic data analysis. The optimal clustering identified through K-means facilitated a nuanced understanding of the genetic factors at play, distinguishing between genes associated with various lipid metabolism processes and those directly impacting the aroma profile of young *‘Nam-Hom’* coconuts.

The significance of this study extends beyond the specific insights into young *‘Nam-Hom’*coconut gene expression. It underscores the pivotal role of advanced clustering techniques in deciphering complex biological data, offering a methodological framework for future genetic research. The findings reinforce the utility of clustering algorithms in bioinformatics, particularly in unraveling the genetic basis of phenotypic traits.

Furthermore, this research illuminates the broader applicability of clustering algorithms in agricultural genetics and post-harvest preservation strategies. By optimizing clustering approaches, researchers can gain deeper insights into the genetic mechanisms governing crop quality traits, enabling targeted interventions to enhance desirable characteristics and mitigate negative ones.

In conclusion, while the study provides valuable insights into the genetic determinants of coconut scent, its broader contribution lies in demonstrating the power and potential of clustering algorithms in genetic research. The comparative analysis between K-means and Fuzzy C-means offers a reference point for selecting appropriate clustering methods in similar genomic studies, paving the way for more refined and effective analyses of the quest to link genetic data with phenotypic traits.

Acknowledgements

This research was supported by Nakhon Pathom Rajabhat University (NPRU), Thailand.

References

Alagukumar, S., & Lawrance, R. (2015). A selective analysis of microarray data using association rule mining. *Procedia Computer Science*, *47*, 3-12. doi:10.1016/j.procs.2015.03.177

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy *c*-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191-203. doi:10.1016/0098-3004(84)90020-7

Hengpraprohm, S., Jungjit, S., Hengpraprohm, K, & Thammasiri, D. (2019). Molecular marker discovery for ovarian maturation level of the black tiger shrimp from microarray data using genetic algorithm. *International Journal of the Computer, the Internet and Management*, *27*(2), 43-51.

Meethaworn, K. (2021). Cracking characteristic on polished young coconut and its prevention. *Proceedings of the 13th NPRU National Academic Conference* (pp. 209-216). Nakhon Pathom, Thailand (in Thai).

Meethaworn, K., Imsabai, W., Zhang, B., Chen, K., & Siriphanich, J. (2022). Off-flavor and loss of aroma in young coconut fruit during cold storage are associated with the expression of genes derived from the LOX pathway and Badh2. *The Horticulture Journal*, *91*(2), 209-220. doi:10.2503/hortj.UTD-309

Meethaworn, K., Luckanatinwong, V., Zhang, B., Chen, K., & Siriphanich, J. (2019). Off-flavor caused by cold storage is related to induced activity of LOX and HPL in young coconut fruit. *LWT*, *114*, 108329. doi:10.1016/j.lwt.2019.108329

Mohpraman, K., & Siriphanich, J. (2012). Safe use of sodium metabisulfite in young coconuts. *Postharvest Biology and Technology*, *65*, 76-78.

Office of Agricultural Economics. (2024). *Amount and economic value of aromatic coconut*. Retrieved from http://www.infoservice@oae.go.th

Pakcharoen, A., Meethaworn, K., & Mohpraman, K. (2012). *The occurrence and deterrence of fruit cracking and off-flavor in aromatic coconut during storage at low temperature* (Report No. KU.R.1/2011). Postharvest Technology Innovation Center (in Thai).

Saensuk, C., Wanchana, S., Choowongkomon, K., Wongpornchai, S., Kraithong, T., Imsabai, W., … Arikit, S. (2016). De novo transcriptome assembly and identification of the gene conferring a "pandan-like" aroma in coconut (*Cocos nucifera* L.). *Plant Science*, *252*, 324-334. doi:10.1016/j.plantsci.2016.08.014

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics* (pp. 747-748). Sydney, Australia. doi:10.1109/DSAA49011.2020.00096

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised k-means clustering algorithm. *IEEE Access*, *8*, 80716-80727. doi:10.1109/ACCESS.2020.2988796

Siriphanich, J., Saradhuldhat, P., Romphophak, T., Krisanapook, K., Pathaveerat, S., & Tongchitpakdee, S. (2011). Coconut (*Cocos nucifera* L.). In E. M. Yahia (Ed.), *Postharvest biology and technology of tropical and subtropical fruits*. Woodhead Publishing.

Yong, J. W. H., Ge, L., Ng, Y. F., & Tan, S. N. (2009). The chemical composition and biological properties of coconut (*Cocos nucifera* L.) water. *Molecules*, *14*, 5144-5164. doi:10.3390/molecules14125144

Zhu, P., Zhu, W., Hu, Q., Zhang, C., & Zuo, W. (2017). Subspace clustering guided unsupervised feature selection. *Pattern Recognition*, *66*, 364-374. doi:10.1016/j.patcog.2017.01.016