Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766307

GENE DATA CLUSTERING USING K-MEANS ALGORITHM

Prateek A. Meshram
Assistant Professor, Dr. D Y Patil Institute of Engineering Managemnt and Resrearch, Akurdi SPPU Pune, Prateekmeshram100@gmail.com

Dr.P.P Halkarnikar
Associate Professor, Dr. D Y Patil Institute of Engineering Managemnt and Resrearch, Akurdi SPPU Pune, pp_halkarnikar@rediffmail.com

Dr. Amol Dhakne
Associate Professor, Dr. D Y Patil Institute of Engineering Managemnt and Resrearch, Akurdi SPPU Pune, dhakne.amol5@gmail.com

Mrs. Pratiksha Shevatekar
Assistant Professor, Dr. D Y Patil Institute of Engineering Managemnt and Resrearch, Akurdi SPPU Pune, p_shevatekar@rediffmail.com

Mr. Shivaji R. Vasekar
Assistant Professor, Dr. D Y Patil Institute of Engineering Managemnt and Resrearch, Akurdi SPPU Pune, vasekarsr@gmail.com

Mr. Jitendra Garud
Assistant Professor, Dr. D Y Patil Institute of Engineering Managemnt and Resrearch, Akurdi SPPU Pune, jsgarud@gmail.com

Abstract: Gene expression data contents vital information about the biological process that takes place in a particular organism under specific environment. Gene expression data is vague, imprecise, and noisy. Therefor to get the information of gene states clustering is vital step. Gene expression clustering is used to find co-systematizes gene groups from large collection of gene, whose collective samples are equal to the expressions. Clustering gene expression data benefits in the identification of homology, this helps in vaccine design. There are many unsupervised clustering algorithms used for this purpose. In this paper we have selected yeast sporulation dataset for clustering. The cluster based on prominent features values convey maximum information of bioprocess developed. In order to cluster such a dataset with many features and large unknown patterns k-means algorithm is effective. Such fast pattern finding method is useful for detecting new viruses and drug simulation. The quality of cluster is important while analysis of gene expression. The effectiveness of our proposed algorithm we compared using adjusted rand index (ARI) with the previously reported algorithms and found satisfactory result on yeast sporulation dataset.
Keywords: gene expression, k-means clustering, bioinformatics, elbow method.

Introduction: Faster and efficient gene expression clustering helps in biological analysis of any organism. Advancement of Biotechnology and bio informatics helped in extraction of knowledge form collection of a very large number of gene expressions of different times and conditions. The use of different machine learning algorithm for clustering helps in study of gene expression under different environment. By clustering method we can extract patterns from large amount of data. Gene database has characteristic of large volume with non-uniform patterns. Gene expression data is usually represented by a matrix, with rows corresponding to genes, and columns corresponding to environment conditions or time points.[1] The content of the matrix is the expression levels of each gene under each condition. Each column contains the results obtained from a single array in a particular condition, and is called the profile of that condition. Each row vector is the expression pattern of a particular gene across all the conditions. Clustering of such huge data need faster and efficient clustering techniques. With advancement of technology with faster hardware and software support, it is possible to find different patterns with similarity of expression.[2] This will help in diagnostics, treatment and drug development. Clustering is an unsupervised learning process which divides the data into number of sample spaces N regions {C1,C2,…,CN}. Each cluster C is based on patterns with similarity or dissimilarity. This similarity or dissimilarity is measured using distance formula like Euclidian distance or Manhattan distance. Or indices like Jaccard similarity index.

Cells pass through distinct transcriptional states while governing the different biological processes. Robust, rapid and reproducible transcriptional states are dispensed throughout the genome. Gene regulatory networks (GRN) are developed to describe the functionally and temporally linked regulatory genes. [3] These GRNs are modular and hierarchical which are further divided into sub networks. Biological outputs of up to thousands of genes are coordinated by creating positive or negative loops of networks. Stable and timely biological responses are produced by these networks in response with cellular and environmental stimuli. [4] These high dense genes are to be identified by means of cluster techniques. Gene clustering has significant role in identifying new biological processes and behaviour on known gene patterns. Gene expression data have vital information that is required to understand the biological process which takes place in a particular organism under certain environment. Functional genomics is analysed by clustering the hidden pattern in gene expressions. [5][6] Due to its imprecision, vagueness and noise clustering techniques provide tool to expose hidden patterns and different structures in huge gene data. This clustering of gene expression data is useful in knowing the natural structure inherent in gene expression data, cellular processes, understanding gene functions. Gene expression is useful for identifying the molecular signature of a disease and for correlating a pharmacodynamics marker with the dose-dependent cellular responses to exposure of a drug.[7] Several transcriptomics technologies can be used to generate the necessary data to analyse. DNA microarrays measure the relative activity of previously identified target genes. Sequence based techniques, like RNA-Seq; provide information on the sequences of genes in addition to their expression level.

Related work:

Heba Saadeh et.al. [8] used K-means for clustering of Haematopoietic stem cells. These stem cells are involved in development of various types of blood cells, such as white blood cells , red blood cells and platelets, by conducting different genetic programmes. Understanding the

pattern in gene expression and understanding the specific genes expressed during different stage of erythropoiesis are important to understand biological development in synthesis of erythroid. Applying K-means algorithm to cluster different genetic expressions authors has identified correlation between gene expressions during various stages of development. By applying K-means clustering on human erythropoiesis data to find a specific lineage of developing haematopoietic stem cells into red blood cells. Gene expression dataset which measures global gene activity across 4 different developmental stages; CFU-E, Pro-E, Int-E and Late-E is used for analysis. As in K-means K is predefined, effective K is decided by Elbow method. The acceptable number of clusters is obtained by using different K values. Authors found 8 clusters are effective using Elbow method.

P. Rajalakshmi et.al. [9] selected K-means algorithm for clustering of gene expression on bases of speed, implementation easiness and quality of clusters. However to overcome the drawback of lazy convergence they have developed rough K- means variation of K-means algorithm. This modification overcome drawback of slow in convergence and sensitivity to the initial value of selection. Rough K-Means algorithm uses lower and upper boundary approximations. They studied both the algorithms on four datasets, taken from National Centre for Biotechnology and Information (NCBI), which are Gene Expression Omnibus datasets. Considering the mathematical property of Rough Set Theory, lower and upper approximations are generated for Rough K Means algorithm. The rough computing model has proposed for improving the gene selection method in a simple and efficient way.

S. Ranathive et.al. [10] used K-means algorithm to cluster Yeast dataset. K-means algorithm starts with random seed values, which result in different cluster time during different run of algorithm. Author suggested modified K-means algorithm to overcome the short coming of this algorithm. The initialization of random centre that is very distant from the sample data may produce instability in algorithm results. To overcome this, select the random initial centres from training data itself or assigning few values randomly which are inside the range of data. Also, the equal active state of centres is not guaranteed. Several samples will be present in few centres, while at the same time few centres will have few samples and that are not updated in further iterations. Proper selection of initial centres results in few iteration to complete the K-means algorithm. To achieve this traditional k-means is modified using graph theory. Minimum spanning tree technique is applied depending on Kruskal's algorithm is to select initial seeds to the conventional scheme of K-means. Applying modified K-means algorithm has improved the quality of clusters and centres of those clusters.

Hui Wen Nies et.al. [11] has reviewed different cluster methods used in clustering gene expression. The study gives comparisons between popular methods for clustering. The inability to decide an optimal number of potential clusters beforehand, many clustering methods misclassifies the gene expression in limited classes. This paper reviews existing methods used for clustering genes, for identifying biologically informative genes. The optimization of the clustering functions used and clustering validation post clustering are also studied. Authors investigate the performance of different clustering techniques using a leukaemia dataset. The results found that grid-based clustering techniques provide better classification accuracy. However k-means and hybrid clustering techniques such as CLIQUE yield high-quality gene clusters.

Hicks SC et.al. [12] address the issue related to K-means clustering algorithm while using the large gene expression dataset. To measure gene expression at the single-cell level in genome, Single-cell RNA-Sequencing (scRNA-seq) is widely. Unsupervised clustering algorithms are used for analyses of scRNA-seq data to detect distinct subpopulations of cells. However, due to recent advances in scRNA-seq dataset resulted range from thousands to millions of expressions. If K-means algorithm is used for clustering typically need the entire dataset to be loaded into memory. This results in slow clustering or impossible execution to run on large datasets. To overcome this short coming of K-means, authors developed the mini-batch k-means algorithm, mbkmeans. In this algorithm it is not necessary to load entire dataset. Ensted the dataset is loaded in batches from disk storage. The K-means iteration take place in similar fashion but using predefined small batched of data. This process continues until convergence.

K-means algorithm

As knowledge regarding the relation between the genes and cells development has no prior assumption, unsupervised learning techniques are preferred in clustering of a set of genes involved in the different stages. K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled dataset into different clusters. The K-means method is one of the simplest and most common clustering algorithms. K-means partition the dataset $D=\{ x_i\}^{N}_{i=1}$, into K clusters denoted by $\{C_1,C_2,\ldots,C_N\}$. each cluster is represented by mean point of cluster i.e. μ_i, which is mean of all samples in cluster. Such algorithms rely on finding the similarity between clusters observations. It follows a greedy approach that aims to minimise the sum of squared error (SSE) over different observations during convergence of algorithm. SSE is defined as

$$SSE = \sum_{i=1}^{n}(C_{new} - C_{old})^2 \qquad \text{---- (1)}$$

Cluster validation

Even though K-mean is unsupervised algorithm, the value of number of final clusters (K) is predefined. To decide number of cluster in gene expression it is necessary to conduct many experiments with different K values. In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

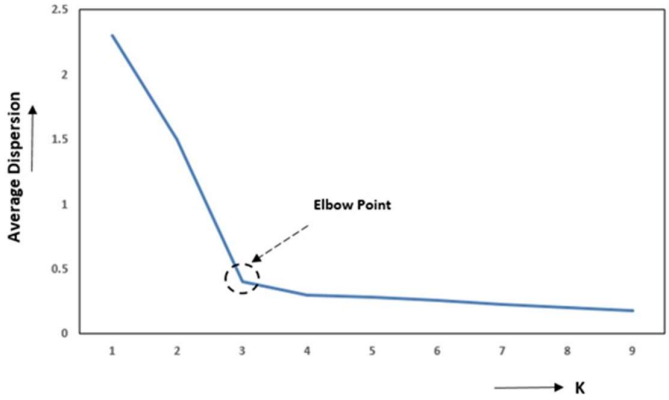


Figure 1: Elbow method to select optimal K value.

Proposed Method

For clustering of gene expression data set we have selected K-means algorithm as it is fast and simple to implement. The way k-means algorithm works is as follows:

Specify number of clusters K.

Initialize centroids randomly selecting K centre points from the gene expression dataset.

Keep iterating until there is no change to the centroids. i.e. assignment of samples to clusters isn't changing. For each iteration do following task

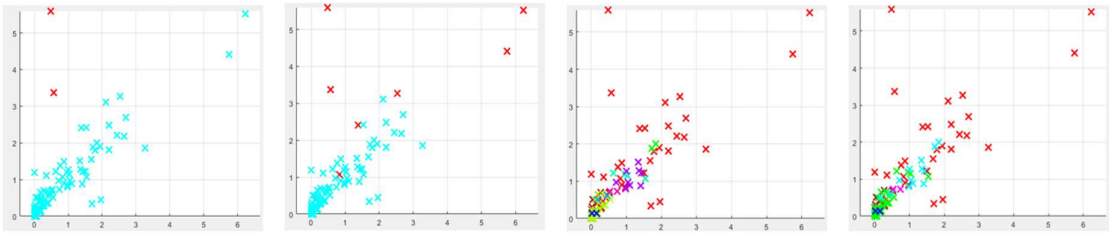
a. Compute the sum of the squared distance between data points and all centroids.

b. Assign each data point to the closest cluster (centroid).

c. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

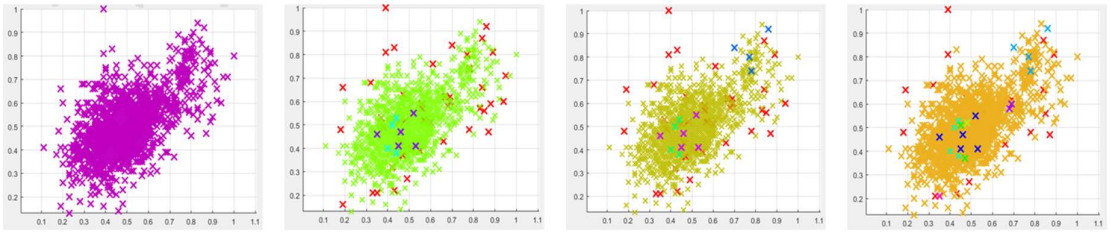
Result analysis

The proposed algorithm is developed and implemented using MATLAB 2019. The dataset used for experiments is Yeast sporulation dataset. The dataset consists of sporulation process genes of the budding yeast consisting of 6118 genes. The dataset is available on the following website: http://cmgm.stanford.edu/pbrown/sporulation. From the total 6118 genes. 13 prominent feature values are taken for clustering. This exhibit the strength of K-means algorithm for scaling purpose. Initial seed values are selected randomly but from the dataset available, in order to avoid the instability of algorithm.We compared our method with the previous algorithms and results reported and found that the proposed algorithm clusters more effectively than the previously reported algorithms in terms of adjusted rand index (ARI). The observed values for our algorithm on yeast gene expression is 0.9800 and on ratCNS is 0.5560 for ARI. The comparative results SiMM-TS, IFCM, VGA, Average Link SOM and CRC are shown in table 3. Figure 2 and Figure 3 shows the graphical outputs of the indexes and the datasets with respect to the number of iterations on validity indexes ARI, Jaccard Index (JI) and Cavg respectively. Figure 5 shows the comparative graph with proposed method on the basis of ARI index obtained from six algorithms.



Iteration 1 Iteration 10 Iteration 20 Iteration 30

Figure 2: Variations in RatCNS dataset at various iterations.



Iteration 1 Iteration 10 Iteration 20 Iteration 30

Figure 3: Variations in Yeast dataset at various iterations.

Table 1: Number of clusters obtained by the proposed algorithm

	Yeast	RatCNS
Original Clusters	10	6
Obtained Clusters	10	6

Table 2: Result of Proposed algorithm

Dataset	Elbow	clstr	Itr	Proposed Algorithm		
				C_{avg}	ARI	JI
RatCNS	12.58	3	1	0.292	0.42	0.384
	2.901	2	10	0.267	0.505	0.498
	0.84	4	20	0.255	0.536	0.537
	0.89	6	30	0.228	**0.556**	**0.532**
Yeast	29.46	5	1	0.201	0.223	0.223
	0.198	6	10	0.124	0.776	0776
	0.19	8	20	0.114	0.917	0.906
	0.20	10	30	0.11	**0.981**	**0.914**

Table 3: Comparison of gene datasets for various algorithms with respect to ARI.

Dataset / Algorithm	Yeast	RatCNS
Our	**0.981**	**0.556**
SiMM-TS	0.6353	0.5147
IFCM	0.4717	0.4032
VGA	0.5800	0.4542
Average Link	0.5007	0.3684
SOM	0.5842	0.4134
CRC	0.5675	0.4455

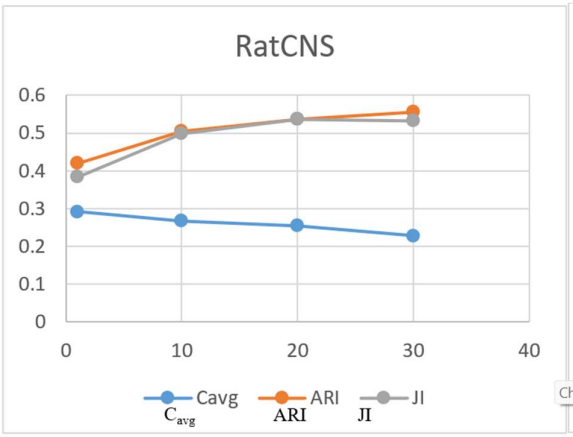


Figure 4: Variations in RatCNS dataset at different intervals

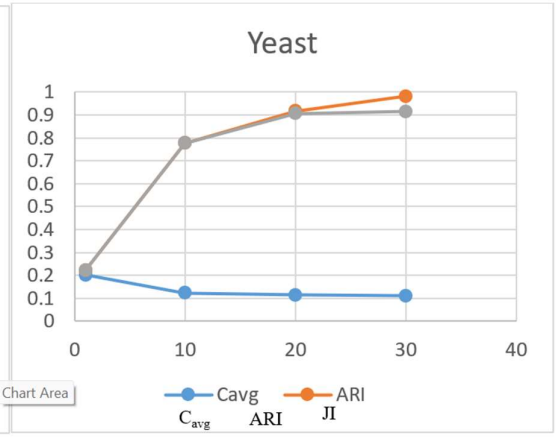


Figure 5: Variations in Yeast dataset at different intervals

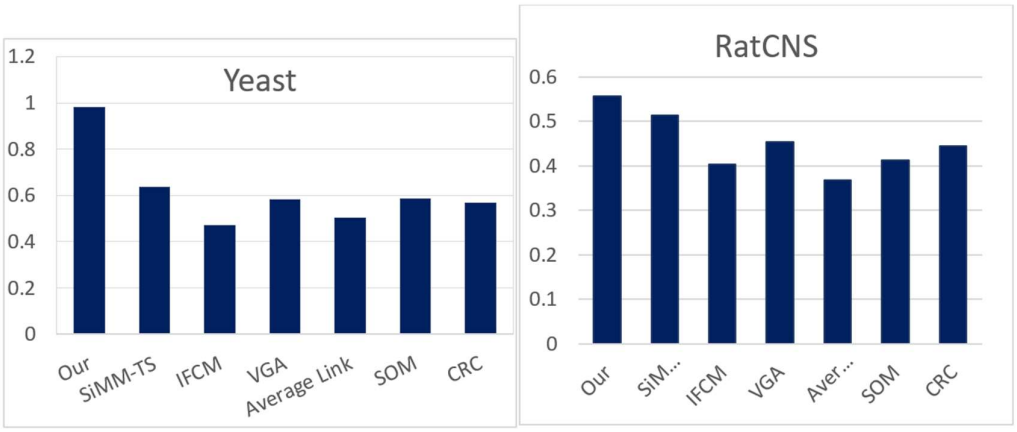


Figure 6: - Comparative graph for the variations in the dataset with respect to ARI in all the algorithms compared.

Conclusion

To understand the biological processes taking place at different simulation it is necessary to find out patterns in gene expressions available in genome. To find the patterns of gene expression at different situation it is necessary to cluster them on similarity index. Such clusters give useful bio information about cell behaviour to different stimuli. In order to find out patters out of large unknown data unsupervised clustering methods are used. In this paper we have used K-means cluster method to analyse yeast sporulation dataset. We have found similar pattern out of large dataset in efficient and fast algorithm. Such analysis of gene expression is necessary to understand unexpected biological processes. Such analysis helps in understanding new variations of viruses and development of drugs. K-means algorithm is effective in gene expression as number of parameters used for finding distinguished features can be incorporated easily in this algorithm. The gene expression having multiple features can be handled effectively by this algorithm. We compared our result with ARI to find effectiveness with multiple parameters for clustering.

References

[1] D. B. Searls, “Using Bioinformatics in Gene and Drug Discovery,” Drug Discov. Today, vol. 5, no. 4, pp. 135–143, 2000.

[2] F. K. Ahmad, Y. Yusof and N. H. Othman, "Gene selection for high dimensional data using k-means clustering algorithm and statistical approach," 2014 International Conference on Computational Science and Technology (ICCST), Kota Kinabalu, Malaysia, pp. 1-6, doi: 10.1109/ICCST.2014.7045188, 2014.

[3] Richard I. Joh , Michael S. Lawrence , Martin J. Aryee and Mo Motamedi, “Gene clustering coordinates transcriptional output of disparate biological processes in eukaryotes”, BioRxiv, doi: https://doi.org/10.1101/2021.04.17.440292; April 19, 2021.

[4] Martin C Nwadiugwu, “Gene-Based Clustering Algorithms: Comparison Between Denclue, Fuzzy-C, and BIRCH”, Bioinformatics and Biology Insights, Volume 14, pg. 1–6 , 2020.

[5] J. Oyelade, Itunuoluwa Isewon, “Clustering Algorithms: Their Application to Gene Expression Data.”, Bioinformatics and Biology Insights, pp 237–253, doi: 10.4137/BBI.S38316, 2016.
[6] patrik D’haeseleer, “ how does gene expression clustering work?”, Nature Biotechnology, Vol. 23, No. 12, pp 1499-1501, December 2005.
[7] Prateek A. Meshram and Pradeep Singh, “An efficient density-based algorithm for clustering gene expressions”, Bioscience Biotechnology Research Communication, Vol. 10, No.2, pp. 1-5, 2017.
[8] Heba Saadeh, Reem Q. Al Fayez, and Basima Elshqeirat, “Application of K-Means Clustering to Identify Similar Gene Expression Patterns during Erythroid Development”, International Journal of Machine Learning and Computing, Vol. 10, No. 3, pp. 452-457, May 2020.
[9] P. Rajalakshmi, K. Thangavel, E. N. Sathishkumar, “Gene Expression Data Analysis using Rough K-Means Clustering Method”, International Journal of Computational Intelligence and Informatics, Vol. 7: No. 1, June 2017.
[10] S. Ranathive, Nelson Kennedy Babu, Miretab Tesfayohanis, S.Sivakumar, “Gene Expression Data Clustering Using Improved K-Means Algorithm”, Turkish Online Journal of Qualitative Inquiry (TOJQI) Volume 12, Issue 7, pp. 12673 – 12686, July 2021.
[11] Hui Wen Nies, Zalmiyah Zakaria, Mohd Saberi Mohamad, “A Review of Computational Methods for Clustering Genes with Similar Biological Functions”, Processes, Vol. 7, pp 550-568; doi:10.3390/pr7090550, 2019.
[12] Hicks SC, Liu R, Ni Y, Purdom E, Risso D, “bkmeans: Fast clustering for single cell data using mini-batch k-means”. PLoS Computational Biology, Vol. 17, No. 1, https://doi.org/10.1371/journal. pcbi.1008625, (2021).
[13] Juan A. Botía, Jana Vandrovcova, Paola Forabosco, Sebastian Guelfi, “An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks”, BMC Systems Biology, Vol. 11, pp. 47- 64, DOI 10.1186/s12918-017-0420-6, 2017.