



Available online at <http://scik.org>

Commun. Math. Biol. Neurosci. 2025, 2025:42

<https://doi.org/10.28919/cmbn/9160>

ISSN: 2052-2541

## EVALUATION OF ORDER PRESERVING TRICLUSTERING FOR 3-DIMENSIONAL GENE EXPRESSION DATA AND FUNCTIONAL INTERPRETATION USING GENE ONTOLOGY IN BREAST CANCER PATIENTS

GHEA DWI APRILIANA<sup>1</sup>, TITIN SISWANTINING<sup>1,\*</sup>, SETIA PRAMANA<sup>2</sup>, HERI KURNIA ANDIKA<sup>1</sup>

<sup>1</sup>Department of Mathematics, Universitas Indonesia, Depok, Indonesia

<sup>2</sup>Department of Statistical Computation, Politeknik Statistika STIS, Jakarta, Indonesia

Copyright © 2025 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Abstract:** Breast cancer accounts for approximately 30% of cancer-related deaths among women globally. Recent advancements in data science have enabled in-depth analysis of various diseases, including breast cancer, through the field of bioinformatics. This study aims to analyze gene expression data from breast cancer patients using the OPTricluster method and to evaluate the biological interpretation of the results through Gene Ontology analysis. To reduce the complexity of breast cancer data, gene filtering techniques are applied. Triclustering, a bioinformatics approach, is particularly effective for processing three-dimensional gene expression data. One such method, Order-Preserving Triclustering (OPTricluster), classifies genes based on similar expression orders across patients over multiple time points. In this research, OPTricluster was employed across various scenarios, utilizing gene filtering simulations with  $\delta$  parameters in comparison to the TD Score. The optimal scenario was identified with an interquartile range (IQR)  $< 0.75$  and  $\delta = 1.1$ . The OPTricluster analysis identified a total of 68 triclusters, consisting of 7 constant triclusters, 46 conserved triclusters, and 15 divergent triclusters. Functional analysis revealed that constant and divergent patterns were associated with protein transport within the Biological Process category. Conserved patterns were linked to apoptotic processes. Furthermore, the Cellular Component analysis highlighted cytosol involvement,

---

\*Corresponding author

E-mail address: [titin@sci.ui.ac.id](mailto:titin@sci.ui.ac.id)

Received January 26, 2025

while the Molecular Function category consistently identified protein binding across all patterns.

**Keywords:** bioinformatics; gene filtering; triclustering approaches; gene expression analysis; tricluster diffusion.

**2020 AMS Subject Classification:** 62P10.

## 1. INTRODUCTION

Breast cancer is the most common type of cancer in Indonesia and is the leading contributor among all cancer types. In addition to its high mortality rate, delayed treatment for cancer patients results in increased costs. According to global cancer statistics, breast cancer accounts for a significant proportion of new cancer cases and deaths annually [1] [2]. Between 2019 and 2020, the Social Health Insurance Administration Body (BPJS) in Indonesia spent over IDR 7.6 trillion on cancer treatment [3]. Furthermore, in 2020, there were 68,858 new breast cancer cases out of 396,914 new cancer cases in Indonesia. This high incidence necessitates technological advancements to facilitate the diagnosis and treatment of breast cancer. Despite advancements in early detection and treatment, the complexity and heterogeneity of breast cancer remain significant challenges [4]. One of the existing technologies used for this purpose is DNA microarrays.

In the field of molecular biology, DNA microarray technology is used to simultaneously monitor gene expression [5]. This technology generates genetic information by employing high-density arrays of DNA or oligonucleotide probes. These arrays are affixed to a semiconductor substrate, commonly referred to as chips [6]. DNA microarrays have a wide range of biological applications, including binding studies, gene expression profiling, and genotyping [7].

Genomics studies genome function, structure, evolution, mapping, and editing. This will enable us to understand biological phenomena such as the role the genome plays in disease [8]. Gene expression profiling has become a cornerstone of cancer research, providing insights into the molecular mechanisms underlying tumor progression and therapy resistance [9]. High-throughput techniques, such as DNA microarrays and RNA sequencing, have enabled researchers to identify gene expression patterns that distinguish cancer subtypes and predict therapeutic responses [10]. However, as the passage of time, Genome analysis generates and requires large data, so the data mining approach is required. The data mining concept used in this research is triclustering.

Triclustering is an advancement of the clustering technique. It necessitates a three-dimensional data matrix, capable of forming a group that includes observation, attribute, and context concurrently [11]. In the realm of gene expression data, the frequently utilized data concept is gene-sample-time (GST).

In 2006, Carroll et al. conducted an extensive study on the binding sites of estrogen receptors throughout the entire genome. MCF-7 cells were exposed to 100 nM estrogen for 0, 3, 6, or 12 hours. The latest probe mapping data were analyzed using the RMA Algorithm. Additionally, the differential expression levels at each time point compared to 0 hours were calculated using the Welch t statistic [12]. In 2012, Tchagang et al. introduced the OPTricluster method, which was successfully applied to four research cases, demonstrating its ability to resist noise and detect similarities and differences between biological samples [13]. In 2020, Siska et al. utilized the OPTricluster triclustering method to investigate the effects of the yellow fever vaccine. This research employed Java programming developed by Tchagang et al. [14]. In 2021, SwathyPriyadharsini and Premalatha used a hybrid cuckoo search with clonal selection to identify co-expressed genes over samples and times using a triclustering solution for breast cancer gene expression data [15]. In 2022, Apriliana et al. researched the application of OPTricluster to breast cancer gene expression data using Java programming developed by Tchagang et al.

This research extends the work of Apriliana et al. (2022) and Siska et al. (2020). There are several key differences between this research and the study by Apriliana et al. (2022). First, in this research, Gene Ontology (GO) has been fully implemented, whereas in Apriliana et al. (2022) it was only partially implemented. Second, Apriliana et al. (2022) did not utilize an evaluation method to test the tricluster results, while this research includes such a method. Third, this research developed the OPTricluster program using the Python programming language, in contrast to Apriliana et al. (2022), which used a Java program developed by Tchagang et al. (2012). Additionally, this research employs more extensive simulations compared to Apriliana et al. (2022), which used only one simulation. The differences between this research and Siska et al. (2020) are also evident in several aspects. Firstly, Siska et al. (2020) did not employ gene filtering, whereas this research uses gene filtering to select genes with more relevant information. Secondly, the simulations in this research are based on gene filtering with a different range of delta selection, ranging from 1.1 to 2.0, while Siska et al. (2020) used a delta range between 1.1 and 1.5. Thirdly, the OPTricluster program in this research was developed using the Python programming language, in contrast to Siska et al. (2020), which used a Java program developed by Tchagang et al. (2012). Therefore, we conducted a study using OPTricluster, a triclustering method, to identify similar patterns in gene expression data relevant to breast cancer patients. The data, obtained on March 2, 2022, is available from the National Center for Biotechnology Information (NCBI) website: <https://www.ncbi.nlm.nih.gov/>. This dataset focuses on MCF-7 cells, which were stimulated with

100 nm estrogen and observed at the 0th, 3rd, 6th, and 12th hours to monitor the effects of estrogen in 3 patients. MCF-7 cells, a widely used breast cancer cell line, offer more relevant patient care data compared to other breast cancer cell lines [15]. The goal of this research is to examine the changes in breast cancer patients following estrogen treatment. OPTricluster is particularly suitable because it can handle 3-dimensional gene expression data, which includes gene, time, and sample dimensions.

## 2. MATERIAL AND METHOD

### 2.1 Data Description

In this study, we used data originally created by Carroll et al., which was updated in 2019. This data set is valuable for gaining insights into biological processes, disease mechanisms, and potential therapeutic targets, making it an excellent resource for research and analysis in fields such as molecular biology, genetics, and medicine. The data consists of three-dimensional gene expression information from breast cancer patients, featuring a matrix with 54,675 probe IDs, 3 patients, and 4 time points (0th, 3rd, 6th, and 12th hours). It was obtained from the Gene Expression Omnibus (GEO) with the serial number GSE11324 and the GPL570 platform. This data can be accessed via <https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE11324> and was imported using RStudio tools.

The description of the data can be seen in Table 1.

**Table 1.** Description of the data

Experimental Condition	Time Point(s)
Patient-1 ( $P_1$ )	0 hour
	3 hour
	6 hour
	12 hour
Patient-2 ( $P_2$ )	0 hour
	3 hour
	6 hour
	12 hour
Patient-3 ( $P_3$ )	0 hour
	3 hour
	6 hour
	12 hour

Table 1 shows a detailed overview of the experimental conditions with the respective time points used in this study. The table classifies all data into the three groups of the three patients diagnosed with breast cancer and identified as Patient-1 ( $P_1$ ), Patient-2 ( $P_2$ ), and Patient-3 ( $P_3$ ). Each patient underwent measurements at four distinct time points: 0<sup>th</sup> hour, 3<sup>rd</sup> hour, 6<sup>th</sup> hour, and 12<sup>th</sup> hour. This framework facilitates the analysis of gene expression patterns over time, allowing investigation of temporal changes and biological processes associated with breast cancer progression. By aligning these time points across all patients, the dataset supports comparative analyzes and triclustering approaches to uncover significant biological insights.

## 2.2 Gene Filtering

Gene Filtering was a method to eliminate the ID Probe to reduce noise and selecting genes that had more relevant information. Also, programming time was shorter since the performed ID Probe gene was smaller. This research employed the inter-quartile range (IQR).

The inter-quartile range (IQR) measure of the expression values is used in the data filtering stage to identify probes with smaller expression ranges. Generally, an empirical analysis or past knowledge-based threshold  $q$  is chosen, and any probes with an  $IQR < q$  are removed. The number of genes that pass the filter steadily decreases as IQR rises [16]. In this study, our total genes are 54,675. We employed three IQR approaches for gene filtering, which are  $IQR < 0.25$ ,  $IQR < 0.50$ , and  $IQR < 0.75$ .

## 2.3 Silhouette Coefficient Method

The silhouette coefficient method is used to determine the delta value for each slice using the k-means clustering method. In this study, the utilized delta is determined based on the obtained coefficient values [17]. The interpretation of the silhouette coefficients can be seen in Table 2.

**Table 2.** Silhouette Coefficients Interpretation

Silhouette Coefficient	Interpretation
$0.7 < \delta \leq 1$	Strong structure
$0.5 < \delta \leq 0.7$	Medium structure
$0.25 < \delta \leq 0.5$	Weak structure
$\delta \leq 0.25$	No structure

In this study, we filtered 54,675 genes with  $IQR < 0.25$ ,  $IQR < 0.50$ , and  $IQR < 0.75$  respectively. According to IQR filtering, we get Silhouette Coefficient 0.5592 for  $IQR < 0.25$ , 0.5767 for  $IQR < 0.50$ , and 0.5318 for  $IQR 0.75$ . These indicates our triclusters have a medium structure.

## 2.4 Triclustering

Triclustering was a development of clustering and biclustering method. Triclustering could be used for three-dimensional data defined in Definition 1, and the definition of triclustering in Definition 2.

**Definition 1.** *Three-dimensional dataset  $A$  was defined by  $n$  observation  $X = \{x_1, \dots, x_n\}$ ,  $m$  attribute  $Y = \{y_1, \dots, y_m\}$ , and  $p$  context  $Z = \{z_1, \dots, z_p\}$  by  $a_{ijk} \in R$ ,  $a_{ijk} \in \Sigma$  ( $\Sigma$  was a set of nominal data or ordinal data), integer ( $a_{ijk} \in Z$ ), or non-identically distributed ( $a_{ijk} \in A_j$ , where  $A_j$  was the domain of  $y_j$ )*

**Definition 2.** *Given a 3D dataset  $A$ , where  $X$  represented  $n$  observations,  $Y$  was  $m$  attributes, and  $Z$  was  $l$  context. Triclustering aimed to find a tricluster set of  $B = \{I, J, K\}$  where  $I \subseteq X$ ,  $J \subseteq Y$ , and  $K \subseteq Z$ .*

The illustration of Triclustering can be seen in Figure 1.

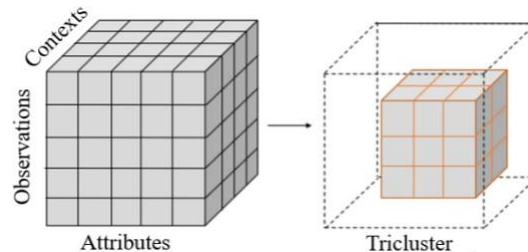


FIGURE 1. Illustration of Triclustering [13]

Triclustering identifies sub-space within a three-dimensional data structure consisting of observations, attributes, and contexts. The full structure represents the entire dataset, while the tricluster highlights a smaller, meaningful sub-space where consistent patterns emerge across these three dimensions simultaneously. This approach is particularly effective for analyzing complex datasets with interrelated elements, such as those in temporal or multi-condition studies.

In the context of bioinformatics, as applied in this research, the observations represent genes, the attributes correspond to time points, and the contexts are samples. Triclustering helps uncover patterns of gene expression that are consistent across specific time intervals and samples, making it a powerful tool for studying dynamic biological processes and sample-specific behaviors.

## 2.5 Order Preserving Triclustering (OPTricluster)

Order Preserving Triclustering (OPTricluster) is a pattern-centric approach for analyzing three-dimensional gene expression data, particularly tailored for short time-series datasets. Developed to identify genes with analogous patterns of expression alterations across various time points and experimental conditions, the OPTricluster algorithm groups genes with similar

expression dynamics and corresponding time points within each subset of experimental conditions [14]. This technique excels in pinpointing subsets of data that display consistent patterns across all three dimensions: rows, columns, and layers, thus offering a notable advantage in pattern recognition across multiple dimensions [13]. By focusing on these subsets, OPTricluster can effectively reduce the data's dimensionality while still capturing relevant patterns, which is beneficial for visualization, analysis, and understanding of complex datasets. In this study, we developed the OPTricluster program using the Python programming language. A visual representation of the OPTricluster algorithm is shown in Figure 2.

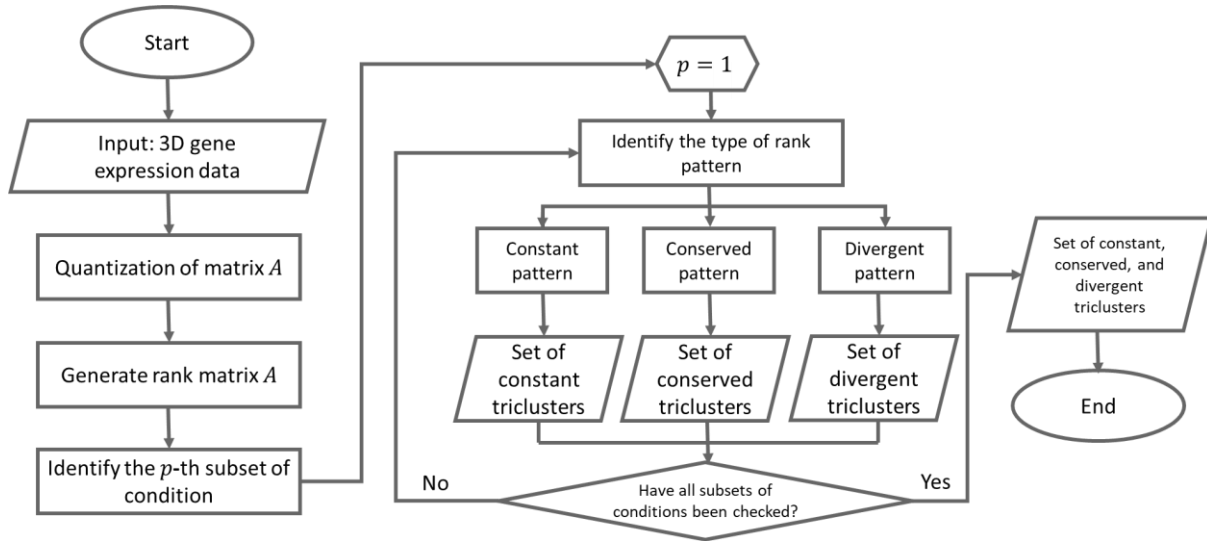


FIGURE 2. Flowchart of OPTricluster's Algorithm [9]

There are an explanation regarding the OPTricluster flowchart:

**1) Quantization:** The implementation of quantization diminishes noise and enhances the precision of triclustering outcomes [13]. OPTricluster employed quantization to partition the values of  $f_{ik}(T)$  into multiple subintervals, with the centroid of each subinterval utilized in the triclustering process. The initial step in quantization involved determining the threshold ( $\delta$ ), which plays a role in establishing the total number of subintervals for each  $f_{ik}(T)$ . The total subintervals are determined by Equation (1).

$$(1) \quad E_{ik} = \left\lceil \frac{b_E - b_0}{\delta} \right\rceil, \delta \neq 0$$

where,

$E_{ik}$ : total of sub-intervals for each  $f_{ik}(T)$ ,

$b_E$  : maximum value of  $f_{ik}(T)$ ,

$b_0$  : minimum value of  $f_{ik}(T)$ ,

$\delta$  : threshold.

The interval  $[b_0, b_E]$  was divided into several sub-intervals based on the results of  $E_{ik}$  using Equation (2).

$$(2) \quad [b_0, b_E] = [b_0, b_0 + \delta), [b_0 + \delta, b_0 + 2\delta), \dots, [b_{e-1}, b_E]$$

where  $b_e = b_0 + e\delta$  and  $e = 1, 2, \dots, E_{ik}$ .

**2) Generate Ranking Matrix:** The next step after quantization is to transform the quantization result matrix into a ranking matrix. The ranking matrix was transformed into 2D matrix defined as  $[r_i(T, C)] = [r_{ijk}]$ . The illustration of the ranking matrix of the quantization results can be seen in Figure 3.

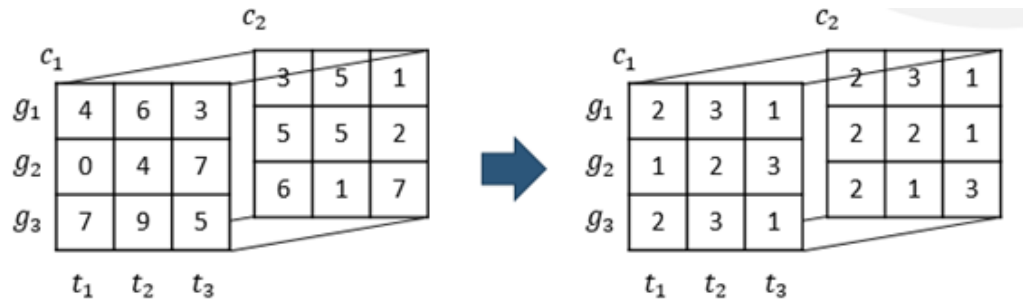


FIGURE 3. Illustration of generating ranking matrix from quantization result [10]

Figure 3 represents a ranking matrix based on the quantization results of each gene across time points and for each condition. From this matrix, the patterns of gene expression can be observed, which may increase, decrease, or fluctuate for each time point of a specific condition.

**3) Pattern Identification for Each Sub-Interval:** OPTricluster recognized rank patterns present in matrix  $R$  for every combination of experimental conditions. Experimental conditions in triclustering typically include various biological or environmental factors, such as different time points, treatments, or combinations of experimental variables. The set of all potential combinations is represented by  $\Omega$ , while the total number of combinations is denoted as  $\Gamma$ , as specified in Equation (3).

$$(3) \quad \Gamma = 2^l - 1$$

where  $l$  is a total of experimental conditions.



**4) Create the Tricluster According to the Pattern:** The final step of the OPTricluster was a grouping of each gene that had similar rank based on pattern type. Thus, the types of the pattern from Tricluster were:

1. Constant Tricluster (CO)

Constant Tricluster is a gene cluster which have expression level or rank did not change along time points in each subset of the experimental condition, such as  $f_{ik}(T) = [1 \ 1 \ 1]$ ,  $f_{ik}(T) = [0.5 \ 0.5 \ 0.5]$ , and other similar patterns. The illustration of constant tricluster can be seen in Figure 4.

Gene	$c_1$			$c_2$			$c_3$		
	$t_1$	$t_2$	$t_3$	$t_1$	$t_2$	$t_3$	$t_1$	$t_2$	$t_3$
$g_1$	1	2	3	1	1	1	1	1	1
$g_2$	1	1	2	3	2	1	1	2	3
$g_3$	1	2	3	1	1	1	1	2	3
$g_4$	3	2	1	1	1	2	1	1	1
$g_5$	1	1	2	1	1	2	1	1	1



No.	Subset of Experimental Conditions	Tricluster	Subset of Gene
1.	$\{c_2\}$	1 1 1	$g_1, g_3$
2.	$\{c_3\}$	1 1 1	$g_1, g_4, g_5$
3.	$\{c_2, c_3\}$	1 1 1	$g_1$

FIGURE 4. Illustration of constant tricluster

2. Conserved Tricluster (CP)


A Conserved Tricluster is a gene cluster found in each subset condition based on rank pattern. To determine the total number of triclusters showing the conserved pattern (CP) within each subset of experimental conditions, one can subtract the number of constant rank patterns from the total number of rank patterns identified in each subset condition. It represented by Equation (4).

$$(4) \quad CP(\Omega_p) = h_p - CO(\Omega_p)$$

where  $h_p$  represents the total number of different rank patterns in the subset of condition in the rank matrix  $r_i(\Omega_p, T)$ . Triclusters with conserved patterns signify groups of genes

that exhibit similar patterns across a subset of experimental conditions over time. An illustration of a conserved tricluster is shown in Figure 5.

Gen	$c_1$			$c_2$			$c_3$		
	$t_1$	$t_2$	$t_3$	$t_1$	$t_2$	$t_3$	$t_1$	$t_2$	$t_3$
$g_1$	1	2	3	1	1	1	1	1	1
$g_2$	1	1	2	3	2	1	1	2	3
$g_3$	1	2	3	1	1	1	1	2	3
$g_4$	3	2	1	1	1	2	1	1	1
$g_5$	1	1	2	1	1	2	1	1	1



Subset Kondisi Eksperimen	Tricluster	Subset Gen
$\{c_1\}$	1 1 2	$g_1, g_2, g_5$
	1 2 3	$g_3$
	3 2 1	$g_4$
$\{c_2\}$	3 2 1	$g_2$
	1 1 2	$g_4, g_5$
$\{c_3\}$	1 2 3	$g_2, g_3$
$\{c_1, c_2\}$	1 1 2	$g_5$
$\{c_1, c_3\}$	1 2 3	$g_3$

FIGURE 5. Illustration of conserved tricluster

### 3. Divergent Tricluster (DP)

A Divergent Tricluster (DP) represents a set of genes that consistently exhibit the same rank pattern across multiple experimental conditions but display a distinct rank pattern under a specific condition. This divergence highlights a group of genes that behave uniformly in the majority of conditions while demonstrating a unique behavior in one particular condition. Mathematically, this can be expressed as shown in Equation (5).

$$(5) \quad DP = CP\{a\} - CP\{b\} = \begin{Bmatrix} I_a - I_b \\ K_a - K_b \end{Bmatrix}$$

where,

$CP\{a\}$  : group of gene that have same rank pattern in  $(l - 1)$  experimental conditions.

$CP\{b\}$  : group of gene that have same rank pattern in  $l$  experimental conditions.

The illustration of divergent tricluster can be seen in Figure 6.

Gen	$c_1$			$c_2$			$c_3$		
	$t_1$	$t_2$	$t_3$	$t_1$	$t_2$	$t_3$	$t_1$	$t_2$	$t_3$
$g_1$	1	2	3	1	1	1	1	1	1
$g_2$	1	2	1	1	1	1	1	1	1
$g_3$	1	2	3	1	1	1	1	2	3
$g_4$	3	2	1	1	1	2	1	1	1
$g_5$	1	1	2	1	1	2	1	1	1

↓

Subset Kondisi Eksperimen	Tricluster	Subset Gen
$\{c_2, c_3\} - \{c_1, c_2, c_3\}$	1 1 1	$g_1, g_2$
$\{c_1, c_3\} - \{c_1, c_2, c_3\}$	1 2 3	$g_3$
$\{c_1, c_2\} - \{c_1, c_2, c_3\}$	1 1 2	$g_5$

FIGURE 6. Illustration of divergent tricluster

## 2.6 Evaluation

The evaluation of triclustering performance was necessarily conducted to determine the quality of the resulting triclusters. In this study, the evaluation of Tricluster Diffusion (TD) was used. The evaluation using the Tricluster Diffusion (TD) score could be done by dividing the mean square residual (MSR) by the tricluster volume [19], which can be seen in Equation (6).

$$(6) \quad TD = \frac{MSR}{|G| \times |T| \times |C|}$$

where,

$$MSR = \frac{1}{|G||T||C|} \sum (m_{gtc} - m_{gTC} - m_{GtC} - m_{GTc} + 2m_{GTC})^2,$$

$m_{gtc}$  =  $g$ -th obsevation,  $t$ -th attribute, and  $c$ -th context,

$$m_{gTC} = \frac{1}{|T||C|} \sum_{t \in T, c \in C} m_{gtc}, \quad \text{mean of } g\text{-th observation,}$$

$$m_{GtC} = \frac{1}{|G||C|} \sum_{g \in G, c \in C} m_{gtc}, \quad \text{mean of } t\text{-th attribute,}$$

$$m_{GTc} = \frac{1}{|G||T|} \sum_{g \in G, t \in T} m_{gtc}, \quad \text{mean of } c\text{-th context.}$$

## 2.7 Gene Ontology

The Gene Ontology (GO) database was created to systematically describe the functional properties of gene products across various species and to aid in the computational prediction of gene functions [20]. GO offers a framework and set of concepts for describing the functions of gene products in all organisms, specifically designed to support computational representations of biological systems. A GO annotation links the product of a specific gene to a GO concept, thus making statements about the function of that gene [21].

There are three aspects of GO [18]:

1. **Molecular Function (MF)** describes the specific activity performed by a gene or its product at the molecular level, such as binding to other proteins or catalyzing chemical reactions.
2. **Biological Process (BP)** refers to a series of biological activities performed by genes, such as cell division or programmed cell death (apoptosis).
3. **Cellular Component (CC)** indicates where the genes or their products are located within the cell, such as in the cytoplasm, nucleus, or cell membrane.

## 3. MAIN RESULTS

The number of remaining genes using the  $IQR < 0.25$ ,  $IQR < 0.50$ , and  $IQR < 0.75$  were 41,006; 27,337; and 13,669; respectively. According to this research, there were 3 patients as experimental condition. The subsets of the experimental condition can be seen in Equation (7).

$$(7) \quad \Omega = \{\{P_1\}, \{P_2\}, \{P_3\}, \{P_1, P_2\}, \{P_1, P_3\}, \{P_2, P_3\}, \{P_1, P_2, P_3\}\}$$

The OPTricluster algorithm formed tricluster based on rank patterns within the subset of the experimental condition on Equation (7), such as constant, conserved, and divergent patterns. In this research, 15 scenarios were performed to find the best tricluster for every gene filtering's scenarios, i.e.  $\delta = 1.1$ ;  $\delta = 1.2$ ;  $\delta = 1.3$ ;  $\delta = 1.4$ ;  $\delta = 1.5$ ;  $\delta = 1.6$ ;  $\delta = 1.7$ ;  $\delta = 1.8$ ;  $\delta = 1.9$ ;  $\delta = 2.0$ ;  $\delta = 2.1$ ;  $\delta = 2.2$ ;  $\delta = 2.3$ ;  $\delta = 2.4$ ;  $\delta = 2.5$ , also  $\delta$  based on Silhouette Coefficient method. We developed a Python Program to execute these scenarios. The comparison of all scenarios can be seen in Figure 7.

OPTRICLUSTER FOR 3-DIMENSIONAL GENE EXPRESSION DATA

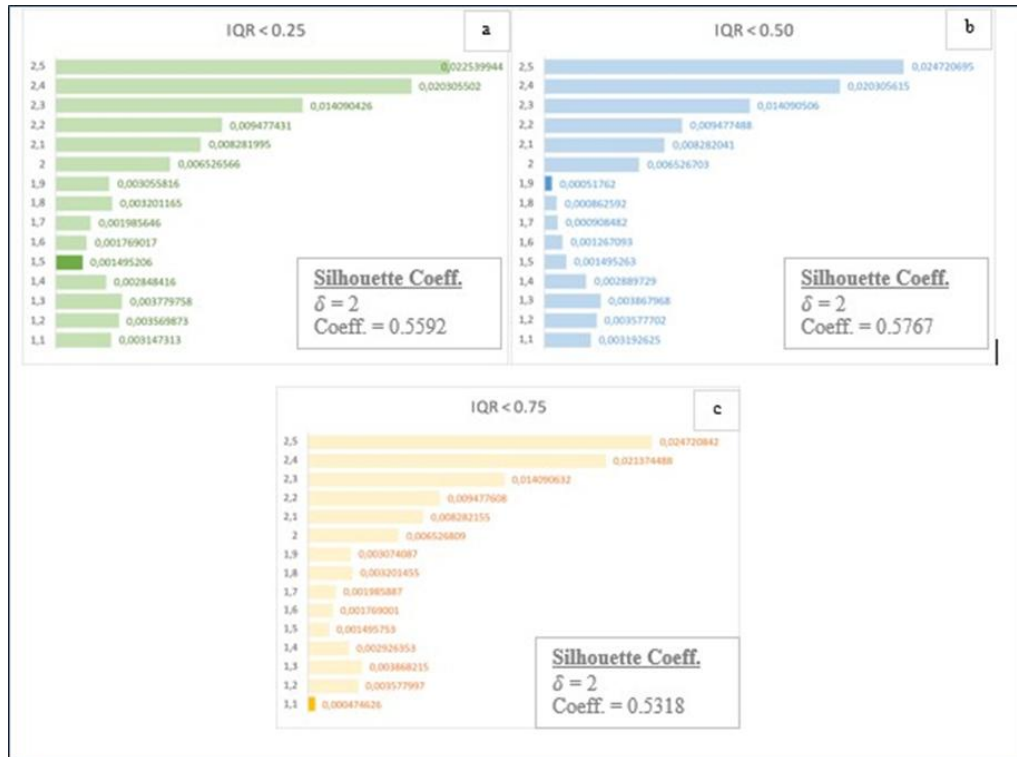


FIGURE 7. Comparison of TD Scores' Average for Each Scenarios

Figure 7 shows the smallest TD score when  $IQR < 0.25$ ,  $IQR < 0.50$ ; and  $IQR < 0.75$  are  $\delta = 1.5$ ;  $\delta = 1.9$ ; and  $\delta = 1.1$ , respectively. Furthermore, we got  $\delta = 2$  with Silhouette Coefficient 0.5592; 0.5767; and 0.5318, respectively, so that our triclusters have a medium structure when we choose  $\delta = 2$ . As we can see, the TD Score of  $\delta = 2$  is not a minimum value, so we can choose  $\delta = 1.5$  for  $IQR < 0.25$ ,  $\delta = 1.9$  for  $IQR < 0.50$ , and  $\delta = 1.1$  for  $IQR < 0.75$ . The comparison of the minimum TD score can be seen in Figure 8.

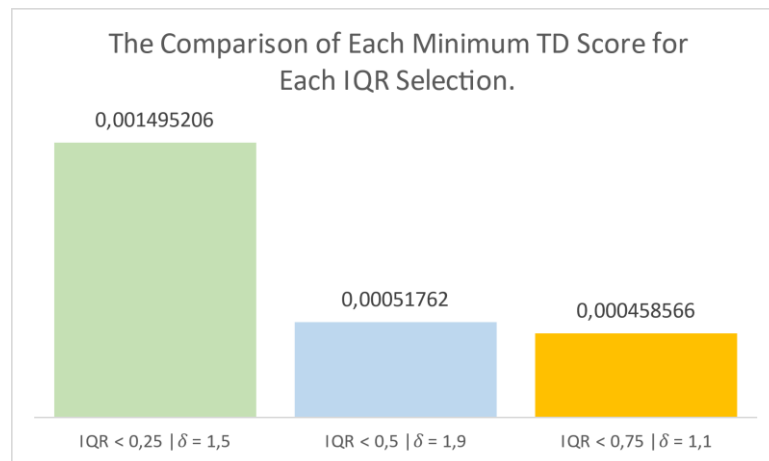


FIGURE 8. Comparison of Minimum TD Scores' for Each IQR Selection

Figure 8 shows the comparison of minimum TD Score's for each IQR Selection. We can see that the minimum TD Score's were found where  $IQR < 0.75$  and  $\delta = 1.1$ . This scenario produced 7 constant triclusters, 46 conserved triclusters, and 13 divergent triclusters, which the total was 66 triclusters.

### 3.1 Constant Tricluster Result

The evaluation results based on the constant-patterned tricluster for each subset of the experimental condition are shown in Figure 9.

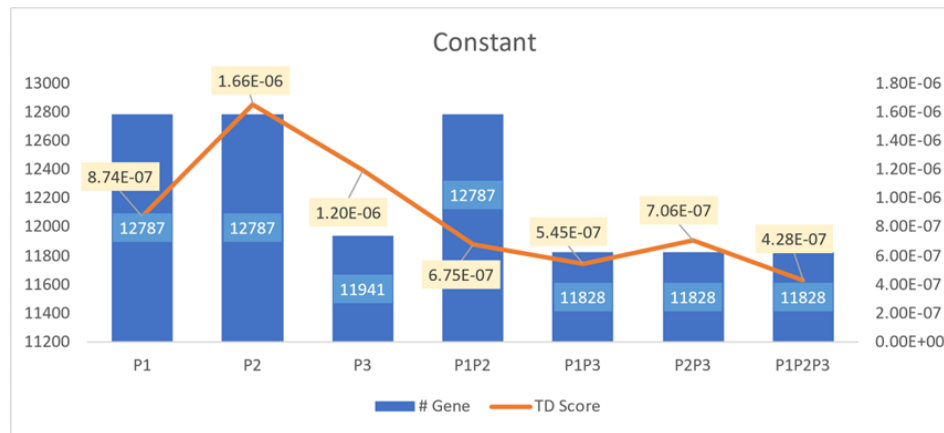


FIGURE 9. Number of genes and TD Score in constant tricluster

Figure 9 shows that the subset of  $\{P_1, P_2, P_3\}$  patient had the lowest TD Score (4.28E-07) in the constant tricluster. It means that the tricluster which formed by the subset of patients  $\{P_1, P_2, P_3\}$  has a better quality compared to tricluster in other patient subsets and comprises 11,828 genes. The gene expression level was relatively unchanged in  $P_1$ ,  $P_2$ , and  $P_3$  body until 12 hours after estrogen stimulation. So, in this case, 11,828 genes in  $P_1$ ,  $P_2$ , and  $P_3$  body is insignificantly affected by the estrogen stimulation. Figure 10 shows a heatmap of genes of the patient  $\{P_1, P_2, P_3\}$ .

IDProbe	P1_0h	P1_3h	P1_6h	P1_12h	P2_0h	P2_3h	P2_6h	P2_12h	P3_0h	P3_3h	P3_6h	P3_12h
1007_s_at	12.95	12.75	12.69	12.42	12.92	12.60	12.54	12.37	12.93	12.71	12.61	12.39
1405_i_at	5.09	4.86	4.59	4.80	5.28	4.97	4.67	4.79	5.08	4.78	4.81	4.64
1487_at	9.44	9.41	9.37	8.72	9.41	9.40	9.06	8.53	9.49	9.54	9.30	8.70
1552263_at	5.91	5.74	5.61	6.05	5.99	6.22	5.70	5.59	5.98	5.80	5.89	6.03
1552264_a_at	8.17	8.15	8.12	8.58	8.18	7.80	8.11	8.62	7.84	7.93	8.21	8.71
1552269_at	6.70	6.30	6.30	6.43	6.71	6.53	6.69	6.36	6.69	6.34	6.51	6.63
1552283_s_at	5.67	5.49	5.68	5.32	5.54	5.67	5.48	5.13	5.44	5.54	5.71	5.36
1552287_s_at	7.50	7.60	7.93	7.90	7.51	8.20	7.81	8.23	7.64	7.94	7.93	8.10
1552288_at	5.36	5.51	5.78	5.60	5.35	5.17	5.68	5.39	5.27	5.53	5.63	5.72
1552291_at	7.99	7.86	8.08	8.31	7.90	8.72	8.05	8.22	7.97	8.39	8.25	8.45
1552321_a_at	4.03	3.86	3.76	3.87	3.92	3.97	3.77	4.10	4.10	4.27	3.72	3.75
1552329_at	7.22	7.19	6.90	6.95	7.26	7.42	6.77	6.80	7.22	7.14	7.28	7.18
1552343_s_at	5.08	5.22	5.04	5.10	5.15	5.59	5.26	5.35	5.13	5.08	5.54	5.52
1552347_at	5.96	5.69	5.68	5.97	5.74	6.13	5.95	5.97	5.85	5.66	5.95	6.11
1552359_at	3.92	4.00	3.59	3.75	3.85	3.79	3.76	3.71	3.53	3.70	3.88	3.52
1552364_s_at	8.19	8.03	8.09	8.26	8.41	8.07	7.98	7.97	8.34	8.16	7.97	8.24
1552365_at	6.31	5.86	6.28	6.09	6.16	6.51	5.89	5.74	6.32	6.30	6.26	6.02
1552424_at	4.38	4.81	4.58	4.36	4.71	4.54	4.73	4.58	4.80	4.66	4.38	4.48
1552427_at	4.02	4.47	4.59	4.32	4.12	4.64	4.56	4.17	4.09	4.50	4.53	4.33

FIGURE 10. Heatmap for constant tricluster on subset of patient  $\{P_1, P_2, P_3\}$

Figure 10 presents the heatmap for the constant tricluster derived from the subset of patients  $\{P_1, P_2, P_3\}$ . The heatmap illustrates the consistent gene expression patterns across all experimental conditions and time points (0h, 3h, 6h, and 12h). The uniformity in expression levels, as indicated by similar color intensities across the matrix, confirms that the identified tricluster maintains constant gene expression regardless of the experimental condition or time. This stability highlights the robustness of the gene subset in the specified tricluster under varying experimental conditions.

### 3.2 Conserved Tricluster Result

The evaluation results based on the conserved-patterned tricluster for each subset of the experimental condition are shown in Figure 11.

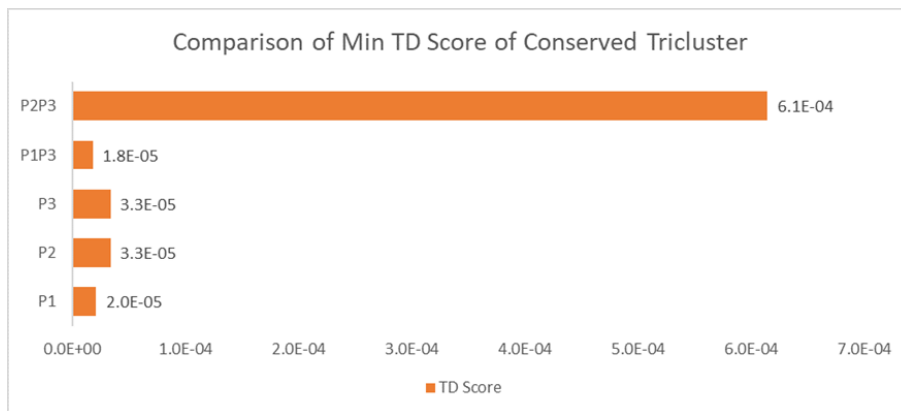


FIGURE 11. Comparison of Minimum TD Score for Each Subset of Patient of Conserved Tricluster

Figure 11 shows that the subset of patient  $\{P_1, P_3\}$  had the lowest TD Score in conserved tricluster. It means the tricluster in subset of patient  $\{P_1, P_3\}$  has a better quality compared to tricluster of another patient subset. In other word, patient  $P_1$  and  $P_3$  has a similar gene that affected by estrogen stimulation. Figure 12 shows the number of genes and the TD Score for each rank pattern in subset of patient  $\{P_1, P_3\}$ .

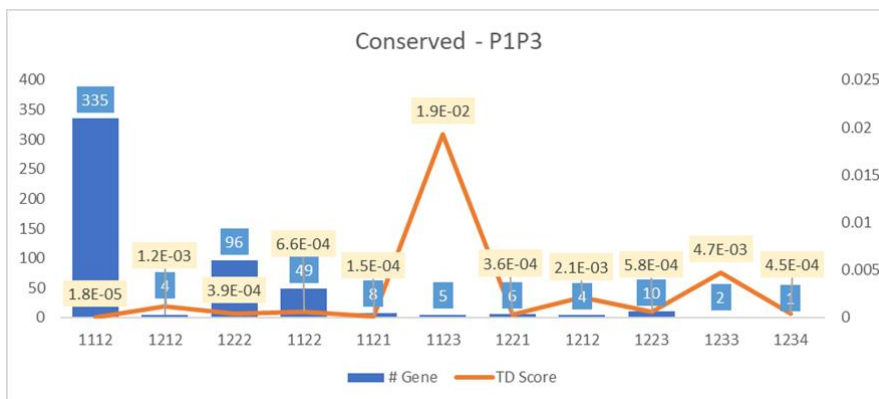


FIGURE 12. Number of Genes and TD Score for Each Rank Pattern of Conserved Tricluster at Subset of Patient  $\{P_1, P_3\}$

Figure 12 shows the rank pattern 1112 have the lowest TD Score ( $1.8E-05$ ) in the subset of patient  $\{P_1, P_3\}$  of conserved tricluster. It means that the tricluster with rank pattern 1112 has a better quality compared to other rank pattern in patient  $P_1$  and  $P_3$ . These tricluster and comprises 335 genes. In other words, 335 gene expression level was relatively unchanged in  $P_1$  and  $P_3$  body until 6 hours after estrogen stimulation but increased at 12 hours after estrogen stimulation. Figure 13 shows a heatmap of genes of the subset of patient  $\{P_1, P_3\}$  with 1112 rank pattern.

IDProbe	P1_0h	P1_3h	P1_6h	P1_12h	P3_0h	P3_3h	P3_6h	P3_12h
1053_at	9.10	8.88	9.20	10.29	9.10	8.93	9.01	10.13
1552277_a_at	8.47	8.34	8.55	8.73	8.47	8.48	8.67	8.80
1553015_a_at	9.69	9.49	9.65	10.38	9.84	9.75	9.60	10.37
1553101_a_at	10.02	9.71	10.14	10.42	9.88	9.69	10.19	10.53
1553103_at	7.39	7.46	7.30	7.59	7.42	7.63	7.53	7.55
1553106_at	5.74	5.81	6.04	6.41	5.38	5.86	6.35	6.55
1553322_s_at	4.97	4.80	5.04	5.55	4.96	4.79	5.19	5.48
1557961_s_at	4.06	4.44	4.33	4.94	3.98	4.25	4.55	5.43
1561042_at	4.78	4.92	4.75	5.26	4.89	4.74	5.03	5.43
1561720_at	3.76	3.72	3.69	3.81	3.84	4.00	3.82	3.99
1562903_at	7.12	7.24	7.04	7.61	7.03	7.08	7.20	7.24

FIGURE 13. Heatmap of Conserved Tricluster at Subset of Patient  $\{P_1, P_3\}$  with 1112 Rank Pattern

Figure 13 displays the heatmap for the conserved tricluster derived from the subset of patients  $\{P_1, P_3\}$  with a 1112 rank pattern. Unlike the constant tricluster, the gene expression levels in this tricluster show noticeable variations across different experimental conditions and time points (0h, 3h, 6h, and 12h). The heatmap reveals that while the expression levels remain relatively stable in earlier time points, there is a tendency for the pattern to change more significantly at the final time point (12h), as reflected by the 1112 rank pattern.

### 3.3 Divergent Tricluster Result

The evaluation results based on the divergent-patterned tricluster for each subset of the experimental condition are shown in Figure 14.

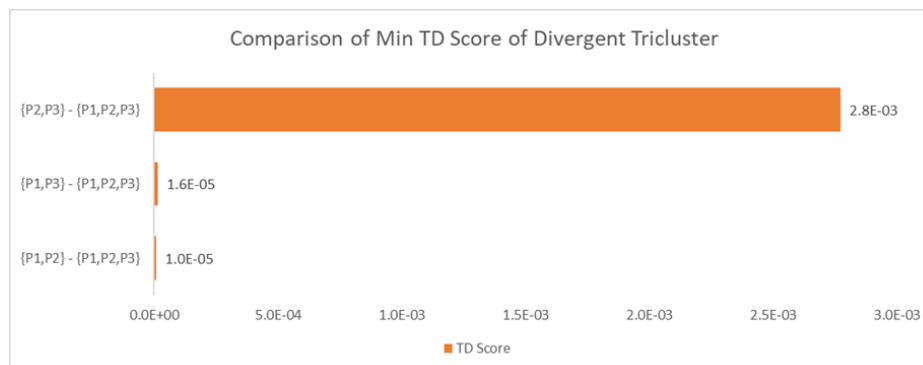


FIGURE 14. Comparison of Minimum TD Score for Each Subset of Patient of Divergent Tricluster



## OPTRICLUSTER FOR 3-DIMENSIONAL GENE EXPRESSION DATA

Figure 14 shows that the subset of patient  $\{P_1, P_2\} - \{P_1, P_2, P_3\}$  has a lowest TD Score in divergent tricluster. It means the tricluster in subset of patient  $\{P_1, P_2\} - \{P_1, P_2, P_3\}$  has a better quality compared to tricluster of another patient subset. In other word, patient  $P_1$  and  $P_2$  has a similar gene expression level's change, but patient  $P_3$  has a different gene expression level's change. Figure 15 shows the number of genes and the TD Score for each rank pattern in subset of patient  $\{P_1, P_2\} - \{P_1, P_2, P_3\}$ .

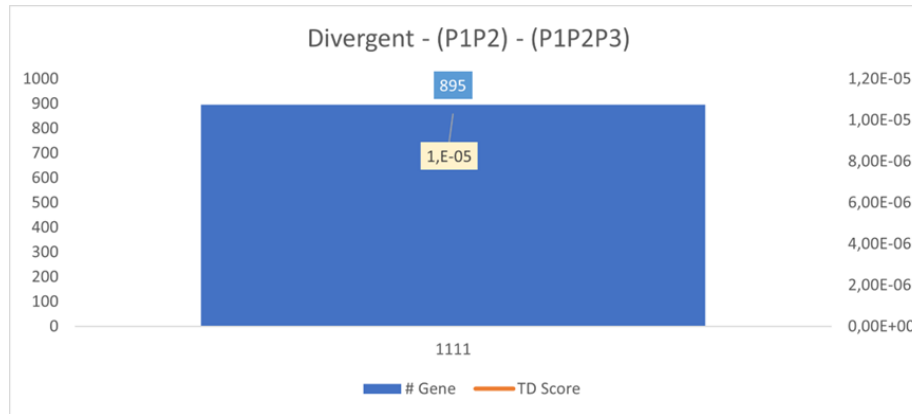


FIGURE 15. Number of Genes and TD Score for Each Rank Pattern of Divergent Tricluster at Subset of Patient  $\{P_1, P_2\} - \{P_1, P_2, P_3\}$

Figure 15 illustrates that there is only one pattern present in the divergent tricluster of patient  $\{P_1, P_2\} - \{P_1, P_2, P_3\}$ , specifically the 1111 pattern. This tricluster comprises 895 genes with a TD Score of  $1 \times 10^{-5}$ . In other words, the gene expression levels in patients  $P_1$  and  $P_2$  remain relatively constant across all time points, whereas they differ in patient  $P_3$ . Figure 16 provides a heatmap representation of this tricluster, highlighting the distinct expression dynamics.

IDProbe	P1_0h	P1_3h	P1_6h	P1_12h	P2_0h	P2_3h	P2_6h	P2_12h	P3_0h	P3_3h	P3_6h	P3_12h
1007_s_at	12,95172	12,74537	12,68614	12,41815	12,91964	12,60402	12,54167	12,36614	13,18524	13,60096	12,37241	13,5031
1405_i_at	5,086716	4,857387	4,58691	4,80141	5,278641	4,965048	4,673724	4,792326	5,963909	5,50368	5,181626	5,277879
1487_at	9,439973	9,414414	9,371814	8,724763	9,40789	9,399282	9,0632	8,526939	8,713588	8,281064	8,165197	8,836901
1552263_at	5,912978	5,737121	5,607969	6,051033	5,99286	6,21693	5,702076	5,594933	5,424058	5,665505	5,084986	5,242708
1552264_a_at	8,168768	8,14627	8,121576	8,577441	8,176735	7,798328	8,113346	8,615748	7,794599	7,473718	7,423852	7,383121
1552269_at	6,701865	6,299928	6,301691	6,432547	6,707431	6,532398	6,693771	6,358243	6,748756	7,291716	6,482405	6,357072
1552283_s_at	5,672753	5,491572	5,677583	5,317681	5,537176	5,669456	5,482725	5,130902	5,972586	5,29158	5,927453	5,039551
1552287_s_at	7,500052	7,602608	7,926951	7,899157	7,507735	8,195002	7,805031	8,232981	7,283427	7,564544	7,514164	7,909828
1552288_at	5,361984	5,505053	5,783534	5,601442	5,351019	5,172075	5,68479	5,391511	5,253409	5,310287	5,540704	5,282945
1552291_at	7,990787	7,864767	8,082109	8,314866	7,897576	8,720583	8,047755	8,221998	7,504114	7,156762	7,612391	7,170322
1552321_a_at	4,030416	3,860134	3,758389	3,873883	3,917184	3,966884	3,765768	4,102809	3,945076	3,091751	3,460245	3,642728
1552329_at	7,217411	7,191647	6,904301	6,945795	7,261313	7,417123	6,773568	6,796491	7,733085	7,730386	7,347663	7,380413

FIGURE 16. Heatmap of Divergent Tricluster at Subset of Patient  $\{P_1, P_2\} - \{P_1, P_2, P_3\}$  with 1111 Rank Pattern

Figure 16 illustrates the heatmap of the divergent tricluster. The heatmap reveals that while the gene expression patterns in  $P_1$  and  $P_2$  remain relatively constant across all time points,  $P_3$  demonstrates a noticeably divergent pattern. The expression levels in  $P_3$  vary significantly over time, deviating from the stable trends observed in  $P_1$  and  $P_2$ . This suggests that the genes in this tricluster behave consistently in  $P_1$  and  $P_2$ , whereas  $P_3$  exhibits dynamic changes that could be attributed to unique regulatory mechanisms or differing experimental responses.

### 3.4 Gene Ontology Result

Based on the constant tricluster results, we can see that the best tricluster is obtained within the patient subset  $\{P_1, P_2, P_3\}$ . Then, according to the conserved tricluster results, the best tricluster is found in the pattern (1112) within the patient subset  $\{P_1, P_3\}$ . Meanwhile, based on the divergent tricluster results, the best tricluster is found in the pattern (1112) within the patient subset  $\{P_1, P_2\} - \{P_1, P_2, P_3\}$ . Table 3 illustrates the biological interpretation of each tricluster result.

**Table 3.** Gene Ontology's Results

Tricluster's Pattern	GO					
	Biological Process		Cellular Component		Molecular Function	
	ID	Name	ID	Name	ID	Name
<b>Constant</b>	GO:0015031	Protein transport	GO:0005829	Cytosol	GO:0005515	Protein binding
<b>Conserved</b>	GO:0006915	Apoptotic process	GO:0005829	Cytosol	GO:0005515	Protein binding
<b>Divergent</b>	GO:0015031	Protein transport	GO:0005829	Cytosol	GO:0005515	Protein binding

Table 3 presents the triclustering results, which reveal biologically relevant patterns among breast cancer-associated genes. Specifically, the analysis identified several key processes and functions. Protein transport (BP) was highlighted as it plays a critical role in the movement of proteins essential for cellular functions, which can become dysregulated in cancer [20] [21] [22]. Apoptotic processes (BP) were also identified, emphasizing genes involved in programmed cell death, a mechanism that can either suppress or promote cancer progression depending on the context [20] [21] [23]. Additionally, the cytosol (CC) was identified as a crucial cellular component, serving as the site of various metabolic and regulatory processes that are often disrupted in breast cancer cells [20] [21] [24]. Lastly, protein binding (MF) was noted, reflecting interactions that

regulate cellular signaling pathways and may influence cancer development, such as those involving hormone receptors in estrogen-sensitive breast cancer [20] [21] [25].

#### **4. CONCLUSIONS**

Based on the findings of this study, several conclusions were drawn in alignment with the research objectives. First, gene expression data with gene-sample-time (GST) dimensions were successfully analyzed using the OPTricluster method, which efficiently grouped genes exhibiting similar expression patterns across samples and time points. Second, the performance of the OPTricluster method was assessed using the TD Score and Silhouette Score. The optimal scenario was achieved under the condition of  $IQR < 0.75$  with  $\delta = 1.1$ , yielding triclusters of the highest quality. Third, the triclustering results were biologically interpreted using Gene Ontology (GO), revealing that the identified genes were involved in specific biological processes, cellular components, and molecular functions. Specifically, these included protein transport and apoptotic processes (Biological Processes), localization in the cytosol (Cellular Components), and protein binding activities (Molecular Functions). This study offers significant insights into gene expression patterns in breast cancer patients and their underlying biological functions, contributing to advancements in diagnostic and therapeutic strategies.

#### **CONFLICT OF INTERESTS**

The authors declare that there is no conflict of interests.

#### **ACKNOWLEDGMENT**

This research gratefully acknowledges financial support from Hibah Publikasi Pascasarjana FMIPA UI 2024.

#### **REFERENCES**

- [1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer Statistics, 2020, CA: Cancer J. Clin. 70 (2020), 7–30.  
<https://doi.org/10.3322/caac.21590>.
- [2] F. Bray, J. Ferlay, I. Soerjomataram, et al. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, CA: Cancer J. Clin. 68 (2018), 394–424.  
<https://doi.org/10.3322/caac.21492>.
- [3] Ministry of Health Republic of Indonesia, Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan, <https://kemkes.go.id/id/kanker-payudara-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan>. Accessed: Apr. 1, 2022.

- [4] K.A. Hoadley, C. Yau, D.M. Wolf, et al. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin, *Cell* 158 (2014), 929–944.  
<https://doi.org/10.1016/j.cell.2014.06.049>.
- [5] R. Fajriyah, Paper Review: An Overview on Microarray Technologies, *Bull. Appl. Math. Math. Educ.* 1 (2021), 21. <https://doi.org/10.12928/bamme.v1i1.3854>.
- [6] D.A. Ihsani, A. Arifin, M.H. Fatoni, Klasifikasi DNA Microarray Menggunakan Principal Component Analysis (PCA) dan Artificial Neural Network (ANN), *J. Teknik ITS* 9 (2020), A124–A129.  
<https://doi.org/10.12962/j23373539.v9i1.51637>.
- [7] J. Wöhrle, S.D. Krämer, P.A. Meyer, et al. Digital DNA Microarray Generation on Glass Substrates, *Sci. Rep.* 10 (2020), 5770. <https://doi.org/10.1038/s41598-020-62404-1>.
- [8] K. Huang, C. Xiao, L.M. Glass, et al. Machine Learning Applications for Therapeutic Tasks with Genomics Data, *Patterns* 2 (2021), 100328. <https://doi.org/10.1016/j.patter.2021.100328>.
- [9] G.W. Sledge, Curing Metastatic Breast Cancer, *J. Oncol. Pract.* 12 (2016), 6–10.  
<https://doi.org/10.1200/JOP.2015.008953>.
- [10] C.M. Perou, T. Sørli, M.B. Eisen, et al. Molecular Portraits of Human Breast Tumours, *Nature* 406 (2000), 747–752. <https://doi.org/10.1038/35021093>.
- [11] R. Henriques, S.C. Madeira, Triclustering Algorithms for Three-Dimensional Data Analysis: A Comprehensive Survey, *ACM Comput. Surv.* 51 (2019), 1–43. <https://doi.org/10.1145/3195833>.
- [12] J.S. Carroll, C.A. Meyer, J. Song, et al. Genome-Wide Analysis of Estrogen Receptor Binding Sites, *Nat. Genet.* 38 (2006), 1289–1297. <https://doi.org/10.1038/ng1901>.
- [13] A.B. Tchagang, S. Phan, F. Famili, et al. Mining Biological Information from 3D Short Time-Series Gene Expression Data: The OPTriclust Algorithm, *BMC Bioinform.* 13 (2012), 54. <https://doi.org/10.1186/1471-2105-13-54>.
- [14] D. Siska, D. Sarwinda, T. Siswantining, et al. Triclustering Algorithm for 3D Gene Expression Data Analysis Using Order Preserving Triclustering (OPTriclust), in: 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), IEEE, Semarang, Indonesia, 2020: pp. 1–6.  
<https://doi.org/10.1109/ICICoS51170.2020.9299101>.
- [15] P. Swathyriyadharsini, K. Premalatha, Hybrid Cuckoo Search with Clonal Selection for Triclustering Gene Expression Data of Breast Cancer, *IETE J. Res.* 69 (2023), 2328–2336.  
<https://doi.org/10.1080/03772063.2021.1911691>.
- [16] S. Chockalingam, M. Aluru, S. Aluru, Microarray Data Processing Techniques for Genome-Scale Network Inference from Large Public Repositories, *Microarrays* 5 (2016), 23.  
<https://doi.org/10.3390/microarrays5030023>.

- [17] A. Atira, B.N. Sari, Penerapan Silhouette Coefficient, Elbow Method dan Gap Statistics untuk Penentuan Cluster Optimum dalam Pengelompokan Provinsi di Indonesia Berdasarkan Indeks Kebahagiaan, *J. Ilm. Wahana Pendidik.* 9 (2023), 76–86.
- [18] D.P. Hill, B. Smith, M.S. McAndrews-Hill, J.A. Blake, Gene Ontology Annotations: What They Mean and Where They Come From, *BMC Bioinform.* 9 (2008), S2. <https://doi.org/10.1186/1471-2105-9-S5-S2>.
- [19] A. Bhar, M. Haubrock, A. Mukhopadhyay, et al. Multiobjective Triclustering of Time-Series Transcriptome Data Reveals Key Genes of Biological Processes, *BMC Bioinform.* 16 (2015), 200. <https://doi.org/10.1186/s12859-015-0635-8>.
- [20] Y. Zhao, J. Wang, J. Chen, et al. A Literature Review of Gene Function Prediction by Modeling Gene Ontology, *Front. Genet.* 11 (2020), 400. <https://doi.org/10.3389/fgene.2020.00400>.
- [21] C. Dessimoz, N. Škunca, eds., *The Gene Ontology Handbook*, Springer, New York, NY, 2017. <https://doi.org/10.1007/978-1-4939-3743-1>.
- [22] European Bioinformatics Institute (EMBL-EBI), GO:0015031 - Protein Transport, QuickGO, (2023). <https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0015031>. Accessed: Apr. 12, 2024.
- [23] European Bioinformatics Institute (EMBL-EBI), GO:0006915 – Apoptotic Process, QuickGO, (2023). <https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0006915>. Accessed: Apr. 12, 2024.
- [24] European Bioinformatics Institute (EMBL-EBI), GO:0005829 – Cytosol, QuickGO, (2023). <https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0005829>. Accessed: Apr. 12, 2024.
- [25] European Bioinformatics Institute (EMBL-EBI), GO:0005515 – Protein Binding, QuickGO, (2023). <https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0005515>. Accessed: April 12, 2024.