Jurnal Ilmu Komputer dan Sistem Informasi (JIRSI)
Volume: 5, Nomor: 2, Mei 2026: 515-524
https://jurnal.unity-academy.sch.id/index.php/jirsi/index

e-ISSN 2830-3954
p-ISSN 2830-6031

Use of Data Visualization Techniques in Bioinformatics for Time-Based Gene Expression Pattern Analysis

M. Khalil Gibran*[1], Mhd Ikhsan Rifki[2], Amir Saleh[3]
[1,2]Computer Science, Universitas Islam Negeri Sumatera Utara, [3]Computer Engineering and Informatics, Politeknik Negeri Medan
E-Mail: [1]m.khalil1100000202@uinsu.ac.id, [2]rifki.mhdikhsan@uinsu.ac.id, [3]amirsaleh@polmed.ac.id

Abstract

This study explores data visualization techniques in bioinformatics for analyzing time-series gene expression patterns. It examines how different visualization approaches support the interpretation of large-scale temporal gene expression data. A dataset comprising 4,381 genes across 24 time intervals was analyzed using heatmaps, Principal Component Analysis (PCA), volcano plots, and dendrograms. Heatmaps were used to observe expression correlations, PCA was applied to reduce dimensionality, volcano plots identified differentially expressed genes between conditions, and dendrograms grouped genes with similar expression profiles. The PCA results showed that the first two principal components accounted for 42.32% of the total variance, indicating that these components captured a substantial but not complete portion of the data structure. Volcano plot analysis detected differentially expressed genes based on log2 fold change > 1 and p-value < 0.05, while dendrogram visualization revealed several major clusters with comparable temporal expression patterns. Overall, the findings suggest that combining multiple visualization methods can improve the exploratory analysis of temporal gene expression data by clarifying patterns, highlighting potentially relevant genes, and supporting further biological interpretation. Rather than serving as standalone evidence for clinical application, these visual approaches provide a useful analytical foundation for subsequent validation, biomarker investigation, and large-scale omics research.

Keywords: *Bioinformatics, Visualization, Gene Expression, Heatmap, PCA*

1. INTRODUCTION

Bioinformatics is a cross-disciplinary discipline that integrates computer science, statistics, and biology to manage and understand complex biological data. One of the important aspects of this field is the analysis of gene expression, which plays a role in uncovering the mechanisms of biological regulation and response at the molecular level[1]. A time-based approach to gene expression analysis allows researchers to monitor the dynamics of changes in gene activity, which is crucial in studying biological processes such as cell development, adaptation to stress, and disease evolution[2].

However, the high volume and dimension of gene expression data is often an obstacle to accurate interpretation. Data visualization is an effective solution by presenting numerical data in a more intuitive graphical form[3]. In this study, various visualization methods were used that have proven to be useful, including heatmaps to display correlation or co-expression relationships between genes. The color scale on the heatmap makes it easier to identify groups of genes that have similar or opposite expression patterns. PCA is used as a dimension reduction

method to simplify data complexity and highlight the main sources of variation[4]. Volcano plots are used to display significant differences in gene expression between two conditions based on log2 fold change and p-value[5]. Meanwhile, dendrograms are applied to group genes based on similarity in expression patterns, which can reflect the functional relationships between genes.

A number of previous studies have used visualization techniques in the study of gene expression, but they are generally still applied individually. [6]utilizing PCA and clustering methods through GeneCloudOmics to analyze microarray and RNA-seq data. Meanwhile, [7]explores spatial transcriptomics using FISH and RNA-seq methods to describe the distribution of genes in tissues, but does not yet cover temporal aspects. [8]integrating PCA and MPSO in the process of classifying microarray data using SVM, with an accuracy of 88%, although it has not yet focused on time-based visualization. [9]combined PCA and artificial neural network (ANN) for cancer classification and obtained 90.02% accuracy, but without an exploratory visualization approach. On the other hand, [10]used Cytoscape to analyze gene expression during dengue infection, and successfully identified key genes, but without combining various visualization methods in an integrated manner.

This study utilizes a time-series-based gene expression dataset consisting of 4381 genes and 24 observation time points. In this study, four visualization techniques were applied in an integrated manner, namely heatmap to see correlations between genes, PCA to reduce dimensions and identify key variability patterns, volcano plots to show significant differences between conditions based on log2 fold change and p-value, and dendrograms to group genes with similar expressions. PCA analysis revealed that the two main components (PC1 and PC2) accounted for 42.32% of the total data variation. The volcano plot showed significantly different genes with a log2 fold change > 1 and a p-value < 0.05. The heatmap displays a strong correlation of gene expression, while the dendrogram manages to form clusters of genes with consistent expression similarity over time.

This approach is expected to provide a more comprehensive understanding of gene expression dynamics and support the development of more effective biological visualization techniques. This study aims to fill a gap in the literature by integrating various visualization methods for comprehensive analysis of temporal gene expression.

2. METHOD

2.1. Approaches and Types of Research

This study uses a quantitative approach with an exploratory method, which aims to evaluate and analyze gene expression patterns visually based on time-based gene expression data. This method was chosen because it is suitable for describing relationships between genes in visual forms that facilitate interpretation, such as heatmaps, PCAs, volcano plots, and dendrograms. Thus, this study is descriptive-analytical, because it describes biological phenomena based on secondary data that is processed computationally.

2.2. Sources and Data Subjects

The dataset used in this study is from the Kaggle website titled "Gene Expression Bioinformatics Dataset" which can be accessed via the link:

e-ISSN 2830-3954
p-ISSN 2830-6031

https://www.kaggle.com/datasets/samira1992/gene-expression-bioinformatics-dataset. This dataset consists of 4,381 rows and 24 columns. Each row represents a unique gene (e.g. YAL001C, YAL014C), while subsequent columns contain the value of gene expression at different points in time (such as 40, 50, 60, up to 260 minutes).

The data is continuous and has been normalized, characterized by a near-zero average value and a small standard deviation. A positive value indicates an increase in gene expression, while a negative value indicates a decrease in expression. This dataset is particularly relevant for time-based gene expression analysis because it reflects the dynamics of changes in gene expression levels over a given time period.

2.3. Research Procedure

This research procedure consists of several main stages that are carried out sequentially to obtain valid and interpretable analysis results. The stages are explained as follows.

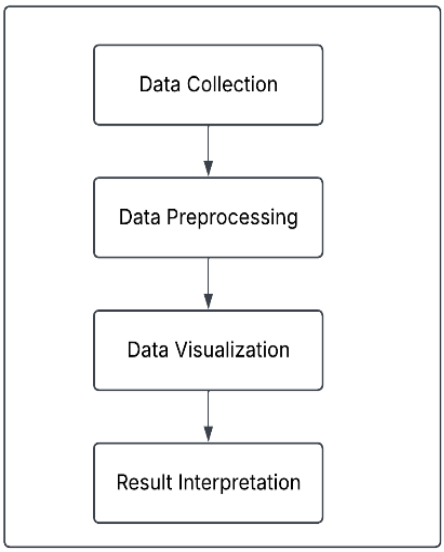


Figure 1. Research Block Diagram

2.4. Instruments and Tools

This research was conducted using Google Colab with the Python 3.10 programming language. The libraries used include Pandas and NumPy for data processing, Matplotlib and Seaborn for visualization, Scikit-learn for PCA and normalization, and SciPy for statistical testing and clustering.

2.5. Data Analysis Techniques

The data analysis technique is carried out by applying various exploratory visualization methods to analyze the relationship and difference in expression patterns between genes. The flow of the analysis technique is described in detail through the flowchart below.

e-ISSN 2830-3954
p-ISSN 2830-6031

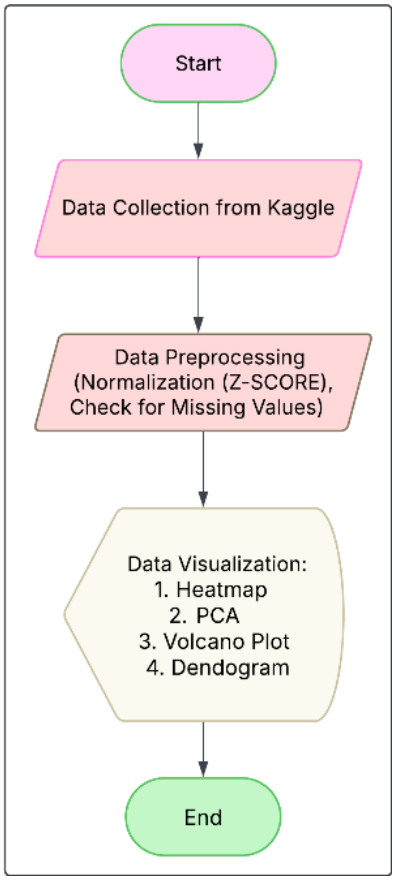


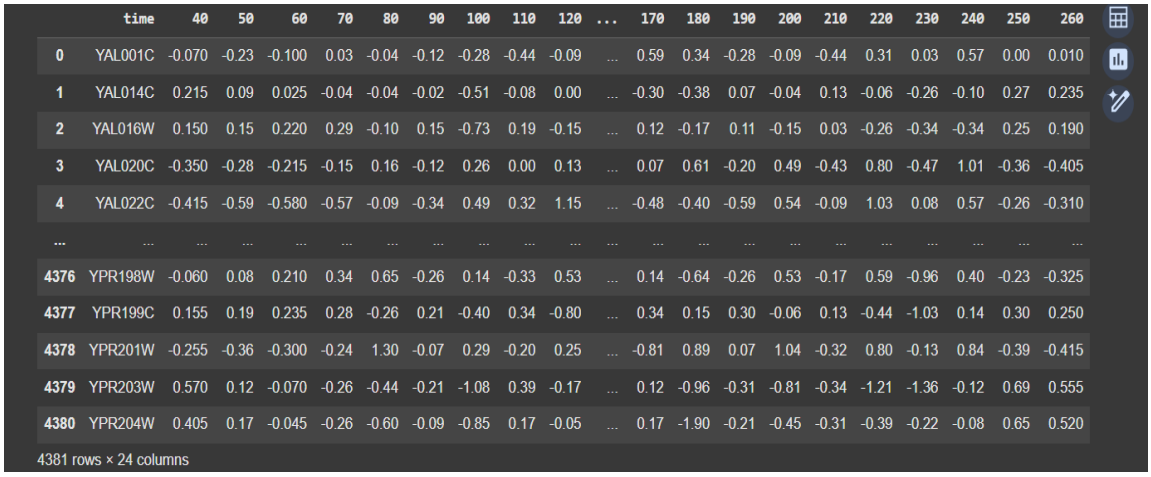
Figure 2. Research Flowchart

3. RESULTS AND DISCUSSION

3.1. Dataset Explanation

The dataset used in this study is time-based gene expression data sourced from the Kaggle website, titled "Gene Expression Bioinformatics Dataset" (https://www.kaggle.com/datasets/samira1992/gene-expression-bioinformatics-dataset). This dataset consists of 4381 rows and 24 columns. Each row represents different genes such as YAL001C and YAL014C, while the columns reflect the level of gene expression at a specific point in time (e.g. 40, 50, 60, up to 260 minutes). A positive value indicates a high level of expression, while a negative value indicates low expression. This dataset has been normalized and contains no empty values, so it is ready for further analysis.

e-ISSN 2830-3954
p-ISSN 2830-6031



	time	40	50	60	70	80	90	100	110	120	...	170	180	190	200	210	220	230	240	250	260
0	YAL001C	-0.070	-0.23	-0.100	0.03	-0.04	-0.12	-0.28	-0.44	-0.09	...	0.59	0.34	-0.28	-0.09	-0.44	0.31	0.03	0.57	0.00	0.010
1	YAL014C	0.215	0.09	0.025	-0.04	-0.04	-0.02	-0.51	-0.08	0.00	...	-0.30	-0.38	0.07	-0.04	0.13	-0.06	-0.26	-0.10	0.27	0.235
2	YAL016W	0.150	0.15	0.220	0.29	-0.10	0.15	-0.73	0.19	-0.15	...	0.12	-0.17	0.11	-0.15	0.03	-0.26	-0.34	-0.34	0.25	0.190
3	YAL020C	-0.350	-0.28	-0.215	-0.15	0.16	-0.12	0.26	0.00	0.13	...	0.07	0.61	-0.20	0.49	-0.43	0.80	-0.47	1.01	-0.36	-0.405
4	YAL022C	-0.415	-0.59	-0.580	-0.57	-0.09	-0.34	0.49	0.32	1.15	...	-0.48	-0.40	-0.59	0.54	-0.09	1.03	0.08	0.57	-0.26	-0.310
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4376	YPR198W	-0.060	0.08	0.210	0.34	0.65	-0.26	0.14	-0.33	0.53	...	0.14	-0.64	-0.26	0.53	-0.17	0.59	-0.96	0.40	-0.23	-0.325
4377	YPR199C	0.155	0.19	0.235	0.28	-0.26	0.21	-0.40	0.34	-0.80	...	0.34	0.15	0.30	-0.06	0.13	-0.44	-1.03	0.14	0.30	0.250
4378	YPR201W	-0.255	-0.36	-0.300	-0.24	1.30	-0.07	0.29	-0.20	0.25	...	-0.81	0.89	0.07	1.04	-0.32	0.80	-0.13	0.84	-0.39	-0.415
4379	YPR203W	0.570	0.12	-0.070	-0.26	-0.44	-0.21	-1.08	0.39	-0.17	...	0.12	-0.96	-0.31	-0.81	-0.34	-1.21	-1.36	-0.12	0.69	0.555
4380	YPR204W	0.405	0.17	-0.045	-0.26	-0.60	-0.09	-0.85	0.17	-0.05	...	0.17	-1.90	-0.21	-0.45	-0.31	-0.39	-0.22	-0.08	0.65	0.520

4381 rows × 24 columns

Figure 3. Dataset gen expression.csv

3.2. Data Preprocessing

The initial stage of analysis begins with data preprocessing. The time column is used as an index, and all the values of the gene expression are converted to a numeric type (float). Data containing missing values is removed to maintain the quality of the analysis. Furthermore, normalization is carried out using the Z-score method to equalize the scale between genes. The formula used:

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

Where x is the value of the expression, as the average, and as the standard deviation. This process generates data that is ready for further analysis such as PCA, heatmap and dendogram.$\mu\sigma$

3.3. Correlation Analysis Between Genes with Heatmap

Once the data is normalized, the next step is to analyze the relationships between genes using the Pearson correlation matrix. This correlation measures the strength and direction of the linear relationship between two genes. The Pearson correlation formula used is:

$$r_{xy} = \frac{\Sigma\,(X_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\Sigma\,(X_i - \bar{x})^2}\,\sqrt{\Sigma\,(y_i - \bar{y})^2}} \quad (2)$$

Where:

x_i and = the expression value of the -I gene of two different genesy_i

$\bar{x}$ and = the average expression value of each gene$\bar{y}$

r_{xy} = Pearson correlation coefficient

The correlation results were visualized using a heatmap, the heatmap in Figure 4.2 shows that some of the observed times had a strong positive correlation (red), indicating similar gene expression patterns between these times. On the other hand, the color blue shows a negative correlation, which indicates the presence of opposite expression patterns. This visualization helps identify interrelated times in gene expression.

e-ISSN 2830-3954
p-ISSN 2830-6031

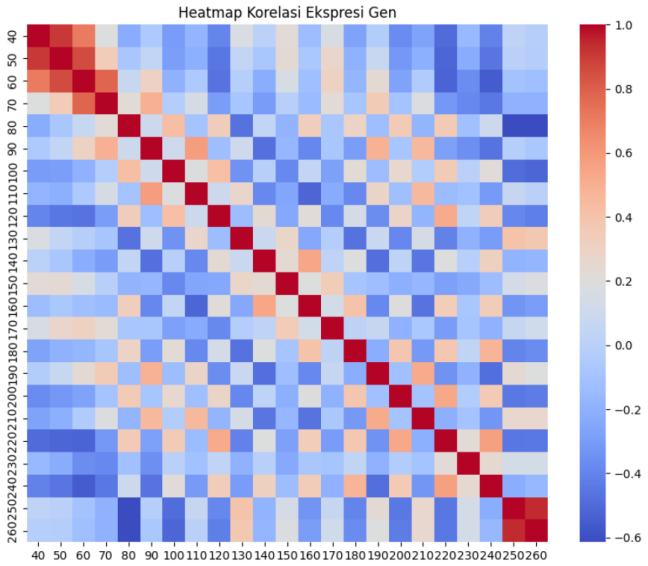


Figure 4. Visualization of Gene Expression Correlation with Heatmap

3.4. Dimension Reduction with PCA (Principal Component Analysis)

Once the relationships between genes are visualized via heatmap, the analysis is followed by dimension reduction using Principal Component Analysis (PCA). This technique aims to simplify the complexity of gene expression data by encapsulating data variations into several key components without losing important information. In this study, two main components (PC1 and PC2) were used to represent the largest variation in the dataset, thus facilitating the visualization and interpretation of the overall gene expression pattern. Mathematically, the transformation of PCA can be explained as follows:

$$X_{PCA} = X \cdot W \tag{3}$$

Where:

X = normalized data matrix

W = Eigenvector matrix of the covariance matrix

X_{PCA} = the result of transforming data into a key component space

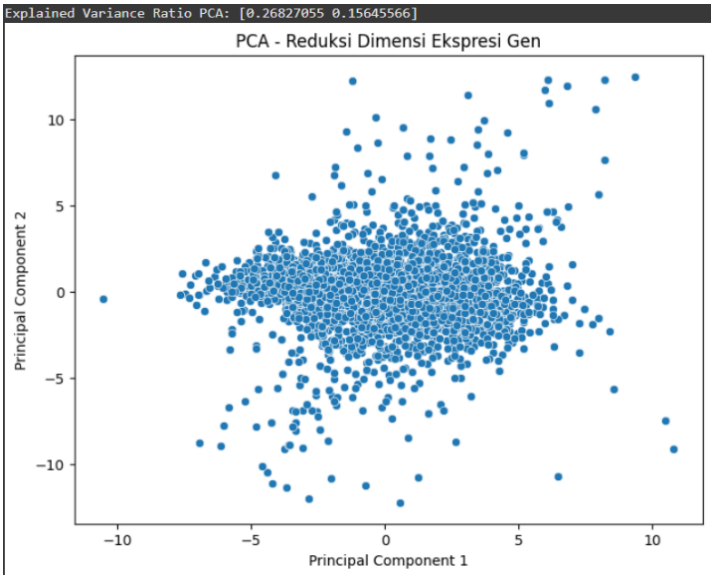


Figure 5. Visualization of PCA Gene Expression Dimension Reduction

Figure 4, above, the PCA results show the distribution of data in two main components that explain most of the dataset variation (26.87% for PC1 and 15.45% for PC2). The scatter plot of the PCA describes how the data is distributed in a two-dimensional space, with specific patterns indicating the presence of clusters or clusters of genes. PCA helps simplify complex datasets and allows the identification of genes or groups of genes that have the greatest contribution to data variation.

3.5. Analysis of Gene Expression Plot Volcano

After dimension reduction using PCA, differential analysis of gene expression is performed to identify genes that have significant changes between two conditions or groups (e.g., case and control). One of the visualization methods used is Volcano Plot. Here is the mathematical formula:

1. Log2 Fold Change ($\log_2$FC): Used to measure changes in gene expression levels between two conditions.

$$log_2 FC = log_2 \left(\frac{\bar{X}\ Case}{\bar{X}\ Control} \right) \quad (4)$$

2. Significance Test (p-value): Conducted using an independent t-test (t-test) to compare two groups.

$$t = \frac{\bar{x}_{1-}\bar{x}_2}{\sqrt{\frac{s_2^1}{n_1} + \frac{s_2^2}{n_2}}} \quad (5)$$

3. Transformation of p-value to logarithmic scale:

$$-log_{10}(\text{p-value}) \quad (6)$$

e-ISSN 2830-3954
p-ISSN 2830-6031

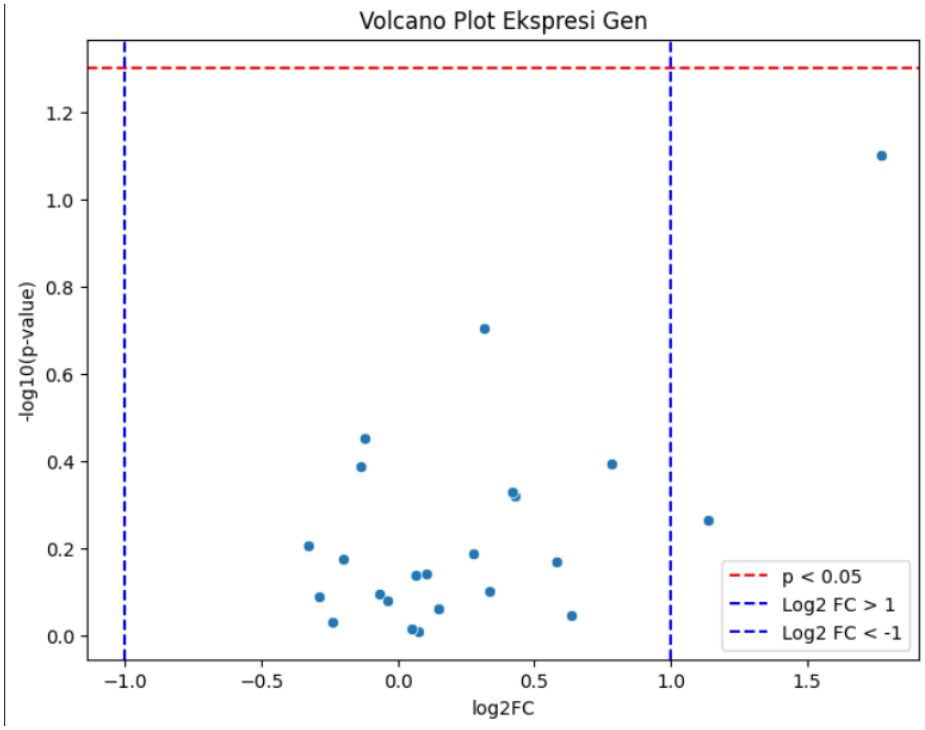


Figure 6. Visualization with volcano plot

Figure 4.4 shows the results of the volcano plot visualization showing significant genes with log2FC > 1 or < -1 and p-value < 0.05. The dots on the scatter plot mark genes that undergo noticeable expression changes, both increasing and decreasing, making it easier to identify biomarkers or molecular targets for further study.

3.6. Gene Clustering with Dendogram

After the gene differentiation is analyzed, the next process is to cluster the genes based on similar expression patterns. The technique used is hierarchical clustering with the Ward linkage method, which aims to group genes into clusters that have high similarity in their expression patterns over time. This method uses the Euclidean distance matrix as the basis for measuring the similarity between genes. The distance formula is:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_{i-y_i})^2} \quad (7)$$

e-ISSN 2830-3954
p-ISSN 2830-6031

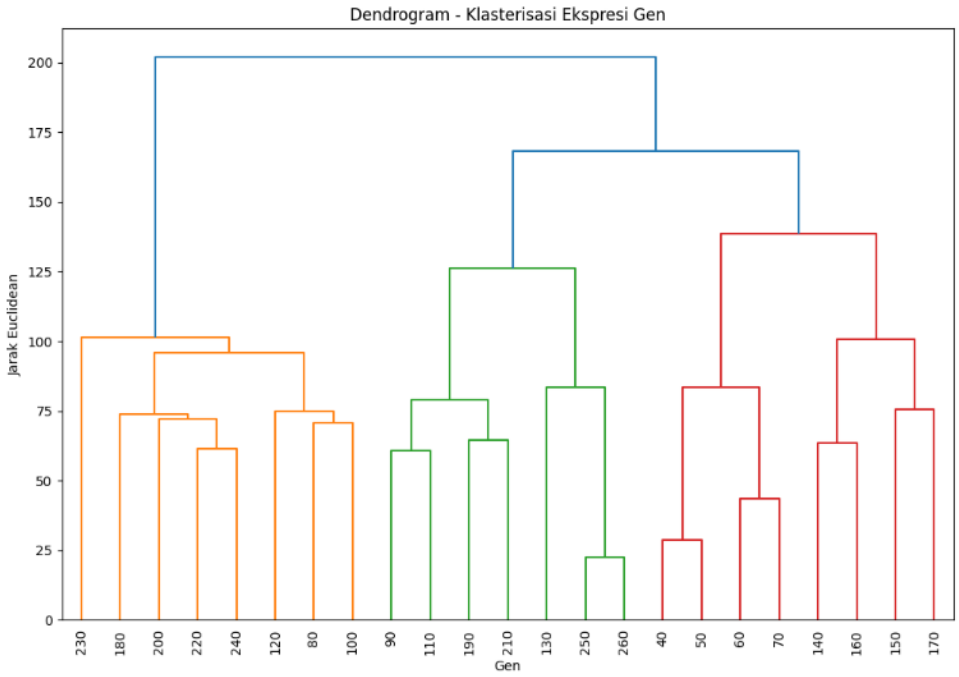


Figure 7. Visualization of clustering with dendograms

Dendrograms are used to illustrate the results of hierarchical clustering based on similarity in gene expression patterns. Adjacent branches show genes with high similarity. This visualization helps identify groups of genes that have the potential to have similar or regulated biological functions simultaneously. Figure 4.5 shows the results of clustering, where genes with similar time expression patterns are incorporated into the same branch.

With this approach, visualization is proving to be an effective early exploratory tool in bioinformatics, particularly for the identification of potential biomarkers. However, limitations such as data limited to one type of biological condition and the absence of experimental validation are concerns. Future research can expand the method with machine learning integration for predictive validation.

CONCLUSION

This study designed and implemented an integrated visualization workflow for analyzing time-series gene expression data. The workflow combines heatmaps, PCA, volcano plots, and dendrograms to support exploratory analysis from correlation identification, dimensionality reduction, differential expression detection, to gene clustering. The results show that the integration of these techniques provides a clearer representation of temporal gene expression patterns. PCA explained 42.32% of the data variation through the first two principal components, while volcano plot analysis identified differentially expressed genes using log2 fold change > 1 and p-value < 0.05. Heatmap visualization highlighted correlation patterns among genes, and dendrogram analysis grouped genes with similar expression profiles. The main contribution of this study lies in demonstrating how multiple visualization techniques can be organized into a systematic workflow to improve the interpretation of large-scale temporal gene expression data. This approach strengthens exploratory bioinformatics analysis by simplifying complex gene

expression patterns and providing a basis for further biological interpretation, biomarker investigation, and future validation studies. Rather than directly establishing clinical outcomes, the proposed workflow offers an analytical foundation for subsequent research in omics-based biomedical studies.

REFRENCES

[1] H. Biran, T. Hashimshony, T. Lahav, O. Efrat, Y. Mandel-Gutfreund, and Z. Yakhini, "Detecting significant expression patterns in single-cell and spatial transcriptomics with a flexible computational approach," *Sci. Rep.*, vol. 14, no. 1, p. 26121, 2024.

[2] R. Mitra and A. L. MacLean, "RVAgene: generative modeling of gene expression time series data," *Bioinformatics*, vol. 37, no. 19, pp. 3252–3262, 2021.

[3] T. Zhao and Z. Wang, "GraphBio: A shiny web app to easily perform popular visualization analysis for omics data," *Front. Genet.*, vol. 13, p. 957317, 2022.

[4] P. Blumenkamp, M. Pfister, S. Diedrich, K. Brinkrolf, S. Jaenicke, and A. Goesmann, "Curare and GenExVis: a versatile toolkit for analyzing and visualizing RNA-Seq data," *BMC Bioinformatics*, vol. 25, no. 1, p. 138, 2024.

[5] A. Razzaque and A. Badholia, "PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification," *Meas. sensors*, vol. 31, p. 100945, 2024.

[6] M. Helmy, R. Agrawal, J. Ali, M. Soudy, T. T. Bui, and K. Selvarajoo, "GeneCloudOmics: a data analytic cloud platform for high-throughput gene expression analysis," *Front. Bioinforma.*, vol. 1, p. 693836, 2021.

[7] B. Liu, Y. Li, and L. Zhang, "Analysis and visualization of spatial transcriptomic data," *Front. Genet.*, vol. 12, p. 785290, 2022.

[8] E. Elhaik, "Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated," *Sci. Rep.*, vol. 12, no. 1, p. 14683, 2022.

[9] S. Gupta, M. K. Gupta, M. Shabaz, and A. Sharma, "Deep learning techniques for cancer classification using microarray gene expression data," *Front. Physiol.*, vol. Volume 13-2022, 2022, [Online]. Available: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2022.952709

[10] J. V. N. Josyula, P. Talari, A. K. B. Pillai, and S. R. Mutheneni, "Analysis of gene expression profile for identification of novel gene signatures during dengue infection," *Infect. Med.*, vol. 2, no. 1, pp. 19–30, 2023.