5 Artificial Intelligence Techniques Based on K-Means$_{\text{Two Way}}$ Clustering and Greedy Triclustering Approach for 3D Gene Expression Data (GED)

N. Narmadha and R. Rathipriya

Abstract

Artificial intelligence (AI) refers to a machine's or robot's capacity to carry out operations that would typically require human comprehension and intelligence. Classification algorithms, regression algorithms, and clustering algorithms have traditionally been the three basic pillars of AI. The use of K-Means$_{\text{Two Way}}$ and Greedy approaches for the triclustering of 3D GED using Artificial Intelligence Techniques is discussed in this chapter. The main goal is to create a triclustering algorithm that extracts triclusters from a given dataset with 100% Tri$_{\text{Gene}}$, Tri$_{\text{Sample}}$, and Tri$_{\text{Timepoint}}$ coverage. This technique is combined with the greedy heuristics to find the ideal tricluster with the highest degree of coherence with the largest volume. On a 3D Yeast Cell Cycle (YCC) dataset, the suggested Greedy-based Triclustering approach is evaluated. In terms of extracting the larger volume tricluster with high MCV, Greedy$_{\text{Tri}}$ outperformed K-Means$_{\text{Two Way}}$ clustering.

Keywords

Artificial intelligence · Triclustering · K-Means$_{\text{Two Way}}$ · Greedy triclustering · MCV · 3D GED

N. Narmadha (✉)
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamil Nadu, India

R. Rathipriya
Department of Computer Science, Periyar University, Salem, Tamil Nadu, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

J. M. Chatterjee, S. K. Saxena (eds.), *Artificial Intelligence in Medical Virology*, Medical Virology: From Pathogenesis to Disease Control,
https://doi.org/10.1007/978-981-99-0369-6_5

5.1 Introduction

Gene expression analysis, to put it simply, is the study of how genes are transcribed into functional gene products. The advancement of 3D data is accelerating daily. Data on 3D gene expression monitoring is particularly challenging. One of the key research issues is the detection of highly expressed gene patterns in this 3D GED using data mining techniques, to extract the numerous patterns from the 3D GED. The most popular methods for analysing the 3D GED are clustering and biclustering. However, because the time point cannot be focused, the clustering and biclustering approaches are unable to extract patterns from the 3D data. Triclustering is introduced to address the issues with clustering and biclustering. This chapter mainly focused on triclustering of 3D GED based on artificial intelligence. The artificial intelligence-based triclustering of 3D GED was the main topic of this chapter. Triclustering is also referred to as 3D data clustering, tradic data clustering, multi clustering, etc. Triclustering, in other terms, is the simultaneous grouping of a gene and a sample at a particular time point (Narmadha 2017).

Contributions of this chapter are all follows:

- GED dataset analysis can find local structures.
- Extracting unique and practical information from a biological perspective.
- The YCC expression dataset contains three categories as follows:
  - CDC15 database contains 8832 genes, 9 samples, and 24 time points
  - Elutriation database contains 7744 genes, 9 samples, and 14 time points
  - Pheromone database contains 7744 genes, 8 samples, and 18 time points.
- Medical Domain—Analysis of this dataset also helps in various ways such as (1) diagnosis, (2) prognosis, (3) treatment planning, as well as (4) drug discovery, (5) drug assessment, and (6) drug development.

Scope

- Treatment of drug development
- Diagnosis diseases/cancer

The structure of this chapter is as follows: The background review required for the research is discussed in Sect. 5.2. The thorough explanations of the proposed task are provided in Sect. 5.3. The findings and discussions are presented in Sect. 5.4. In Sect. 5.5, a summary of this chapter is provided.

5.2 Background Study

The work that is pertinent to 3D GED analysis is discussed in this area, and it is specifically related to Greedy-based clustering, biclustering, and triclustering methods.

The OAC-triclustering techniques to implement certain suggested modifications depending on the prime operators (Gnatyshak 2015). To make minor adjustments depending on clustering processes in order to maximize the effectiveness of the specialist-generalist categorization scheme (Gnatyshak 2014). To enhance search efficiency, the binary Particle Swarm Optimization (BPSO) technique incorporates a pattern-driven local search operator (Rathipriya et al. 2011). Gene class sensitivity (GCS) BPSO utilized largely for gene selection. Gene selection efficiency is achieved by using (1) K-Nearest Neighbour (KNN) and (2) Support Vector Machine (SVM) classifiers to predict microarray data with high accuracy (Han et al. 2017).

The bicluster in the GED is eliminated using the Particle Swarm Optimization (PSO) technique. This method's main objective is to include each gene expression data with matrix component in the overlapping bicluster (Li et al. 2014). Using the biclustering method, the coherent bicluster for GED is defined with low MSR (Mean Square Residue) and large row variance (Pontes et al. 2015). This type of problem is solved using a variety of optimization techniques, including (1) Nelder Mead with Levy Flight (NMLF) and (2) Nelder Mead Forbidden Search (NMFS), which are both introduced and contrasted. Comparing NM with Levy Flight to Nelder Mead's Tabu search, NM with Levy Flight demonstrates greater efficiency and offers a more optimal global answer (Balamurugan et al. 2016)

By lowering the residue or merit function of the biclusters, the coclustering techniques are utilized to group the data on gene expression. The stochastic heuristic technique is more applicable because of this merit function. It is advised that biclusters be optimized using a Parallel Genetic Algorithm (PGA) in order to decrease local optima in the biclustering approach and improve the possibility of global optima (Wei Shen 2012). Utilizing hybrid EDA-GA, can't be easily converges but also offers the full solution, the GED is evaluated (Liu 2006). To more accurately handle the biclustering problem, hybrid PSO-GA is provided (Xie et al. 2007)

The K-Means one-way clustering algorithm is used to develop the highly clustered small disjoint submatrices first. Second, the algorithm for greedy searches, which is mostly utilized for enlargement of seed. The Greedy Search (GS) algorithm is used to extract the output and starting binary PSO population in order to categorize the bicluster (Shyama and Idicula 2010). K-nearest neighbour (K-NN) as an IBPSO evaluator to resolve the GED classification problem techniques help reduce the overall number of functions as necessary (Chuang et al. 2008)

For clustering microarray data, hierarchical methods, FCM, and K-means are employed. But for clustering microarrays, PSO based on K-means provides the outperformance well (Lopamudra Dey 2014). The two-clustering approach scaled and shifted the advantage of the function, which is mostly used to enlarge the bicluster, in accordance with a pattern. But in order to establish scaling patterns and promote coherent evolution to build the bicluster, this measure has drawbacks (Thangavel et al. 2012). A coherent pattern based on tricluster with a higher MCV

and greater volume is to be found using greedy two-way K-Means clustering (Narmadha and Rathipriya 2019a, b).

Greedy algorithms combined with other approaches are frequently used to analyse GED and to create high-quality biclusters from online usage data. In order to determine the quality of the tricluster, this chapter introduces the greedy approach using the triclustering technique with a coherent pattern.

5.2.1 Issues in the Literature

Some negative aspects of the literature exist. Following is a list of some of them:

- The effectiveness of the search is only relevant for local searches.
- In order to cover all gene expression elements, the overlapping will take place in the biclusters.
- In terms of defining scaling patterns and maintaining bicluster evolution, qualitative measurements have limitations of their own. For example, the accuracy of the high consistency of the given cluster in the literature is only about 70–80%.

5.3 Proposed Work

5.3.1 Mean Correlation Value Equation for Tricluster ($\mathbf{Tri_{MCV}}$)

Mean Correlation Value equation for deriving a Tricluster (Tri_{MCV}) is discussed in Eq. (5.1)

$$\sum_{a}\sum_{b}\left(X_{ab}-\overline{X}\right) \times \left(Y_{ab}-\overline{Y}\right) \Big/ \sqrt{\left(\sum_{\mathrm{a}}\sum_{b}\left(X_{ab}-\overline{X}\right)^{2}\right)\left(\sum_{a}\sum_{b}\left(Y_{ab}-\overline{Y}\right)^{2}\right)} \tag{5.1}$$

where $\overline{X}=\frac{\sum_{a}\sum_{b}(X_{ab})}{a*b}$, $\overline{Y}=\frac{\sum_{a}\sum_{b}(Y_{ab})}{a*b}$

Tri_{MCV} has a range of [0, 1]. If the X value is close to 1 (when the tricluster is highly correlated) otherwise, it is low-correlated or null-correlated (Narmadha and Rathipriya 2018).

5.3.2 Function for Fitness

Finding triclusters with a high Mean Correlation Value (Tri_{MCV}) and a higher volume is the major goal ($\text{Tri}_{\text{Volume}}$). To extract the best tricluster, the fitness function F (Tri_{Gene}, $\text{Tri}_{\text{Sample}}$, $\text{Tri}_{\text{Timepoint}}$) is specified in Eq. (5.2).

$$F\left(\text{Tri}_{\text{Gene}}, \text{Tri}_{\text{Sample}}, \text{Tri}_{\text{Timepoint}}\right) = \left\{ \left|\text{Tri}_{\text{Gene}}\right| * \left|\text{Tri}_{\text{Sample}}\right| * \left|\text{Tri}_{\text{Timepoint}}\right|, \text{ if } \text{Tri}_{\text{MCV}}(\text{Tricluster}) \geq \delta \; 0, \text{ Otherwise} \right. \tag{5.2}$$

where $|\text{Tri}_{\text{Gene}}|$, $|\text{Tri}_{\text{Sample}}|$, $|\text{Tri}_{\text{Timepoint}}|$ are the no. of genes: Tri_{Gene}, no. of samples $\text{Tri}_{\text{Sample}}$, and the no. of time points $\text{Tri}_{\text{Timepoint}}$ of tricluster.

5.3.3 Description of K-Means$_{\text{Two Way}}$

5.3.3.1 Tricluster Generation Using K-Means$_{\text{Two Way}}$ Clustering

In order to create the tricluster, of K-Means$_{\text{Two Way}}$ clustering algorithms are as follows:

- To Use the K-Means$_{\text{Two Way}}$ clustering algorithm to create kg and ks clusters at each $\text{Tri}_{\text{timepoint}}$ 'T' along with the row and column dimensions of the data. For each time point, combine these clusters to produce the k_{gene}*k_{sample} initial bicluster, where $\text{Data}_{\text{Timepoint}}$ is the 2D data of the Tri_{Gene} and $\text{Tri}_{\text{Sample}}$ at $\text{Tri}_{\text{Timepoint}}$ 'T' and T = 1, 2,...nT.
- The binary string for these triclusters has the length n_{binary}*(n_{Gene}+n_{Sample}). $\text{Binary}_{\text{bicluster}}$ is used to indicate it. The encoded bicluster's length is shown in Table 5.1.
- Create a random binary string of size n_{binary}*$n_{\text{Timepoint}}$, where $n_{\text{Timepoint}}$ is the number of Data_{GST} time points and $n_{\text{bicluster}}$ is the number of biclusters. It has the symbol $\text{Binary}_{\text{Timepoint}}$.

Then, combine two binary strings with the same number of rows, $\text{Binary}_{\text{bicluster}}$ and $\text{Binary}_{\text{Timepoint}}$, to create a new binary string with the size $n_{\text{bicluster}}$* (n_{Gene}+n_{Sample}+$n_{\text{Timepoint}}$), which is then utilized as a binary-encoded tricluster for further processing. Table 5.2 displays a single (n_{Gene}+n_{Sample}+$n_{\text{Timepoint}}$)-length binary-encoded tricluster. The notation and its descriptions are displayed in Table 5.3. Algorithm 5.1 describes the Tricluster Seed Formation phase utilizing K-means: Two-Way Clustering (Narmadha and Rathipriya 2020).

Table 5.1 Length of the encoded bicluster (n_{Gene}+n_{Sample})

$Trigene_1$	$Trigene_2$	...	$Trigene_{nG-1}$	$Trigene_{nG}$	$Trisample_1$	$Trisample_2$	...	$Trisample_{nS}$

Table 5.2 The encoded tricluster's length (n_{Gene}+n_{Sample}+$n_{Timepoint}$)

$Trigene_1$	$Trigene_2$	. . .	$Trigene_{nG-1}$	Trigenen	$Trisample_1$	$Trisample_2$	. . .	$Trisample_{nS}$	TritimepointI	Tritimepoint2	. . .	TritimepointnT
n_{Gene}-8832					n_{Sample}-9				$n_{Timepoint}$-24			
n_{Gene}-7744					n_{Sample}-9				$n_{Timepoint}$-14			
n_{Gene}-7744					n_{Sample}-9				$n_{Timepoint}$-18			

Table 5.3 Notations and representations

Notations	Representations
$Data_T$	Data in two dimensions for the gene and the sample at time 'T'
k_{gene}	'k' is $gene_{clusters}$
k_{sample}	'k' is $sample_{cluster}$
$Binary_{bicluster}$	Bicluster—Encoded with Binary Value
$Binary_{Timepoint}$	Time points—Encoded with Binary Value
$n_{bicluster}$	No. of bicluster's
n_{Gene}	No. of Genes
n_{Sample}	No. of Samples
$n_{Timepoint}$	No. of time points
$Data_{GST}$	Dataset with three dimensions of data such as Genes, Samples, and Time points
Tri_{gene}	A subset of specific tricluster's genes
Tri_{sample}	A subset of specific tricluster's samples
$Tri_{timepoint}$	A subset of specific tricluster's timepoint
$Tri_{Optimal}$	To find the Optimal tricluster
$Tri_{population}$	Population Size
Tri_{Gene}	No. of the tricluster's Genes
Tri_{Sample}	No. of tricluster's Samples
$Tri_{Timepoint}$	No. of tricluster's Timepoint
Tri_{MCV}	Mean Correlation Value in the Tricluster
Tri_{Volume}	Volume of the tricluster
Tri_{Gene} coverage	No. of Genes Covered by Tricluster
Tri_{Sample} coverage	No. of Sample Covered by Tricluster
$Tri_{Timepoint}$ coverage	No. of Time Points Covered by Tricluster
$K\text{-}Means_{Two\ Way}$	Two-way K-means
$Greedy_{Tri}$	Greedy Triclustering

Algorithm 5.1 Tricluster Seed Formation Step Using $K\text{-}Means_{Two\ Way}$ Clustering

Input: $Data_{GST}$, 3D GED

Output: Optimal Tricluster ($Tri_{Optimal}$)

Step 1: Initialize the correlation distance measure

Step 2: For each 't' in $n_{Timepoint}$

(a) Apply the **$K\text{-}Means_{Two\ Way}$** Clustering algorithm on the no. of genes (Tri_{Gene}) and generate 'k_{gene}' gene clusters.

(b) Apply a two-way K-Means Clustering algorithm on the no. of Samples (Tri_{Sample})and generate 'k_{sample}' sample clusters.

(c) Combine 'k_{gene}' is represented as gene cluster and 'k_{sample}' is represented as a sample cluster to get $k_{gene}{}^{*}k_{sample}$ as initial biclusters.

(continued)

Algorithm 5.1 (continued)

(d) Encode the bicluster into a binary string of length $n_{bicluster}^{*}(n_{Gene}+n_{Sample})$ and it is represented as $Binary_{bicluster}$

(e) Generate a random binary string of length $n_{bicluster}^{*}n_{Timepoint}$ and it is represented as $Binary_{Timepoint}$

(f) Concatenate ($Binary_{bicluster}$ and $Binary_{Timepoint}$).

(g) Encode the Tricluster into a binary string of length $n_{bicluster}^{*}(n_{Gene}+n_{Sample}+n_{Timepoint})$

End

Step 3: Call Greedy Triclustering()

Step 4: Return the Optimal Tricluster ($Tri_{Optimal}$)

5.3.4 Specification of Greedy Triclustering

A heuristic is a greedy algorithm that uses this method of problem-solving to achieve the global optimum by looking for the local optimal solution at each stage. The $Greedy_{Tri}$ algorithm was used to determine the ideal tricluster using the initial population as input. In the proposed $Greedy_{Tri}$ method, a list of genes, samples, and time points that were eliminated from the tricluster were preserved separately. More $TriG_{enes}$, $Tri_{Samples}$, and $Tri_{Timepoints}$ are individually added to each tricluster to increase its size (Narmadha and Rathipriya 2019a, b).

The finest element is chosen from the given list such as (1) gene list, (2) sample list, or (3) time point list in this procedure, and it is additional to the tricluster. The value of Tri_{Volume} and Tri_{MCV} included in the element determines the quality of the tricluster. The best element is that which results in a greater Tri_{MCV} being additional to the tricluster. The initial tricluster expands from the given list such as (1) gene list, then (2) the sample list (3) the time point list, until the tricluster's Tri_{MCV} value increases. The next gene, sample, or time point are chosen using what is referred to as the 'greedy strategy', which results in an ideal tricluster with a greater Tri_{MCV} value. In Algorithm 5.2, this $Greedy_{Tri}$ is described.

Algorithm 5.2 Greedy Triclustering ($Greedy_{Tri}$)

Input : Initialize of ($n_{Gene}+n_{Sample}$)

Output: $Gene_{Enlargement}$ and $Gene_{Refinement}$ tricluster

Step 1: Generate the random population using the algorithm1

Step 2: For each and every Tri_{Gene}

(i) Call $Gene_{Enlargement}$ (gene (G', S', T'))

(ii) Call $Gene_{Refinement}$ (gene (G', S', T'))

Step 3: Return the $Gene_{Enlargement}$ and $Gene_{Refinement}$ tricluster

Step 4:

// Subfunctions of $Gene_{Refinement}$ and $Gene_{Refinement}$ tricluster

(continued)

Algorithm 5.2 (continued)

Call $\text{Gene}_{\text{Enlargement}}$ (gene (G', S', T'))

Step 1: Set of genes as 'g' not in G'

Step 2: Set of sample as 's' not in S'

Step 3: Set of time point as 't' not in T'

Step 4: For each node g/s/t

If Tri_{MCV} (union (gene, (g/s/t)))> Tri_{MCV} (gene (G,S,T)) then

1. Add g/s/t to gene (G,S,T)
2. End (if)

End (for)

Step 5: Return the $\text{Enlarged}_{\text{gene}}$ set

Call $\text{Gene}_{\text{Refinement}}$ (gene (G', S', T'))

Step 1: For each node gene/sample/timepoint in the Enlarged gene

Remove node gene/sample/timepoint in the Enlarged gene

G"/S"/T' be set of rows as r/columns as c/time point as t in G'/S'/T' but not contained g/s/t

If Tri_{MCV} $\text{Enlarged}_{\text{gene}}$ (G", S", T")> Tri_{MCV} ($\text{Enlarged}_{\text{gene}}$ (G'/S'/T'))

Update G'/S'/T'

End (if)

End (for)

Step 2: Return $\text{Refined}_{\text{gene}}$ set G", and A (G", S', T') as refined tricluster.

5.4 Result and Analysis

The performance level of the $\text{Greedy}_{\text{Tri}}$ and $\text{K-Means}_{\text{Two Way}}$ triclustering algorithms for three datasets was shown in Table 5.4. It is well known that the $\text{Greedy}_{\text{Tri}}$ technique significantly boosted both the mean volume and the mean Tri_{MCV} of the triclusters. According to these findings, the $\text{Greedy}_{\text{Tri}}$ technique did a good job of extracting the higher volume tricluster for the given Tri_{MCV}.

The clustered vertical bar chart for the $\text{Greedy}_{\text{Tri}}$ and K-Means: Two-Way approaches' volume-based performance is shown in Fig. 5.1. Visually, it is evident that the $\text{Greedy}_{\text{Tri}}$ method extends the K-Means: Two-Way triclustering method's triclusters, and this is reflected in the volume of the resulting clusters. The

Table 5.4 Mean presentation of $\text{K-Means}_{\text{Two Way}}$ and $\text{Greedy}_{\text{Tri}}$ triclustering methods

Datasets	Methods	Mean $\text{Tri}_{\text{Volume}}$	Mean Tri_{MCV}
$\text{Data}_{\text{CDC15}}$	$\text{K-Means}_{\text{Two Way}}$	1192.125	0.941
	$\text{Greedy}_{\text{Tri}}$	47,390	0.951
$\text{Data}_{\text{Elutriation}}$	$\text{K-Means}_{\text{Two Way}}$	763.7143	0.940
	$\text{Greedy}_{\text{Tri}}$	38,705.79	0.940
$\text{Data}_{\text{Pheromone}}$	$\text{K-Means}_{\text{Two Way}}$	799.5556	0.945
	$\text{Greedy}_{\text{Tri}}$	70,951.56	0.959

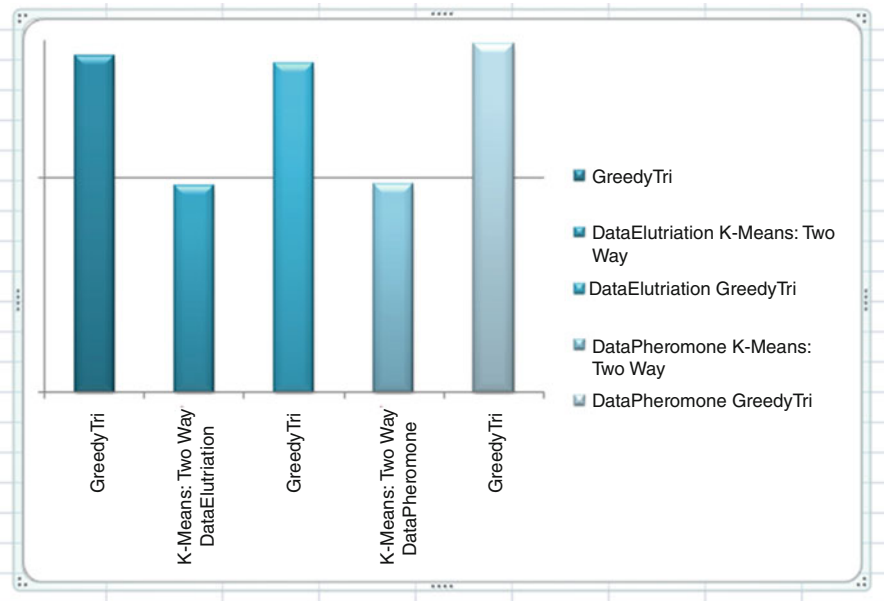

Fig. 5.1 Mean volume of K-Means: two-way and Greedy$_{\text{Tri}}$ triclustering methods

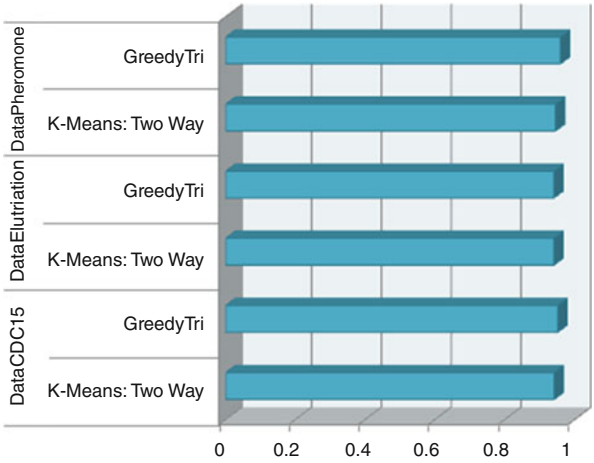

Fig. 5.2 Mean performance of K-Means: two-way and Greedy$_{\text{Tri}}$ triclustering methods

performance of the Greedy$_{\text{Tri}}$ and K-Means: Two-Way triclustering approaches, based on Tri$_{\text{MCV}}$, is shown in Fig. 5.2 as a clustered horizontal bar chart. It is discovered that while there is a significant difference in their Tri$_{\text{Volume}}$, there is little variation in their Tri$_{\text{MCV}}$. That is what makes the Greedy Triclustering approach special. Tricluster validation may be biological or statistical, depending on the features of the produced triclusters and the genes annotated in the various triclusters. This chapter primarily focused on the biological verification of the artificial intelligence-assisted tricluster.

- Utilize the GoTermFinder web tool to assess the triclusters physiologically.
- In order to determine what the genes in a given list may have in common, it looks for substantial shared GO keywords.
- In fact, the biological criteria make it possible to assess the quality of the generated triclusters by determining if their genes share any biological traits (Narmadha and Rathipriya 2019a, b).

5.4.1 CDC15 Experiment Using 3D GED

- The YCC data, $\text{Greedy}_{\text{Tri}}$, is capable to extract extremely correlated subset genes with larger volume. From the input list 3336 genes are known and not ambiguous. Also, 288 duplicates were removed from the input list. The removed genes are identified to be either unknown or ambiguous in the dataset.
- 7166 genes in all were utilized to determine the background distribution of GO keywords. Out of the 3435 terms that were found, 24 were shown.
- The total number of $\text{Tri}_{\text{Timepoint}}$ in the CDC15 experiment is represented as 10–290 m and the difference between the time points is represented as 30 m.
- Figure 5.3 shows the representation of $\text{Tri}_{\text{Optimal}}$ for the $\text{Data}_{\text{CDC15}}$. Table 5.5 shows the Highly Correlated Genes for the $\text{Data}_{\text{CDC15}}$
- $\text{Data}_{\text{CDC15}}$ from YCC: Biological Significant for Biological Process shown in Table 5.6.

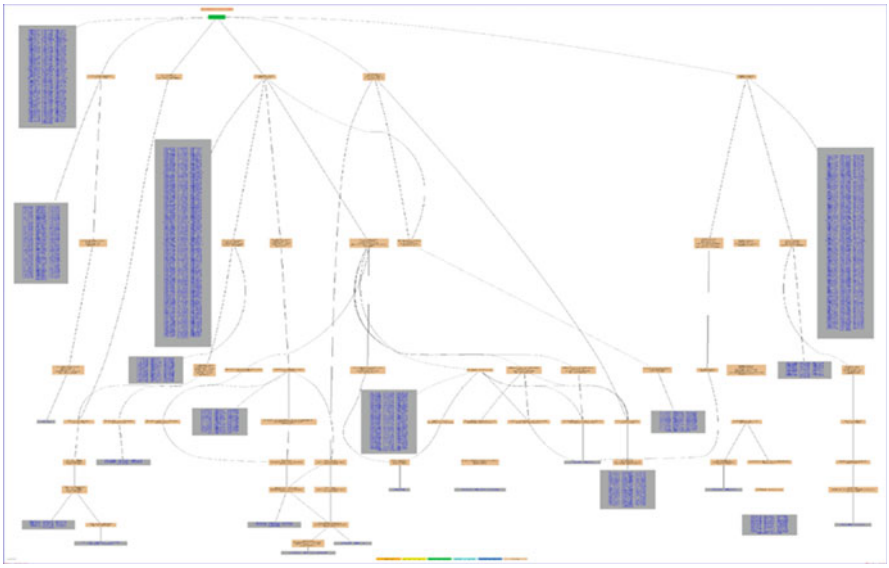

Fig. 5.3 Representation of $\text{Tri}_{\text{Optimal}}$ for the $\text{Data}_{\text{CDC15}}$

Table 5.5 Highly correlated genes for the Data$_{\text{CDC15}}$

YHR197W	YKL203C	YEL009C	YDR201W	YHR119W	YLL003W
YOL071W	YPL139C	YEL015W	YDL127W	YMR072W	YCL048W
YCR060W	YIL115C	YCR092C	YAL018C	YLR009W	YGL175C
YBR120C	YHR115C	YLR218C	YJL159W	YDR130C	YOR250C
YIL075C	YLR467W	YLR408C	YKL126W	YKR100C	YLR307W
YGL131C	YDL164C	YDR258C	YGL066W	YLR248W	YKR076W
YFL049W	YDR328C	YKR086W	YDR190C	YMR305C	YHR102W
YEL077C	YAL002W	YPR049C	YLR356W	YCL031C	YBR202W
YKL190W	YOR058C	YDR309C	YGR133W	YCL056C	YPL074W
YOL109W	YLR399C	YPR179C	YOR152C	YDR356W	YMR190C
YBL031W	YBL043W	YDR069C	YOR330C	YBR135W	YPL029W
YAL029C	YPL004C	YGL049C	YLR067C	YDR295C	YAL007C
YLR168C	YJL180C	YIL041W	YNL084C	YLL026W	YGR239C
YNL218W	YHR039C	YNL021W	YOR056C	YOL126C	YLR461W
YDL195W	YPR194C	YNL006W	YML072C	YIL062C	YBL006C
YNL121C	YGL120C	YNL039W	YIL132C	YDL139C	YMR260C
YHL027W	YKL137W	YLR276C	YIR010W	YMR265C	YDR440W
YGL115W	YEL056W	YHR160C	YBR114W	YKL087C	YDR319C
YJR151C	YDL093W	YDR065W	YAL048C	YHR041C	YDR486C
YOR073W	YNL066W	YDR397C	YBR017C	YLR341W	YER106W
YBR237W	YKL192C	YMR163C	YMR211W	YER169W	YNL310C
YLR006C	YDR150W	YBR091C	YML010W	YDR437W	YBL078C

5.4.2 Elutriation Experiment Using 3D GED

- The YCC data, Greedy$_{\text{Tri,}}$ is capable to extract extremely correlated genes with larger volume. From the input list 3320 genes are known and not ambiguous. Also, 301 duplicates were removed from your input list. The removed genes are identified to be either unknown or ambiguous in the dataset. 7166 genes in all were utilized to determine the background distribution of GO keywords. Ten terms are being displayed out of the 3474 total identified. The total number of Tri$_{\text{Timepoint}}$ in the elutriation is represented as 0–390 m and the difference between the time points is represented as 30 m.
- Figure 5.4 shows the Representation of Tri$_{\text{Optimal}}$ for the Data$_{\text{Elutriation}}$. Table 5.7 shows the highly correlated genes for the Data$_{\text{Elutriation}}$. Data$_{\text{Elutriation}}$ from YCC: Biological Significant for Biological Process shown in Table 5.8.

5.4.3 Pheromone Experiment Using 3D GED

- The YCC data, Greedy$_{\text{TriPSO}}$, is able to extract highly correlated genes with larger volumes of genes. From the input list 3297 genes are known and not ambiguous. Also, 270 duplicates were removed from your input list. The removed genes are

Table 5.6 $Data_{CDC15}$ from YCC: biological significant for biological process

Biological process				Molecular function				Cellular component			
Gene Ontology term	Cluster frequency with Percentage	Genome frequency with Percentage	Corrected *P*-value	Gene Ontology term	Cluster frequency with Percentage	Genome frequency with Percentage	Corrected *P*-value	Gene Ontology term	Cluster frequency with Percentage	Genome frequency with Percentage	Corrected *P*-value
Cellular component organization	1120 of 3333: genes, 33.6%	2087 of 7166 genes, 29.1%	1.57E-11	Catalytic activity	1258 of 3333 genes, 37.7%	2434 of 7166 genes, 34.0%	2.05E-07	Organelle	2409 of 3333 genes, 72.3%	4824 of 7166 genes, 67.3%	2.77E-14
Biological regulation	1099 of 3333 genes, 33.0%	2055 of 7166 genes, 28.7%	1.42E-10	Ion binding	851 of 3333 genes, 25.5%	1603 of 7166 genes, 22.4%	1.44E-06	Intracellular organelle	2408 of 3333 genes, 72.2%	4822 of 7166 genes, 67.3%	2.90E-14
Intracellular membrane-bounded organelle	2225 of 3333 genes, 66.8%	4467 of 7166 genes, 62.3%	2.53E-10	Nucleotide binding	478 of 3333 genes, 14.3%	879 of 7166 genes, 12.3%	0.00041	Membrane-bounded organelle	2255 of 3333 genes, 67.7%	4526 of 7166 genes, 63.2%	7.79E-11
Cell	2831 of 3333: genes-, 84.9%	5834 of 7166: genes-, 81.4%	3.08E-10	Nucleoside phosphate binding	478 of 3333 genes, 14.3%	879 of 7166 genes, 12.3%	0.00041	Intracellular membrane-bounded organelle	2225 of 3333 genes, 66.8%	4467 of 7166 genes, 62.3%	2.53E-10
Cell part	2828 of 3333 genes, 84.8%	5829 of 7166 genes, 81.3%	4.50E-10	Small molecule binding	511 of 3333 genes, 15.3%	948 of 7166 genes, 13.2%	0.00067	Cell	2831 of 3333 genes, 84.9%	5834 of 7166 genes, 81.4%	3.08E-10
Membrane	1077 of 3333 genes, 32.3%	2070 of 7166 genes, 28.9%	1.12E-06	Anion binding	513 of 3333 genes, 15.4%	966 of 7166 genes, 13.5%	0.00677	Cell part	2828 of 3333 genes, 84.8%	5829 of 7166 genes, 81.3%	4.50E-10

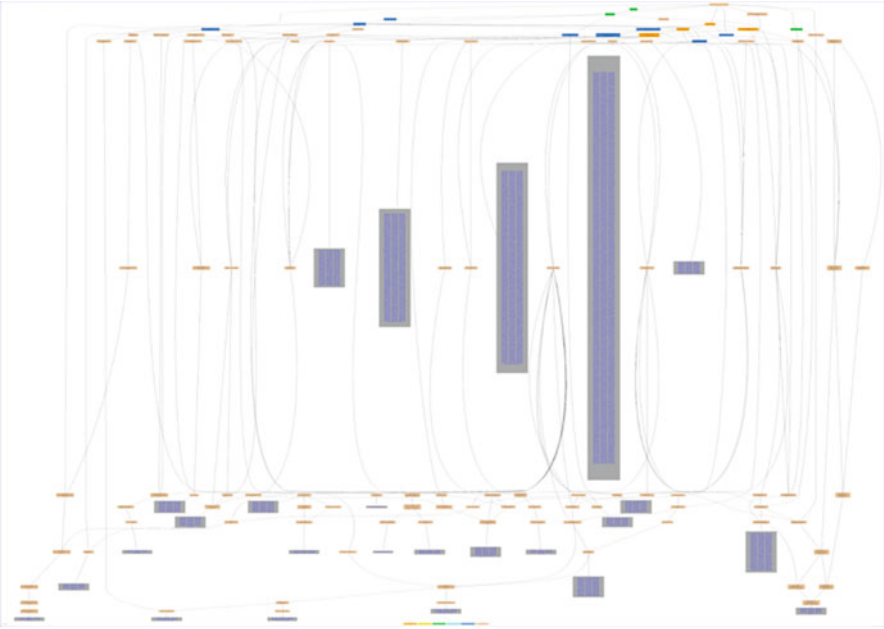

Fig. 5.4 Representation of $\text{Tri}_{\text{Optimal}}$ for the $\text{Data}_{\text{Elutriation}}$

identified to be either unknown or ambiguous in the dataset. 7166 genes in all were utilized to determine the background distribution of GO keywords. Forty seven terms are being displayed out of the 3534 total discovered.

- The total number of $\text{Tri}_{\text{Timepoint}}$ in the pheromone experiment is represented as 000–119 m and the difference between the time points is represented as 007 m.
- Figure 5.5 shows the Representation of $\text{Tri}_{\text{Optimal}}$ for the $\text{Data}_{\text{Pheromone}}$. Table 5.9 shows the Highly Correlated Genes for the $\text{Data}_{\text{Pheromone}}$.
- The Biological Significant of the Biological Process for $\text{Data}_{\text{Pheromone}}$ from YCC shown in Table 5.10.

5.5 Summary

This chapter introduces a new algorithm for extracting tricluster from 3D GED using artificial intelligence methods. K-Means$_{\text{TwoWay}}$ clustering to create tricluster seeds. The greedy triclustering method is then used to enlarge these seeds. The primary goal of the optimization issue is to extract the highly correlated tricluster that has a bigger volume. The greedy triclustering method, which is based on the objective

Table 5.7 Highly correlated genes for the $Data_{Elutriation}$

YOR361C	YDR144C	YOL083W	YKL033W	YIR019C	YER013W
YDR099W	YKL074C	YOL009C	YBR034C	YHR086W	YDR171W
YAR003W	YDR225W	YNL264C	YMR183C	YLR424W	YDR060W
YHR038W	YMR078C	YDR273W	YBL055C	YKL049C	YMR002W
YIL095W	YPL242C	YIR018W	YGR280C	YOR189W	YPL152W
YJR021C	YML058W	YHR132C	YPR175W	YIL011W	YDR372C
YHR004C	YGR240C	YGL098W	YDR093W	YDR469W	YML015C
YGL112C	YIL176C	YBR234C	YKL117W	YFL013C	YPR016C
YGR074W	YPL184C	YGR105W	YJL164C	YKL164C	YPL047W
YJR033C	YJL100W	YML019W	YPL138C	YDR334W	YBL035C
YBR275C	YMR086W	YJL058C	YOR094W	YDR043C	YMR268C
YLL038C	YOR035C	YBR044C	YPL133C	YPR161C	YHR186C
YGL172W	YLR163C	YKR010C	YHL025W	YDR493W	YPL127C
YHR181W	YHR205W	YIL126W	YDR528W	YNL258C	YGR131W
YHL030W	YKL058W	YBR214W	YAL054C	YKL050C	YLR090W
YHL024W	YJL036W	YOL047C	YMR231W	YGL207W	YLR433C
YER122C	YNR031C	YGL194C	YKR054C	YLR200W	YDR359C
YGL059W	YKL156W	YNL138W	YER155C	YPR135W	YLR343W
YNL312W	YDR208W	YKL013C	YLR038C	YIR041W	YOR295W
YDL064W	YGL169W	YBL067C	YMR037C	YJR066W	YBR152W
YGR136W	YPL120W	YDL223C	YHR081W	YOR246C	YJL021C
YNL216W	YPR048W	YGR150C	YPL250C	YDL120W	YOR293W
YDR485C	YOR147W	YDR517W	YIR021W	YGL189C	YFL024C
YLR249W	YNL182C	YPL211W	YML046W	YKL145W	YPR041W

function, may extract highly linked genes with significant volume. The proposed $Greedy_{Tri}$ is tested on YCC datasets using GO ontology tool. The findings of the GO validation are highly correlated and the levels of significance for the retrieved terms are high. The biological mechanism, molecular function, and cellular component of triclusters have all been graphically shown. It is observed that $Greedy_{Tri}$ is capable to extract highly correlated tricluster with larger volume and biological significance also. Triclusters with scaling pattern have still more biological significance. Above all, triclusters with high correlation degree have more biological significance which is the finding of this work.

Table 5.8 $Data_{Elutriation}$ from YCC: biological significant for biological process

Biological process				Molecular function				Cellular component			
Gene Ontology term	Cluster frequency with Percentage	Genome frequency with Percentage	Corrected *P*-value	Gene Ontology term	Cluster frequency with Percentage	Genome frequency with Percentage	Corrected *P*-value	Gene Ontology term	Cluster frequency with Percentage	Genome frequency with Percentage	Corrected *P*-value
Biological regulation	1091 of 3308 genes, 33.0%	2055 of 7166 genes, 28.7%	1.96E-10	Catalytic activity	1250 of 3308 genes, 37.8%	2434 of 7166 genes, 34.0%	1.72E-07	Intracellular organelle	2359 of 3308 genes, 71.3%	4822 of 7166 genes, 67.3%	7.89E-09
Regulation of biological process	903 of 3308 genes, 27.3%	1690 of 7166 genes, 23.6%	1.58E-08	Ion binding	836 of 3308 genes, 25.3%	1603 of 7166 genes, 22.4%	3.26E-05	Membrane-bounded organelle	2226 of 3308 genes, 67.3%	4526 of 7166 genes, 63.2%	8.34E-09
Regulation of cellular process	845 of 3308 genes, 25.5%	1579 of 7166 genes, 22.0%	7.22E-08	Coenzyme binding	94 of 3308 genes, 2.8%	144 of 7166 genes, 2.0%	0.0027	Organelle	2359 of 3308 genes, 71.3%	4824 of 7166 genes, 67.3%	1.07E-08
Cellular component organization	1079 of 3308 genes, 32.6%	2087 of 7166 genes, 29.1%	3.53E-06					Intracellular membrane-bounded organelle	2190 of 3308 genes, 66.2%	4467 of 7166 genes, 62.3%	1.76E-07
Regulation of macromolecule metabolic process	602 of 3308 genes, 18.2%	1122 of 7166 genes, 15.7%	9.36E-05					Endomembrane system	574 of 3308 genes, 17.4%	1091 of 7166 genes, 15.2%	0.00169
Regulation of metabolic process	634 of 3308 genes, 19.2%	1192 of 7166 genes, 16.6%	0.00021					Nucleus	1221 of 3308 genes, 36.9%	2458 of 7166 genes, 34.3%	0.00751

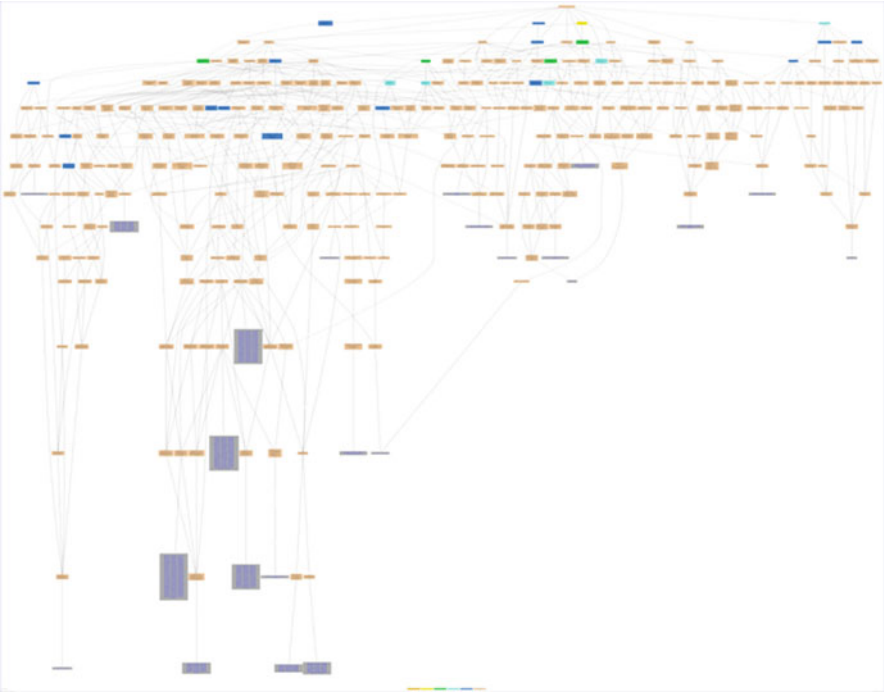

Fig. 5.5 Representation of $Tri_{Optimal}$ for the $Data_{Pheromone}$

Table 5.9 Highly correlated genes for the $Data_{Pheromone}$

YHR102W	YEL077C	YAL002W	YPR049C	YLR356W	YCL031C
YBR202W	YKL190W	YGR095C	YOR058C	YDR309C	YGR133W
YGR103W	YCL056C	YPL074W	YOL109W	YLR399C	YKL014C
YPR179C	YOR152C	YDR356W	YMR190C	YBL031W	YBL043W
YDR069C	YOR330C	YBR135W	YPL029W	YAL029C	YGR218W
YPL004C	YOR224C	YGL049C	YBL061C	YLR067C	YLL011W
YDR295C	YAL007C	YLR168C	YJL180C	YIL041W	YNL084C
YHR087W	YLL026W	YGR239C	YNL218W	YHR039C	YNL021W
YOR056C	YOL126C	YLR461W	YDL195W	YPR194C	YNL006W
YML072C	YIL062C	YBL006C	YBL018C	YNL121C	YGL120C
YNL039W	YIL132C	YDL139C	YMR260C	YHL027W	YKL137W
YLR276C	YDR021W	YIR010W	YMR265C	YDR440W	YNL075W
YGL115W	YNL232W	YEL056W	YHR160C	YBR114W	YKL087C
YOR210W	YJR002W	YDR319C	YJR151C	YDL093W	YPR187W
YDR065W	YAL048C	YHR041C	YDR486C	YOR073W	YNL066W
YDR397C	YBR017C	YLR341W	YAL059W	YER106W	YBR237W
YOR099W	YKL192C	YMR163C	YMR211W	YER169W	YNL310C
YLR006C	YDR150W	YDR449C	YBR091C	YML010W	YDR437W
YBL078C	YGR145W	YAL013W	YNL110C	YDR224C	YCL034W
YBL068W	YBR043C	YIL118W	YNL313C	YDR229W	YOR228C
YIL146C	YDR132C	YBR186W	YNL131W	YLL027W	YNL206C

Table 5.10 $Data_{Pheromone}$ from YCC: biological significant for biological process

Biological process				Molecular function				Cellular component			
Gene Ontology term	Cluster frequency with Percentage	Genome frequency with Percentage	Corrected *P*-value	Gene Ontology term	Cluster frequency with Percentage	Genome frequency with Percentage	Corrected *P*-value	Gene Ontology term	Cluster frequency with Percentage	Genome frequency with Percentage	Corrected *P*-value
Cellular component organization	1122 of 3291 genes, 34.1%	2087 of 7166 genes, 29.1%	3.41E-14	Catalytic activity	1263 of 3291 genes, 38.4%	2434 of 7166 genes, 34.0%	2.65E-10	Intracellular organelle	2431 of 3291 genes, 73.9%	4822 of 7166 genes, 67.3%	2.27E-25
Biological regulation	1101 of 3291 genes, 33.5%	2055 of 7166 genes, 28.7%	4.03E-13	Ion binding	852 of 3291: genes, 25.9%	1603 of 7166 genes, 22.4%	3.30E-08	Organelle	2431 of 3291 genes, 73.9%	4824 of 7166 genes, 67.3%	3.71E-25
Organelle organization	817 of 3291 genes, 24.8%	1496 of 7166 genes, 20.9%	8.37E-11	Protein binding	509 of 3291 genes, 15.5%	929 of 7166 genes, 13.0%	4.68E-06	Membrane-bounded organelle	2282 of 3291 genes, 69.3%	4526 of 7166 genes, 63.2%	5.89E-21
Regulation of biological process	911 of 3291 genes, 27.7%	1690 of 7166 genes, 23.6%	1.19E-10	Enzyme activator activity	90 of 3291 genes, 2.7%	136 of 7166 genes, 1.9%	0.00143	Intracellular membrane-bounded organelle	2253 of 3291 genes, 68.5%	4467 of 7166 genes, 62.3%	2.41E-20
Regulation of cellular process	853 of 3291 genes, 25.9%	1579 of 7166 genes, 22.0%	6.27E-10	Metal ion binding	443 of 3291 genes, 13.5%	830 of 7166 genes, 11.6%	0.00334	Cell	2796 of 3291 genes, 85.0%	5834 of 7166 genes, 81.4%	4.28E-10
Cellular component organization or biogenesis	1260 of 3291 genes, 38.3%	2447 of 7166 genes, 34.1%	2.14E-08	Cation binding	447 of 3291 genes, 13.6%	841 of 7166 genes, 11.7%	0.00541	Cell part	2793 of 3291 genes, 84.9%	5829 of 7166 genes, 81.3%	6.33E-10

References

Gnatyshak DV (2014) Greedy modifications of OAC-triclustering algorithm. Proc Comput Sci 31: 1116–1123. https://doi.org/10.1016/j.procs.2014.05.367

Gnatyshak DV (2015) A single-pass triclustering algorithm. Automat Doc Math Ling 49(1):27–41. https://doi.org/10.3103/s0005105515010057

Han F, Yang C, Wu Y, Zhu J, Ling Q, Song Y, Huang D (2017) A gene selection method for microarray data based on binary PSO encoding gene-to-class sensitivity information. IEEE/ACM Trans Comput Biol Bioinform 14(1):85–96. https://doi.org/10.1109/tcbb.2015.2465906

Rathipriya R, Thangavel K, Bagyamani J (2011) Binary particle swarm optimization based biclustering of web usage data. Int J Comput Appl 25(2):43–49. https://doi.org/10.5120/3001-4036

Li Y, Tian X, Jiao L, Zhang X (2014) Biclustering of gene expression data using Particle Swarm Optimization integrated with pattern-driven local search. In: 2014 IEEE congress on evolutionary computation (CEC). https://doi.org/10.1109/cec.2014.6900323

Pontes B, Girldez R, Aguilar-Ruiz JS (2015) Quality measures for gene expression biclusters. PLoS One 10(3):e0115497. https://doi.org/10.1371/journal.pone.0115497

Balamurugan R, Natarajan A, Premalatha K (2016) Biclustering microarray gene expression data using modified Nelder-Mead method. Int J Inf Commun Technol 9(1):43. https://doi.org/10.1504/ijict.2016.077686

Wei Shen GL (2012) A novel biclustering algorithm and its application in gene expression profiles. J Inf Comput Sci 9(11):3113–3122

Liu F (2006) Biclustering of gene expression data using EDA-GA hybrid. In: IEEE international conference on evolutionary computation. https://doi.org/10.1109/CEC.2006.1688499

Xie B, Chen S, Liu F (2007) Biclustering of gene expression data using PSO-GA hybrid. In 2007 1st international conference on bioinformatics and biomedical engineering. https://doi.org/10.1109/icbbe.2007.81

Shyama D, Idicula SM (2010) Greedy search-binary PSO hybrid for biclustering gene expression data. Int J Comput Appl 2(3):1–5. https://doi.org/10.5120/651-908

Chuang L, Chang H, Tu C, Yang C (2008) Improved binary PSO for feature selection using gene expression data. Comput Biol Chem 32(1):29–38. https://doi.org/10.1016/j.compbiolchem.2007.09.005

Lopamudra Dey AM (2014) Microarray gene expression data clustering using PSO based K-means algorithm. Int J Comput Sci Appl 1(1):232–236. issn: 2250-3765

Thangavel K, Bagyamani J, Rathipriya R (2012) Novel hybrid PSO-SA model for biclustering of expression data. Proc Eng 30:1048–1055. https://doi.org/10.1016/j.proeng.2012.01.962

Narmadha N, Rathipriya R (2019a) Greedy K-means: two way clustering for optimal coherent tricluster. Int J Sci Technol Res 8(10):1916–1921. issn:2277-8616

Narmadha N, Rathipriya R (2018) Triclustering algorithm for 3D gene expression data using correlation measure. Int J Eng Res Comput Sci Eng 5(2):221–227

Narmadha N (2017) Query based tri-clustering (QBTC). Int J Comput Intell Inform 7(2):117–121

Narmadha N, Rathipriya R (2019b) Gene ontology analysis of 3D microarray gene expression data using hybrid PSO optimization. Int J Innov Technol Explo Eng 8(11):3890–3896

Narmadha N, Rathipriya R (2020) An optimized three-dimensional clustering for microarray data. In: Handbook of research on big data clustering and machine learning. IGI Global, pp 366–377