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

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

Rara Fazira[1], Dimas Yudistira[2], Kelvyn Rosan[3]

[1,2]Universitas Islam Negeri Sumatera Utara; rarafazira27@gmail.com, yudistira@uinsu.ac.id
[3]Universitas Pembangunan Panca Budi; kelvynrosanbintang2003@gmail.com

ABSTRACT

This study aims to explore data visualization techniques in bioinformatics to analyze time-based gene expression patterns. The research seeks to answer how different visualization approaches can improve the interpretation of large-scale temporal gene expression data. A time-series gene expression dataset consisting of 4381 genes across 24 time intervals was used. The methods applied include heatmaps to identify gene correlations, Principal Component Analysis (PCA) for dimensionality reduction, volcano plots to detect significant expression changes between conditions, and dendrograms to classify genes into functional clusters. The PCA results revealed that two principal components (PC1 and PC2) accounted for 42.32% of the total variance. Volcano plot analysis identified differentially expressed genes with log2 fold change > 1 and p-value < 0.05, while the dendrogram visualization revealed several major gene clusters with similar expression behaviors. These findings demonstrate that combining multiple visualization methods provides comprehensive insights into temporal gene expression dynamics. The application of these methods offers a solution to the challenges of interpreting complex biological data by simplifying correlation patterns, identifying candidate biomarkers, and supporting the development of personalized therapeutic strategies. This research confirms the value of visual analytics in bioinformatics and recommends the integration of these tools in future large-scale omics studies.

Keywords : *Bioinformatics, Visualization, Gene Expression, Heatmap, PCA, Dendrogram, Volcano Plot;*

Corresponding Author:
Author Name: Rara Fazira
Affiliation: Universitas Islam Negeri Sumatera Utara
Email: rarafazira27@gmail.com

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 (Biran et al., 2024). 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 (Maclean, 2021).

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 (Zhao et al., 2022). 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 (Blumenkamp et al., 2024). Volcano plots are used to display significant differences in gene expression between two conditions based on log2 fold change and p-value (Razzaque et al., 2024). 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. (Helmy et al., 2021) utilizing PCA and clustering methods through GeneCloudOmics to analyze microarray and RNA-seq data. Meanwhile, (Liu at al., 2022) explores spatial transcriptomics using FISH and RNA-seq methods to describe the distribution of genes in tissues, but does not yet cover temporal aspects. (Razzaque et al., 2024) 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. (Ihsani et al., 2020) 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, (Josyula et al., 2023) 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. LITERATURE REVIEW

2.1. Bioinformatics and Gene Expression

Bioinformatics is a multidisciplinary field that combines biology, computers, mathematics, and statistics to manage and analyze large-scale biological data, particularly those related to DNA and proteins. As technology advances, genomic and proteomics data collection can be done more systematically and quickly (Du et al., 2023). Through a computational approach, bioinformatics helps researchers in uncovering DNA sequences, detecting genetic mutations, and understanding the role of gene and protein structure and function. Proteins, as the end result of gene expression, have an important role in various cellular processes and are a key target in medical therapy research (Baharuddin, 2025).

One of the important aspects of bioinformatics is the analysis of gene expression, which is the process of converting genetic information from the base sequence of DNA or RNA into functional proteins. The expression of this gene greatly determines the biological characteristics of an organism (Koh et al., 2023). This activity occurs through metabolic reactions in cells mediated by enzymes as catalysts, accelerating and directing biochemical reactions to run efficiently. By analyzing gene expression, researchers can understand gene regulation patterns, cell development mechanisms, and responses to stress or disease (Geovana, 2020).

2.1. Data Visualization

Data visualization is a technique to present information in graphic form that aims to facilitate understanding and support the data-based decision-making process. By presenting data in visual form, such as graphs or diagrams, hidden patterns, relationships between variables, and trends in raw data from various sources can be revealed more clearly (Alfia et al., 2022). In the field of bioinformatics, visualization plays an important role in handling the complexity of biological data such as gene expression, which is difficult to analyze manually in numerical form. The application of visual methods such as heatmaps, Principal Component Analysis (PCA), and dendrograms allows researchers to explore gene clusters, analyze correlations between genes, and effectively understand data variations (Kleverov et al., 2024).

2.2. Heatmap

Heatmap is a matrix visualization method that uses color variations to represent the difference in value or intensity of data in each cell. This technique is very effective in handling large-scale and complex data with many categories, as the different colors make it easier for users to quickly recognize patterns and trends (Sulianta, 2024). In the field of bioinformatics, heatmaps are often used to describe the level of gene expression in various biological conditions, such as normal and sick conditions. Colors like red usually indicate high levels of gene expression, while blue indicates low expression. Through this approach, researchers can easily group genes that have similar expression patterns, thereby accelerating the process of comprehensive genetic data analysis and interpretation.

2.3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a multivariate analysis method used to reduce the dimensions of data while maintaining the covariance structure between variables. By converting the original variables into principal components, PCA helps simplify the data so that relationships between genes or between biological conditions can be visualized in low-dimensional spaces (Elhaik, 2022). These key components are calculated in such a way as to explain as many variances in the data as possible, with the first component (PC1) explaining the largest variance, followed by the second component (PC2), and so on. PCA is an important tool in the initial exploration of data, the identification of groups of genes that have similar expressions, as well as as a basis for advanced analyses such as clustering or anomaly detection (Ihsani et al., 2020)

2.4. Volcano Plot

A volcano plot is a type of two-dimensional scatter plot commonly used in differential expression analysis, mapping the change in gene expression $\log_2$ fold change on the x-axis against the level of statistical significance ($-\log_{10}$ p-value or t-statistic) on the y-axis. The horizontal axis depicts a $\log_2$ fold change, where the genes on the right represent up-regulated and the left represent down-regulated, while the vertical axis $-\log_{10}$ p-value ensures that genes with high significance appear at the top of the plot. With this combination, volcano plots make it easy to quickly identify biologically relevant candidate genes (Goedhart, 2020).

2.5. Dendogram

Dendrograms are graphical representations that describe the process of grouping objects based on their similarities. The visualization comes from the results of hierarchical cluster analysis and is generally arranged in the form of a tree diagram. The dendrogram is constructed using a distance matrix measuring n × n to measure the degree of similarity between objects (Fan et al., 2024). The initial stage of forming a dendrogram usually begins by sorting the objects based on the highest level of similarity so that the most similar objects will be grouped first. Furthermore, the groups formed will be gradually combined with other groups that have a fairly high level of similarity, until an overall hierarchical structure is formed (Muflihan et al., 2022).

3. METHOD

3.1. 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.

3.2. 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: 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.

3.3. 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.

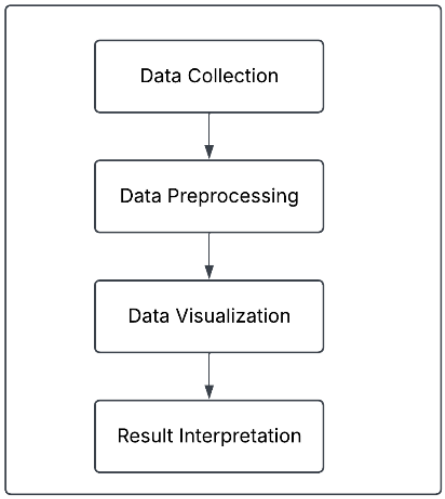

Figure 1. Research Block Diagram

3.4. 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.

3.5. 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.

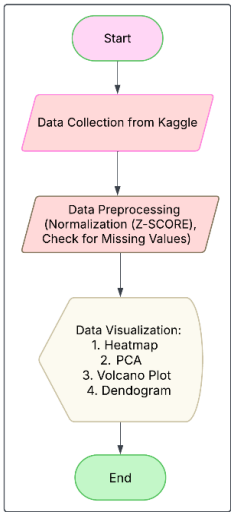

Figure 2. Research Flowchart

4. RESULTS AND DISCUSSION

4.1 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.

	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

4.2 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:

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

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.$\mu\sigma$

4.3 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:

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

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.

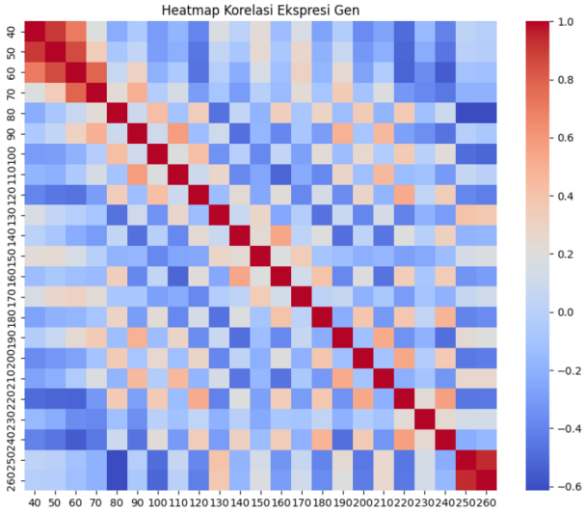

Figure 4. Visualization of Gene Expression Correlation with Heatmap

4.4 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

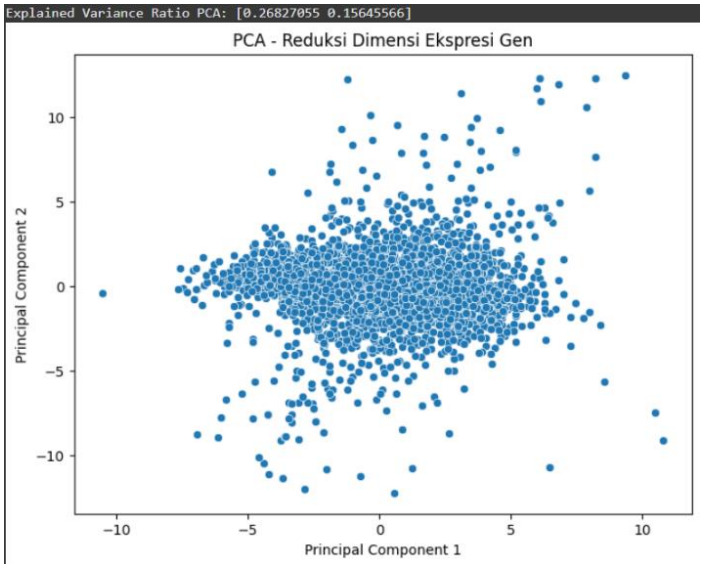

Figure 5. Visualization of PCA Gene Expression Dimension Reduction

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.

4.5 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_2FC): Used to measure changes in gene expression levels between two conditions.

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

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

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_2^1}{n_1} + \frac{s_2^2}{n_2}}} \tag{5}$$

3. Transformation of p-value to logarithmic scale:

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

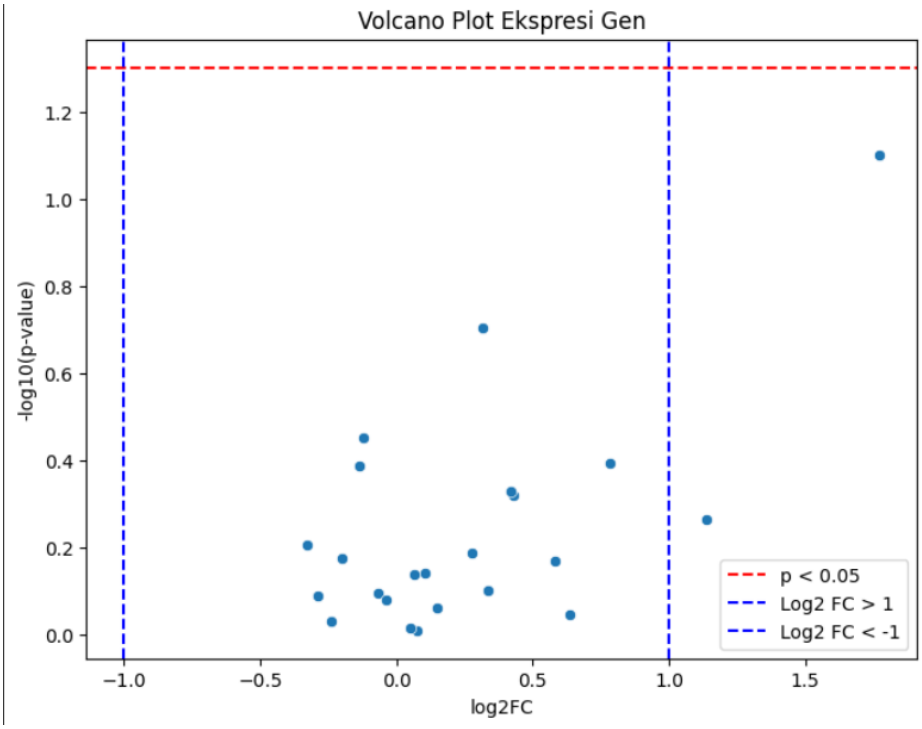

Figure 6. Visualization with volcano plot

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.

4.6 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:

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

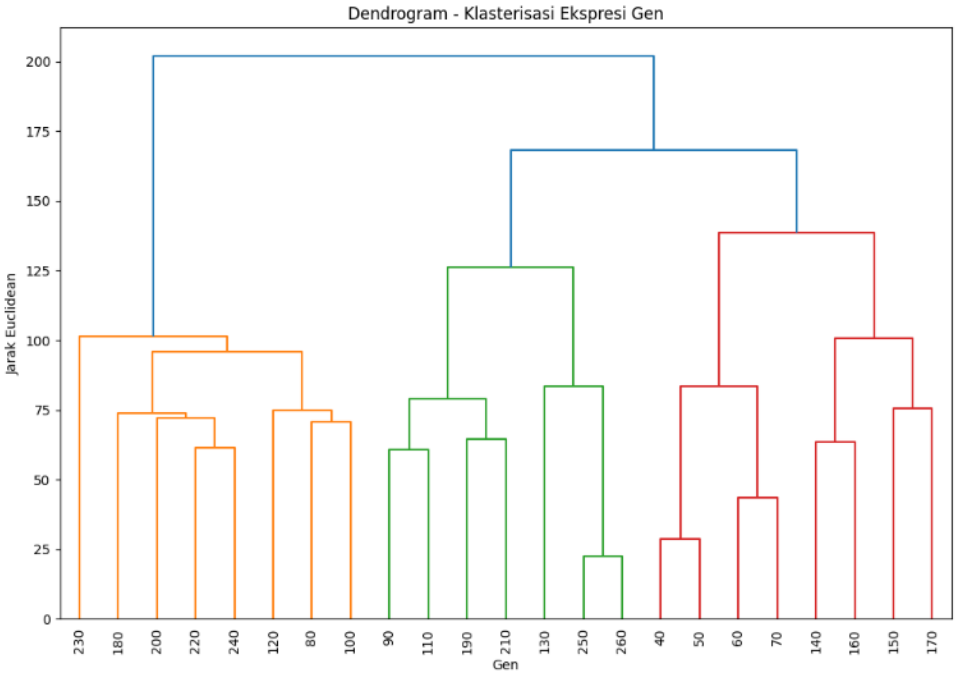

Figure 7. Visualization of clustering with dendograms

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.

5. CONCLUSION

This study designed and implemented a visualization system workflow to analyze time-based gene expression. Starting from the preprocessing of time-series data, the system involves correlation analysis using heatmaps, dimension reduction with PCA, differential analysis through volcano plots, and gene clustering with dendrograms. The main objective of the study was to explore the effectiveness of the integration of the four techniques in understanding temporal gene expression patterns. The results showed that the combination of these methods successfully described significant

variations, relationships, and differences between genes—where PCA explained 42.32% of the data variation through two main components, while the volcano plot identified significantly different genes based on the log2 fold change > 1 and p-value < 0.05. The heatmap shows a strong correlation, and the dendrogram forms clusters of similar expressions. The application of this approach strengthens the exploratory visual analysis of bioinformatics, supports the understanding of biological dynamics, and contributes to the identification of biomarkers and the development of more targeted medical therapies.

REFRENCES

Alfia, F. S., & Agussalim. (2022). Literature Review of Data Visualization and Geographic Information Systems. COMSERVA : Journal of Research and Community Service, 2(8), 1494–1500. https://doi.org/10.59141/comserva.v2i8.493

Baharuddin, M. (2025). DNA Protein Bioinformatics. Rizmedia Pustaka Indonesia.

Biran, H., Hashimshony, T., Lahav, T., Efrat, O., Mandel-Gutfreund, Y., & Yakhini, Z. (2024). Detecting significant expression patterns in single-cell and spatial transcriptomics with a flexible computational approach. Scientific Reports, 14(1), 26121. https://doi.org/10.1038/s41598-024-75314-3

Blumenkamp, P., Pfister, M., Diedrich, S., Brinkrolf, K., Jaenicke, S., & Goesmann, A. (2024). Curare and GenExVis : a versatile toolkit for analyzing and visualizing RNA - Seq data. BMC Bioinformatics, 1–12. https://doi.org/10.1186/s12859-024-05761-2

Du, J., Yang, Y. C., An, Z. J., Zhang, M. H., Fu, X. H., Huang, Z. F., … Hou, J. (2023). Advances in spatial transcriptomics and related data analysis strategies. Journal of Translational Medicine, 21(1), 1–21. https://doi.org/10.1186/s12967-023-04150-2

Elhaik, E. (2022). Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. In Scientific Reports (Vol. 12). Nature Publishing Group UK. https://doi.org/10.1038/s41598-022-14395-4

Fan, Y., Li, L., & Sun, S. (2024). Powerful and accurate detection of temporal gene expression patterns from multi-sample multi-stage single-cell transcriptomics data with TDEseq. Genome Biology, 25(1), 1–31. https://doi.org/10.1186/s13059-024-03237-3

Geovana, D. (2020). Mechanism of Gene Expression in Organisms and Enzymes That Play a Role. University of Muhammadiyah Prof.Dr.Hamka: Jakarta, (January), 6–26.

Goedhart, J. (2020). VolcaNoseR is a web app for creating , exploring , labeling and sharing volcano plots. Scientific Reports, 1–5. https://doi.org/10.1038/s41598-020-76603-3

Helmy, M., Agrawal, R., Ali, J., Soudy, M., Bui, T. T., & Selvarajoo, K. (2021). GeneCloudOmics: A Data Analytic Cloud Platform for High-Throughput Gene Expression Analysis. Frontiers in Bioinformatics, 1(November), 1–14. https://doi.org/10.3389/fbinf.2021.693836

Ihsani, D. A., Arifin, A., & Fatoni, M. H. (2020). Classification of DNA Microarray Using Principal Component Analysis (PCA) and Artificial Neural Network (ANN). ITS Engineering Journal, 9(1). https://doi.org/10.12962/j23373539.v9i1.51637

Josyula, J. V. N., Talari, P., Pillai, A. K. B., & Mutheneni, S. R. (2023). Analysis of gene expression profile for identification of novel gene signatures during dengue infection. Infectious Medicine, 2(1), 19–30. https://doi.org/10.1016/j.imj.2023.02.002

Kleverov, M., Zenkova, D., Kamenev, V., Sablina, M., Artyomov, M. N., & Sergushichev, A. A. (2024). Phantasus, a web application for visual and interactive gene expression analysis. ELife, 13, 1–34. https://doi.org/10.7554/eLife.85722

Koh, C. W. T., Ooi, J. S. G., Ong, E. Z., & Chan, K. R. (2023). STAGEs : A web - based tool that integrates data visualization and pathway enrichment analysis for gene expression studies. Scientific Reports, 1–12. https://doi.org/10.1038/s41598-023-34163-2

Liu, B., Li, Y., & Zhang, L. (2022). Analysis and Visualization of Spatial Transcriptomic Data. Frontiers in Genetics, 12(January), 1–15. https://doi.org/10.3389/fgene.2021.785290

Maclean, A. L. (2021). Gene expression RVAgene : generative modeling of gene expression time series data. 37(May), 3252–3262. https://doi.org/10.1093/bioinformatics/btab260

Muflihan, Y., Retnawati, H., & Kristian, A. (2022). Cluster analysis with a hierarchical method for the grouping of high schools based on school quality report cards in Nagan Raya Regency. Measurement in Educational Research, 2(1), 22–33.

Razzaque, A., & Badholia, D. A. (2024). PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification. Measurement: Sensors, 31(January 2023), 100945. https://doi.org/10.1016/j.measen.2023.100945

Sulianta, F. (2024). Data visualization for beginners. Sulianta Ferry.

Zhao, T., & Wang, Z. (2022). GraphBio: A shiny web app to easily perform popular visualization analysis for omics data. Frontiers in Genetics, 13(September), 1–5. https://doi.org/10.3389/fgene.2022.957317