UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

Three-Dimensional Clustering Method on Gene Expression Dataset Using the Gene Cube Approach

Almaira Nabila Ayudhiya | Paper ID: 210124
Dra. Saskya Mary Soemartojo, M.Si
Dr. Dra. Titin Siswantining, D.E.A

Outline

Introduction

Introduction

Data Mining

Source: vecteezy.com

- Converting large data sizes into useful information.
- Clustering is one of the methods used in data mining.

Reference: Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concept and techniques.* ProQuest Ebook Central.

Introduction

Clustering vs Biclustering

Clustering

- Grouping only the observation (rows) or attributes (columns) dimension only.
- Find the observation group on all attributes.

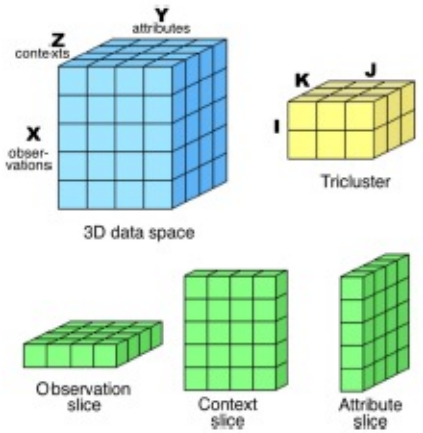
Biclustering

- Grouping the observation and attributes dimensions together.
- Find the observation group on several attributes.

Reference: Madeira, S.C., & Oliveira, A.L. (2004). *Biclustering algorithms for biological data analysis: a survey.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 1, No. 1, pp.24-45.

Introduction

Triclustering



Source: Henriques, 2018

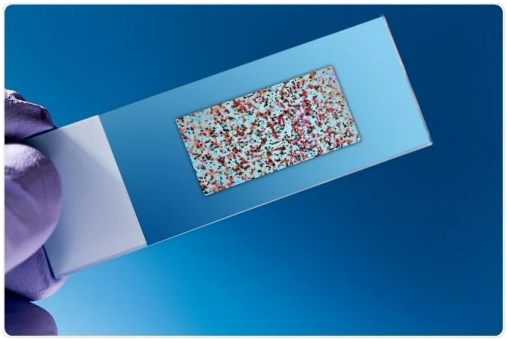
Grouping the observation, attributes, and context dimensions simultaneously.

Generates a subspace consisting of a subset of observations, a subset of attributes, and a subset of context.

Reference: Henriques, R., & Madeira, S. C. (2018). *Triclustering Algorithms for Three-Dimensional Data Analysis: A Comprehensive Survey.* ACM Computing Surveys, Vol. 51, No. 5, Article 95.

Introduction

Microarray Gene Expression Data



Measures the level of expression of thousands of genes simultaneously under certain conditions.

Producing data arranged in a numeric matrix known as an expression matrix.

Each element of this data matrix indicates the level of numerical expression of a gene under certain conditions.

Reference: Baldi, P., & Hatfield, G. W. (2011). *DNA Microarrays and Gene Expression – From Experiments to Data Analysis and Modeling*. Cambridge University Press, pp. 1-213.

Introduction

Triclustering Applications

- Microarray time-series data with dimension Time × Gene × Observation
- Data from different species with dimension Gene × Observation × Organism
- Chemical data from protein structure with dimension Residue × Position × Time
- and many more…

none have yet taken into account the role of chromosomes.

Reference: Henriques, R., & Madeira, S. C. (2018). *Triclustering Algorithms for Three-Dimensional Data Analysis: A Comprehensive Survey.* ACM Computing Surveys, Vol. 51, No. 5, Article 95.

Introduction

Chromosome

Why it is necessary to take into account chromosomes?

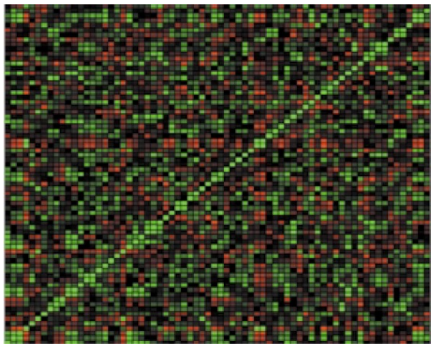
- Expression of gene data is controlled by **regulatory element** which can be located alongside a chromosome, in some cases, even located on other chromosomes. Said regulatory elements are proteins produced by a gene regulator, namely genes whose expression products play a role in regulating the expression of other genes.
- **Chromosome mapping** has proven to be a promising method in finding expression patterns between genes (Cohen, 2000).

Source: news-medical.net

Reference: Cohen, B. A., Mitra, R. D., Hughes, J. D., & Church, G. M. (2000). *A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression*. Nature Genetics, Vol. 26, pp. 183-186.

Introduction

Chromosome Mapping



Source: Cohen, 2000

Map genes to their chromosome region.

If correlation exists between genes, then this correlation will be seen from the location of the gene in the chromosome region.

Genes that have a similar expression pattern tend to be in close positions along the chromosome.

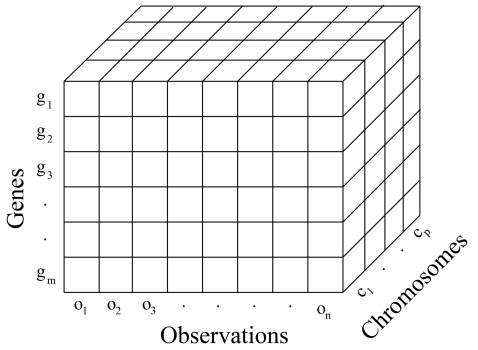
Reference: Cohen, B. A., Mitra, R. D., Hughes, J. D., & Church, G. M. (2000). *A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression*. Nature Genetics, Vol. 26, pp. 183-186.

Introduction

Gene Cube

The formation of a three-dimensional structure with the form of Chromosomes x Genes x Observations.

The K-means algorithm which initial steps are optimized using the K-means++ algorithm and δ-Trimax triclustering are applied.



Materials and Method

- Dataset
- K-Means Clustering Method
- K-Means++ Clustering Algorithm
- Davies Bouldin Index
- δ-Trimax Triclustering Algorithm
- Tricluster Diffusion (TD)

Dataset

Gene ID	Gene Symbol	Chromosome	Observation				
			Control 1	Control 2	Control 3	...	Bladder Cancer T2+ 3
1007_s_at	DDR1	6	8,518418	8,250695	8,769348	...	9,611918
1053_at	RFC2	7	5,233081	5,158875	4,923864	...	5,918295
117_at	HSPA6	1	5,722707	5,935722	4,730148	...	5,12154
121_at	PAX8	2	6,165746	6,272942	6,159483	...	6,272729
1255_g_at	GUCA1A	6	2,996766	2,881827	2,778352	...	2,853586

Source: *Gene Expression Omnibus* (GEO) – *National Center for Biotechnology Information* (NCBI)

Gene expression of bladder cancer dataset

Consists of gene expression profile on 3 normal (control) bladder tissue observations, 3 Ta tumor observations, 3 T1 tumor observations, and 3 T2+ tumor observations

Consists of 12 observations, each of which consisted of 45,746 genes

K-Means Clustering Method

1. Select k observations as the initial centroids.
2. Calculating the distance between each observation x_i to the centroid using the Euclidean distance.
3. Grouping observations into the closest centroid.
4. Determine the new centroid by calculating the average of observations in each cluster.
5. Repeating step 2, 3, and 4 until the object no longer moves to another cluster.
6. Algorithm ends.

Reference: Wang, J., & Su, X. (2011). *An improved K-means clustering algorithm*. 2011 IEEE 3rd International Conference on Communication Software and Networks, pp. 44-46.

K-Means++ Clustering Algorithm

1 Choose one observation from the data randomly. The selecter observation is the initial centroid and is denoted as c_1.

2 Calculate the distance between each observation to c_1. The distance between the mth observation (x_m) and the jth centroid (c_j) is denoted by $d(x_m, c_j)$.

3 Choose the next centroid, c_2, randomly with probability

$$\frac{d^2(x_m, c_1)}{\sum_{j=1}^{n} d^2(x_j, c_1)}$$

4 To select the jth centroid (c_j):

- Compute the distance between each observation and each centroid and assigns each observation to the nearest centroid.
- For $m = 1, \dots, n$ and $p = 1, \dots, j-1$, choose the jth centroid randomly with probability

$$\frac{d^2(x_m, c_p)}{\sum_{\{h; x_h \in C_p\}} d^2(x_h, c_p)}$$

where C_p is the set of all observations closest to c_p.

5 Repeat step 4 until k centroids have been selected.

Reference: Arthur, D., & Vassilvitskii, S. (2007). *K-Means++: The Advantages of Careful Seeding*. Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms, Vol. 8, pp. 1027-1035.

Davies Bouldin Index

Cohesion

Sum of the data proximity to the centroid of the cluster.

$$SSW_i = \frac{1}{m_i}\sum_{j=1}^{m_i} d(x_j, c_i)$$

Separation

Distance between the centroids of the cluster.

$$SSB_{i,j} = d(c_i, c_j)$$

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}}$$

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j}(R_{i,j})$$

Reference: Davies, D., & Bouldin, D. (1979). *A Cluster Separation Measure*. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1, Vol. 2, pp. 224-227.

δ-Trimax Triclustering Method

- Bhar (2013) previously applied the δ-Trimax triclustering algorithm on the dimensions of time, genes, and observations.

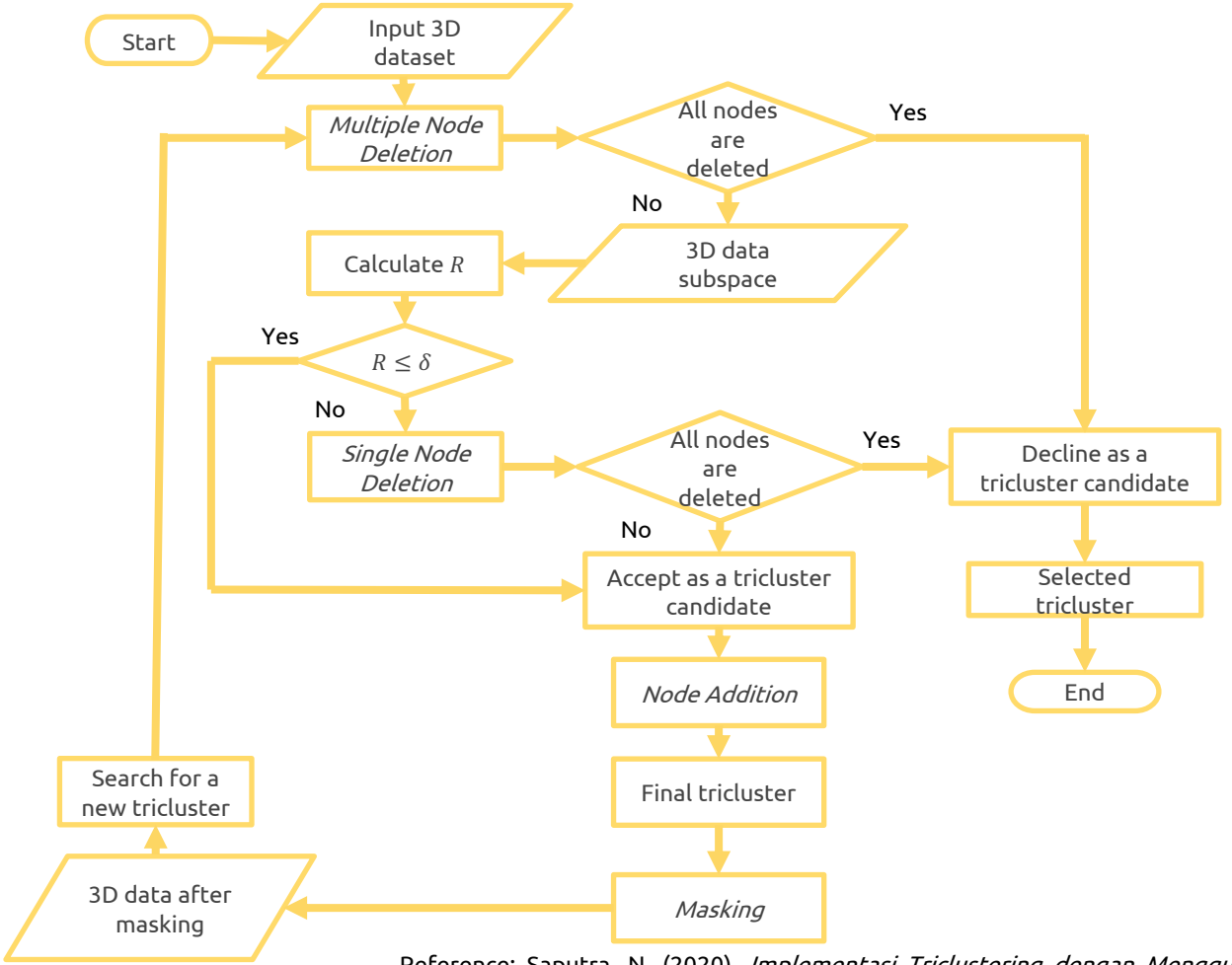
- The time dimension was replaced by the chromosome dimension.

Mean Squared Residual

$$R = \frac{1}{|\boldsymbol{G}||\boldsymbol{O}||\boldsymbol{C}|} \sum_{g \in \boldsymbol{G}, o \in \boldsymbol{O}, c \in \boldsymbol{C}} \left(m_{goc} - m_{gOC} - m_{GoC} - m_{GOc} + 2m_{GOC}\right)^2$$
$$= \frac{1}{|\boldsymbol{G}||\boldsymbol{O}||\boldsymbol{C}|} \sum_{g \in \boldsymbol{G}, o \in \boldsymbol{O}, c \in \boldsymbol{C}} r_{goc}^2$$

Reference: Bhar, A., *et al.* (2013). *Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell.* Algorithms for Molecular Biology, Vol. 8, No. 9.

δ-Trimax Algorithm



Reference: Saputra, N. (2020). *Implementasi Triclustering dengan Menggunakan Metode δ-Trimax pada Data Ekspresi Gen Microarray*. Skripsi. Universitas Indonesia.

Tricluster Diffusion

$$TD_i = \frac{MSR_i}{Volume_i} = \frac{MSR_i}{|\boldsymbol{G}_i||\boldsymbol{O}_i||\boldsymbol{C}_i|}$$

MSR_i = mean squared residual of the ith tricluster

$|\boldsymbol{G}_i|$ = number of dimensions of genes in the ith tricluster

$|\boldsymbol{O}_i|$ = number of dimensions of observations in the ith tricluster

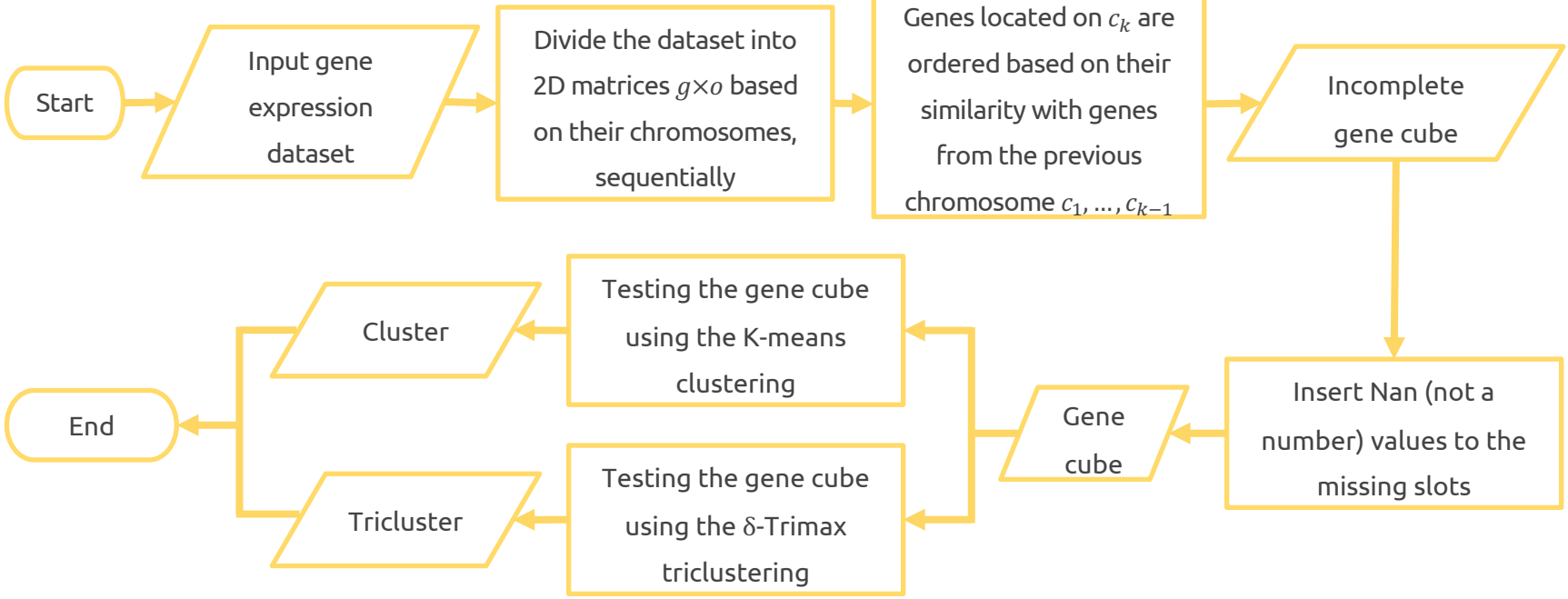
$|\boldsymbol{C}_i|$ = number of dimensions of chromosomes in the ith tricluster

Reference: Bhar, A., *et al.* (2015). *Multiobjective triclustering of time-series transcriptome data reveals key genes of biological processes*. BMC Bioinformatics, Vol. 16, No. 200.

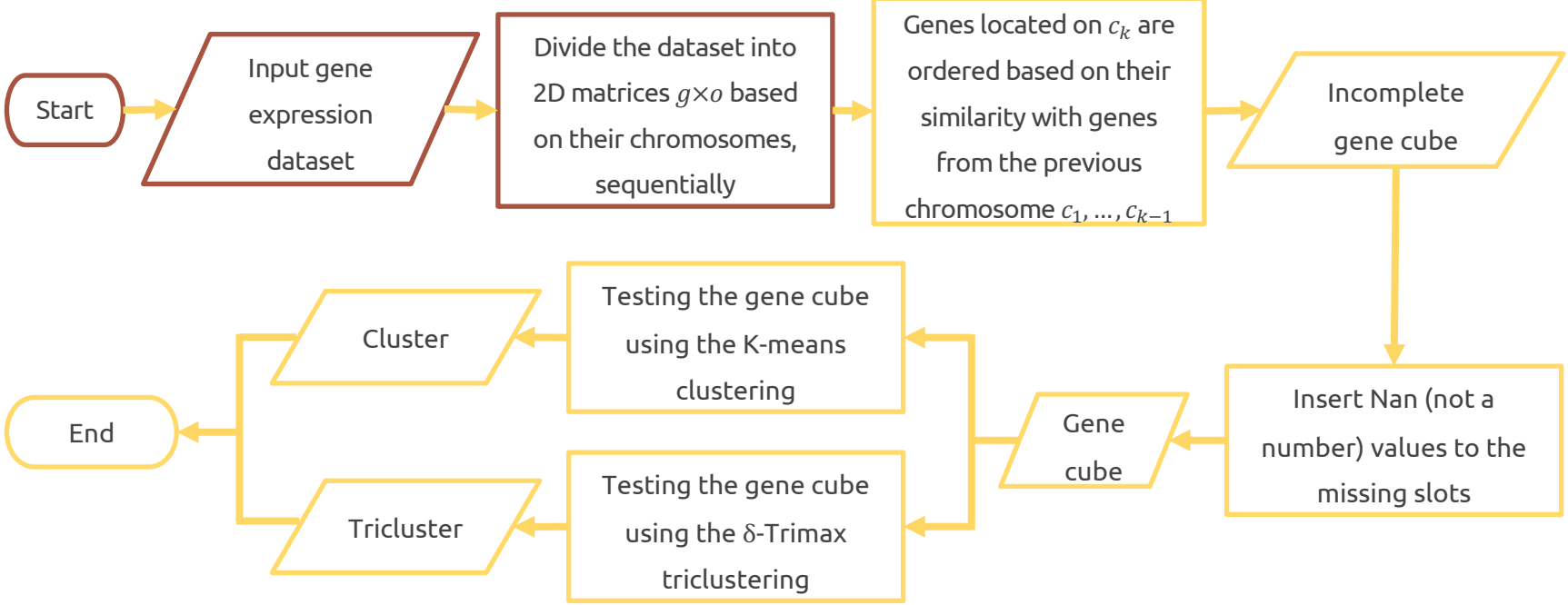
Results and Discussion

- Formation of the Gene Cube
- Results
- Discussion

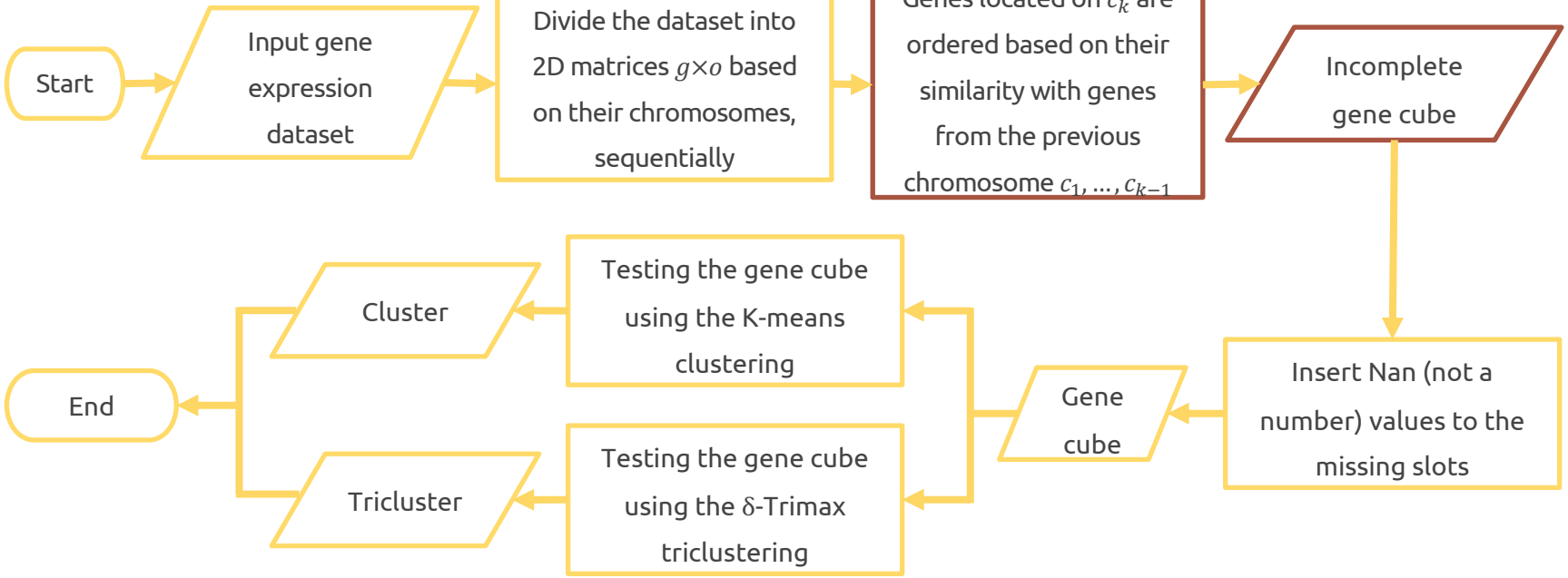
Formation of the Gene Cube



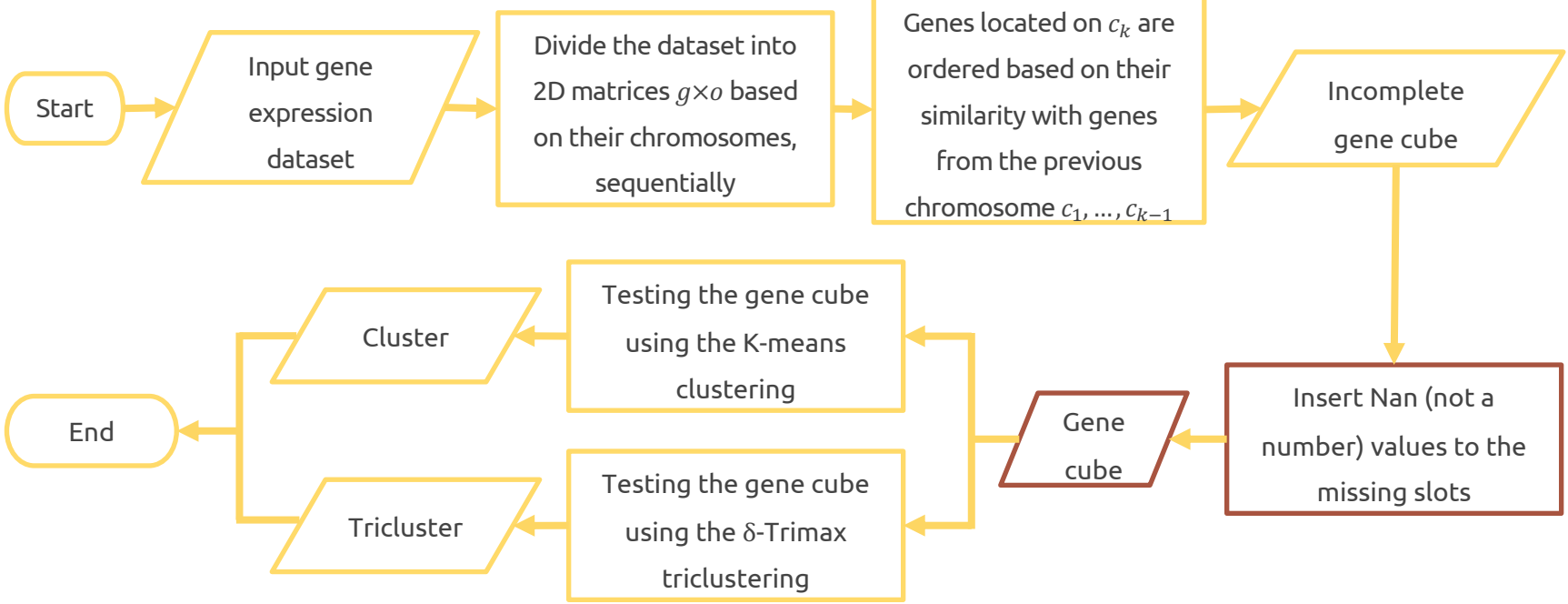
Formation of the Gene Cube



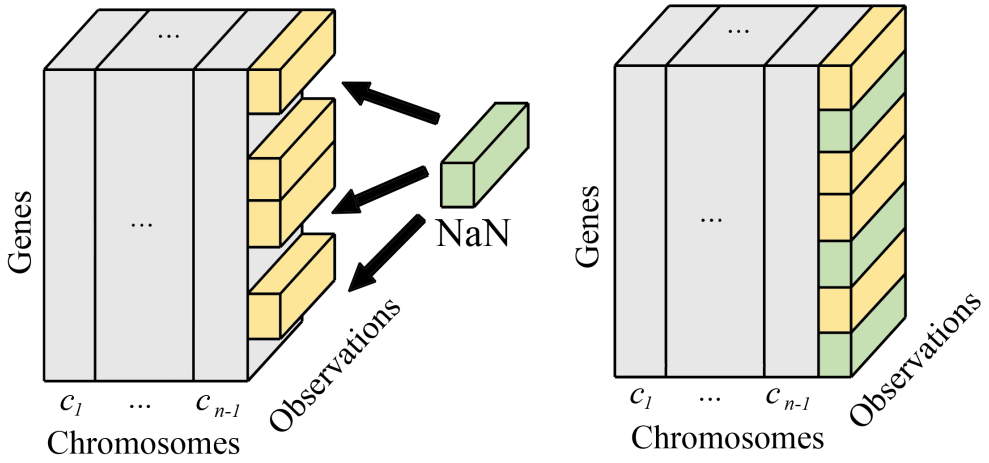
Formation of the Gene Cube



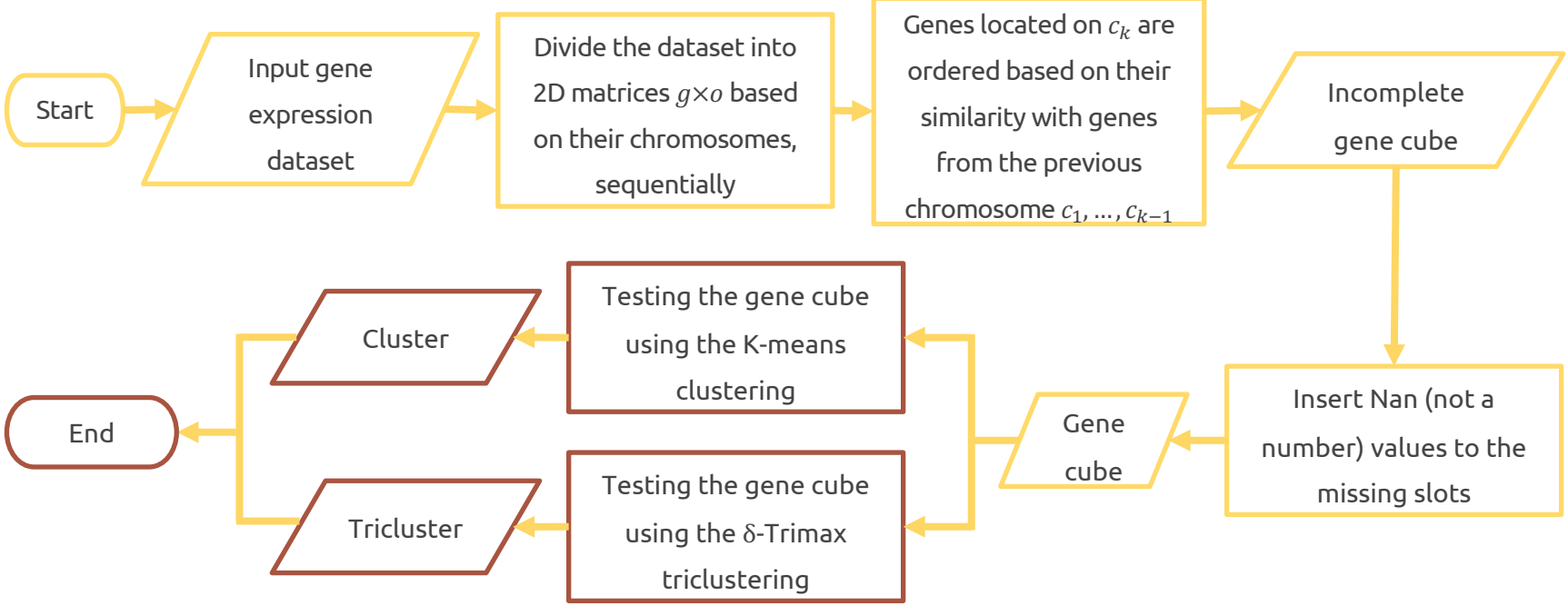
Formation of the Gene Cube



Illustration



Formation of the Gene Cube



Formation of the Gene Cube

Chromosome	Number of Genes
Chromosome 1	4510 genes
Chromosome 2	3184 genes
Chromosome 3	2648 genes
Chromosome 4	1838 genes
Chromosome 5	2141 genes
Chromosome 6	2551 genes

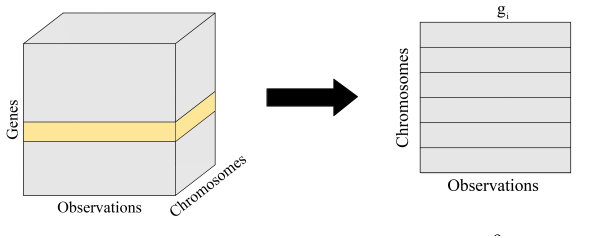
Chromosome	Number of Genes
Chromosome 7	2272 genes
Chromosome 8	1661 genes
Chromosome 9	1812 genes
Chromosome 10	1798 genes
Chromosome 11	2487 genes
Chromosome 12	2410 genes

Chromosome	Number of Genes
Chromosome 13	987 genes
Chromosome 14	1528 genes
Chromosome 15	1492 genes
Chromosome 16	1852 genes
Chromosome 17	2527 genes
Chromosome 18	806 genes

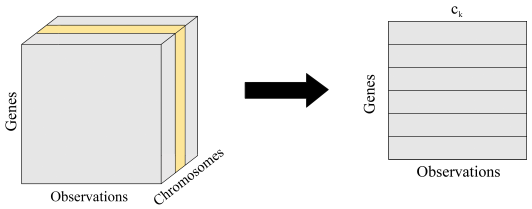
Chromosome	Number of Genes
Chromosome 19	2668 genes
Chromosome 20	1281 genes
Chromosome 21	600 genes
Chromosome 22	1093 genes
Chromosome 23 (X)	1512 genes
Chromosome 24 (Y)	88 genes

Results

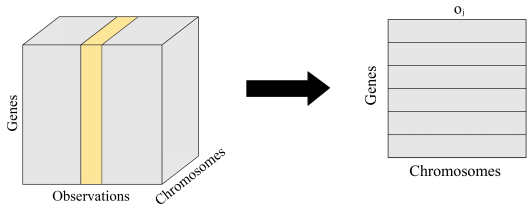
Clustering Analysis Using the K-Means Algorithm on the Gene Cube



Cross section across the gene axis



Cross section across the chromosome axis

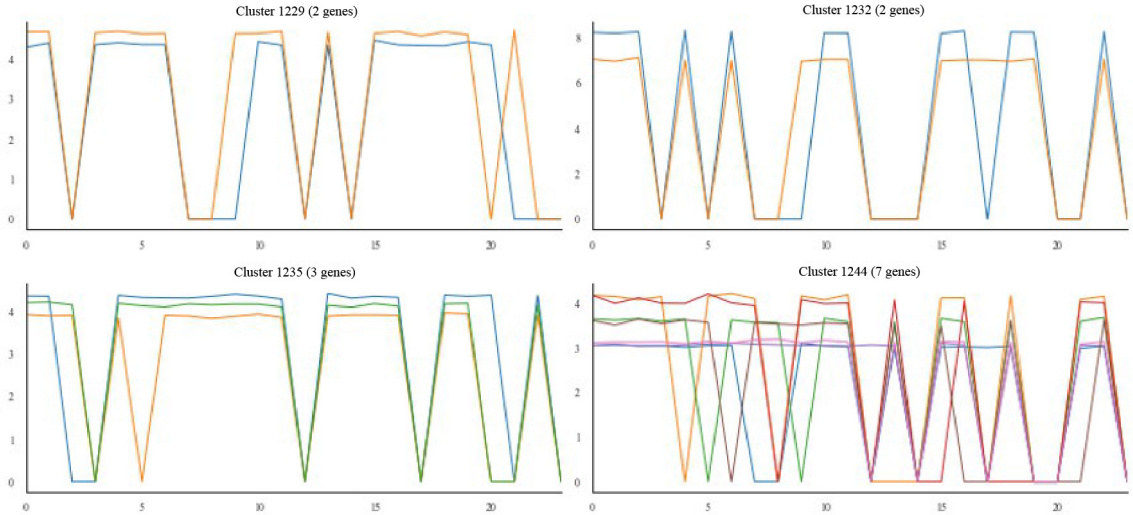


Cross section across the observation axis

Results

K-Means Clustering Across the Gene Axis → dimension = 24 x 12

4 clusters

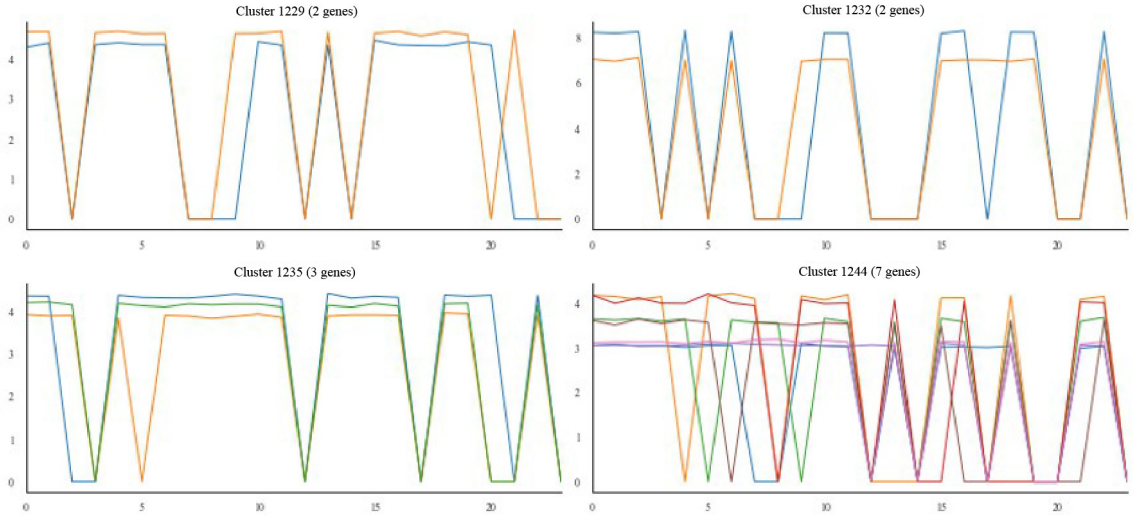


DBI = 0,33062

Results

K-Means Clustering Across the Observation Axis → dimension = 4510 x 24

2500 clusters

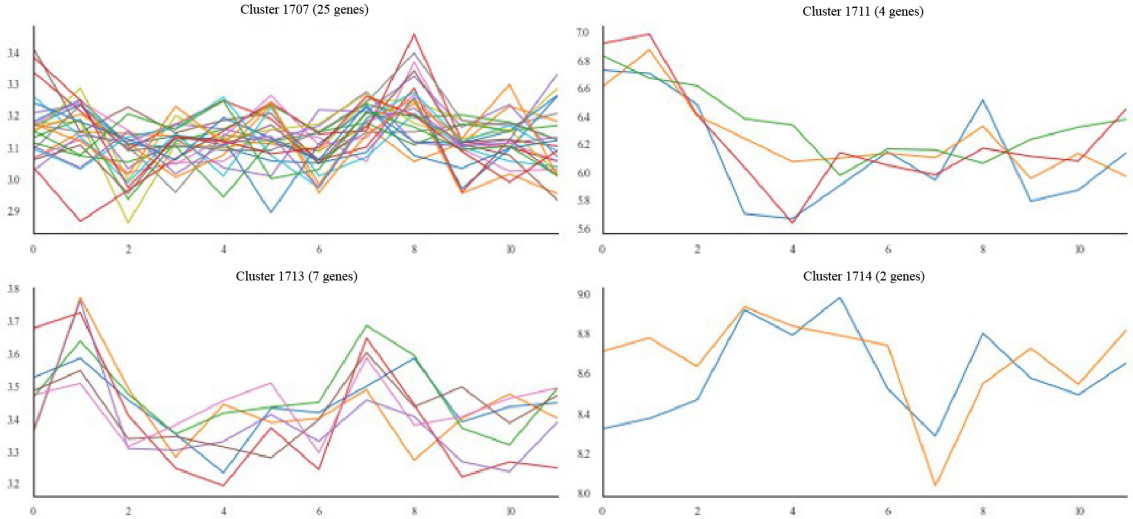


DBI = 0,63860

Results

K-Means Clustering Across the Chromosome Axis → dimension = 4510 x 12

1800 clusters



DBI = 0,66134

Results

Triclustering Analysis Using the δ-Trimax on the Gene Cube

Simulation Design

Simulation 1: $\delta = 0{,}0466$ and $\lambda = 1{,}05$

Simulation 2: $\delta = 0{,}0566$ and $\lambda = 1{,}05$

Simulation 3: $\delta = 0{,}0732$ and $\lambda = 1{,}05$

Simulation 4: $\delta = 0{,}0765$ and $\lambda = 1{,}05$

Simulation 5: $\delta = 0{,}0817$ and $\lambda = 1{,}05$

Simulation 6: $\delta = 0{,}0466$ and $\lambda = 1{,}15$

Simulation 7: $\delta = 0{,}0566$ and $\lambda = 1{,}15$

Simulation 8: $\delta = 0{,}0732$ and $\lambda = 1{,}15$

Simulation 9: $\delta = 0{,}0765$ and $\lambda = 1{,}15$

Simulation 10: $\delta = 0{,}0817$ and $\lambda = 1{,}15$

Simulation 11: $\delta = 0{,}0466$ and $\lambda = 1{,}25$

Simulation 12: $\delta = 0{,}0566$ and $\lambda = 1{,}25$

Simulation 13: $\delta = 0{,}0732$ and $\lambda = 1{,}25$

Simulation 14: $\delta = 0{,}0765$ and $\lambda = 1{,}25$

Simulation 15: $\delta = 0{,}0817$ and $\lambda = 1{,}25$

Results

Comparison of Simulation Results

Simulation	λ	δ	Average TD Value
1	1,05	0,0466	1,88507E-07
2		0,0566	3,10619E-06
3		0,0732	1,67872E-07
4		0,0765	1,77406E-07
5		0,0817	2,08116E-07

Simulation	λ	δ	Average TD Value
6	1,15	0,0466	2,23432E-07
7		0,0566	1,90620E-06
8		0,0732	1,65044E-07
9		0,0765	1,52081E-07
10		0,0817	1,86735E-07

Simulation	λ	δ	Average TD Value
11	1,25	0,0466	7,89120E-07
12		0,0566	1,38477E-06
13		0,0732	1,86130E-07
14		0,0765	1,73917E-07
15		0,0817	1,83219E-07

Results

No.	Tricluster Diffusion	Dimensions (Gene x Observation x Chromosome)
1.	1,18767E-07	2106 x 12 x 24
2.	1,20914E-07	1979 x 12 x 23
3.	1,23117E-07	1889 x 12 x 19
4.	1,26634E-07	1358 x 12 x 18
5.	1,30937E-07	2076 x 12 x 20
6.	1,34220E-07	1452 x 12 x 18
7.	1,38313E-07	1910 x 12 x 16
8.	1,38879E-07	1416 x 12 x 12
9.	1,39655E-07	1338 x 12 x 12
10.	1,79960E-07	1620 x 12 x 24
11.	1,90850E-07	1442 x 12 x 10
12.	2,11273E-07	1699 x 12 x 14
13.	2,23538E-07	1577 x 12 x 21

Results

No.	Tricluster Diffusion	Dimensions (Gene x Observation x Chromosome)
1.	1,18767E-07	2106 x 12 x 24
2.	1,20914E-07	1979 x 12 x 23
3.	1,23117E-07	1889 x 12 x 19
4.	1,26634E-07	1358 x 12 x 18
5.	1,30937E-07	2076 x 12 x 20
6.	1,34220E-07	1452 x 12 x 18
7.	1,38313E-07	1910 x 12 x 16
8.	1,38879E-07	1416 x 12 x 12
9.	1,39655E-07	1338 x 12 x 12
10.	1,79960E-07	1620 x 12 x 24
11.	1,90850E-07	1442 x 12 x 10
12.	2,11273E-07	1699 x 12 x 14
13.	2,23538E-07	1577 x 12 x 21

Discussion

- The implementation of the gene cube approach in this study is using the gene expression data of bladder cancer patients. From the results of the implementation, it is known that K-means algorithm produces groups of gene expressions that have similar pattern on each axis of each dimension. K-means clustering was successful in finding optimal clusters with small Davies Bouldin index values on the gene, observation, and chromosome axis.

- Further on, based on 15 simulations conducted by δ-Trimax triclustering using different δ and λ, the best simulation is obtained, where the tricluster generated from this simulation has the smallest average tricluster diffusion value. In this simulation, optimal tricluster is produced, some of which are thought to be a group of gene expression that has the characteristics of bladder cancer. Therefore, the gene group in this tricluster can be used by medical experts as a target to stimulate the development of therapy on these genes.

- The gene cube structure manifests a greater opportunity for gene expression analysis as it facilitates the possibility of examining different combinations of experimental variables.

Conclusion

Conclusion

In this research, a data structure known as gene cube has been used. Gene cube is a three-dimensional matrix, where the dimensions consist of genes, observations, and chromosomes. The advantage of this gene cube approach is that the data structure is formed by considering the chromosomes of each gene. This approach can be useful in understanding the mechanisms of disease and tumors in general. By testing the gene cube using K-means algorithm which initial steps is optimized using K-Means++ algorithm and δ-Trimax triclustering algorithm, it has been proven that gene cube structure could provide information about groups of gene expression that have similar pattern.

Reference

Reference

- Arthur, D., & Vassilvitskii, S. (2007). K-Means++: The Advantages of Careful Seeding. Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms, Vol. 8, pp. 1027-1035.
- Baldi, P., & Hatfield, G. W. (2011). *DNA Microarrays and Gene Expression – From Experiments to Data Analysis and Modeling*. Cambridge University Press, pp. 1-213.
- Bhar, A., Haubrock, M., Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., & Wingender E. (2013). *Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell.* Algorithms for Molecular Biology, Vol. 8, No. 9.
- Bhar, A., Haubrock, M., Mukhopadhyay, A., & Wingender, E. (2015). *Multiobjective trilcustering of time-series transcriptome data reveals key genes of biological processes.* BMC Bioinformatics, Vol. 16, No. 200.
- Cohen, B. A., Mitra, R. D., Hughes, J. D., & Church, G. M. (2000). *A computatuional analysis of whole-genome expression data reveals chromosomal domains of gene expression*. Nature Genetics, Vol. 26, pp. 183-186.
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concept and techniques*. ProQuest Ebook Central.
- Henriques, R., & Madeira, S. C. (2018). *Triclustering Algorithms for Three-Dimensional Data Analysis: A Comprehensive Survey*. ACM Computing Surveys, Vol. 51, No. 5, Article 95.

Reference

- Lambrou, G., Sdraka, M., & Koutsouris, D. (2019). *The "Gene Cube": A Novel Approach to Three-dimensional Clustering of Gene Expression Data*. Current Bioinformatics, Vol. 14, pp. 721-727.
- Madeira, S. C., & Oliveira, A. L. (2004). *Biclustering algorithms for biological data analysis: a survey*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 2, No. 4, pp. 719-725.
- Miele, A., & Dekker, J. (2008). *Long-range chromosomal interactions and gene regulation*. Molecular buiSystems, Vol. 4, No. 11, pp. 1046-1057.
- Saputra, N. (2020). *Implementasi Triclustering dengan Menggunakan Metode δ-Trimax pada Data Ekspresi Gen Microarray*. Skripsi. Universitas Indonesia.
- Turkheimer, F. E., Roncaroli, F., Hennuy, B., Herens, C., Nguyen, M., Martin, D., Evrard, A., Bours, V., Boniver, J., & Deprez, M. (2006). *Chromosomal patterns of gene expression from microarray data: methodology, validation and clinical relevance in gliomas.* BMC Bioinformatics, Vol. 7, No. 526.
- Wang, J., & Su, X. (2011). *An improved K-Means clustering algorithm*. 2011 IEEE 3rd International Conference on Communication Software and Networks, pp. 44-46.

Thank You