The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 380 – 391

# Machine Learning Methodologies for Clustering Gene Expression Data in Cancer

**Alka A.Kumbhar**
JSPM's Rajarshi Shahu College of Engineering
**P. B. Kumbharkar**
JSPM's Rajarshi Shahu College of Engineering
**D. T. Mane**
JSPM's Rajarshi Shahu College of Engineering
alka.kumbhar@gmail.com

**Abstract**
Gene expression data hide vital information required to understand the biological process that takes place in a particular organism. Extracting the hidden patterns in gene expression data helps to strengthen the understanding of functional genomics. The complexity of biological networks and the volume of genes present increase the challenges of comprehending and interpretation of the resulting mass of data, which consists of millions of measurements; these data also inhibit vagueness, imprecision, and noise. Therefore, thousands of gens can be analyzed at a time using clustering techniques is a first step toward addressing these challenges, which is essential in the data mining process to understand natural structures and identify interesting patterns in the underlying gene expression data [2]. The clustering of gene expression data has been proven to be useful in making known the natural structure inherent in gene expression data, understanding gene functions, cellular processes, and subtypes of cells, finding useful information from noisy data, and understanding gene regulation. The other benefit of clustering gene expression data is the identification of homology, which is very important in drug design. Clustering is a useful method that groups items based on certain similarity measures for understanding the structures, functions, regulation of genes, and cellular processes obtained from gene expression data and providing more insight on a given data set [13].
This review examines the various clustering algorithms applicable to the gene expression data in order to discover and provide useful knowledge of the appropriate clustering technique that will guarantee stability and high degree of accuracy in its analysis procedure.
**Keywords:** Unsupervised Machine Learning , Statistical significance, gene expression, ALL

## 1. Introduction

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, or nearly one in six deaths. The most common cancers are breast, lung, colon and rectum and prostate cancers. Cancer is a generic term for a large group of diseases that can affect any part of the body. One defining feature of cancer is the rapid creation of abnormal cells that grow beyond their usual boundaries, and which can then invade adjoining parts of

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 380 – 391

the body and spread to other organs; the latter process is referred to as metastasis. Widespread metastases are the primary cause of death from cancer[3].

Luekemia a type of bone marrow cancer , normally bone marrow produces too many white blood cells that helps our body to fight the infection but in ALL bone marrow produces abnormal cells and they crowd out the healthy cells which can lead to anemia and bleeding. This may spread the infection in other part of the body like brain and spinal cord .Cancer can be described as a disease of altered gene expression. There are many proteins that are turned on or off (gene activation or gene silencing) that dramatically alter the overall activity of the cell. A gene that is not normally expressed in that cell can be switched on and expressed at high levels.

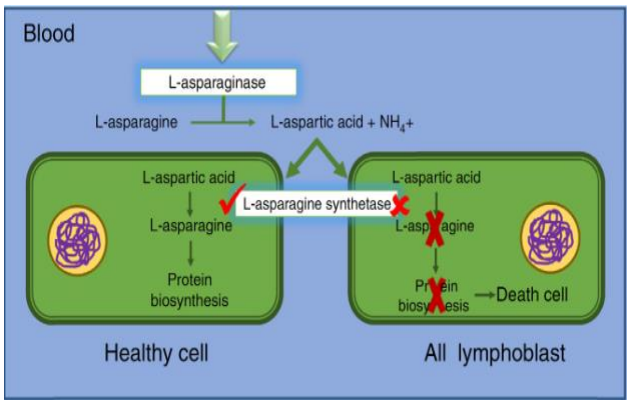



**Fig 1. ( Antineoplastic action of L-asparaginase )**

L-asparagine is transformed into Aspartate during L-asparagine therapy, and L'Aspartate then enters cells via an amino acid transporter. L-aspartate will be transformed back to L-asparagine in healthy cells by the enzyme L'Asparaginase Synthetase (ASNS). On the other hand, cancer cells are unable to synthesize asparagine because they express ASNS either not at all. Asparagine depletion by L-asparaginase causes these cancer cells to undergo apoptosis.

## 2. Comprehensive Analysis Of Previous Works

| Year | Paper Title | Description | Shortcoming |
|---|---|---|---|
| Bioinformatics. 2021 Sep | Bipartite graph-based approach for clustering of cell lines by gene expression–drug response associations. Bioinformatics, 37(17), pp.2617-2626 | In pharmacogenomic studies, the biological context of cell lines influences the predictive ability of drug-response models and the discovery of biomarkers. Thus, similar cell lines are often studied together based on prior | Author presents a procedure to compare cell lines based on their gene–drug association patterns. Starting with a grouping of cell lines from biological annotation, the model gene–drug association patterns for each group as a bipartite graph |

| | | | |
|---|---|---|---|
| | | knowledge of biological annotations. However, this selection approach is not scalable with the number of annotations, and the relationship between gene–drug association patterns and biological context may not be obvious. | between genes and drugs. This is accomplished by applying sparse canonical correlation analysis (SCCA) to extract the gene–drug associations, and using the canonical vectors to construct the edge weights. Then, paper introduce a nuclear norm-based dissimilarity measure to compare the bipartite graphs. Accompanying our procedure is a permutation test to evaluate the significance of similarity of cell line groups in terms of gene–drug associations. |
| Wiley Interdiscip Rev Syst Biol Med . 2020 Nov | Molecular networks in Network Medicine: Development and applications. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 12(6), p.e1489. | Network Medicine applies network science approaches to investigate disease pathogenesis. Many different analytical methods have been used to infer relevant molecular networks, including protein-protein interaction networks, correlation-based networks, gene regulatory networks, and Bayesian networks. Network Medicine applies these integrated approaches to Omics Big Data | Author discuss briefly the types of molecular data that are used in molecular network analyses, survey the analytical methods for inferring molecular networks, and review efforts to validate and visualize molecular networks. Successful applications of molecular network analysis have been reported in pulmonary arterial hypertension, coronary heart disease, diabetes mellitus, chronic lung diseases, |

| | | | |
|---|---|---|---|
| | | (including genetics, epigenetics, transcriptomics, metabolomics, and proteomics) using computational biology tools and, thereby, has the potential to provide improvements in the diagnosis, prognosis, and treatment of complex diseases. | and drug development. Important knowledge gaps in Network Medicine include incompleteness of the molecular interactome, challenges in identifying key genes within genetic association regions, and limited applications to human diseases. |
| Journal of computational biology Volume 28, Number 5, 2021 | Deep large-scale multitask learning network for gene expression inference. Journal of Computational Biology, 28(5), pp.485-500. | Gene expression profiling makes it possible to conduct many biological studies in a variety of fields due to its thorough characterization of cellular states under various experimental conditions. Despite recent advances in high-throughput technology, profiling an entire set of genomes is still difficult and expensive. Due to the high correlation between expression patterns of different genes, the aforementioned problem can be solved with a cost-effective approach that collects only a small subset of genes, called landmark | In this study, we introduced a new MTL method for training deep inference models for estimating the gene expressions. Our algorithm improves the generalizations of multitask predictors by effectively discovering the task correlations. To do so, we proposed a seamless regularization for deep neural networks that is scalable to a huge number of tasks. Experimental results confirmed the effectiveness of our proposed algorithm compared to alternative models, where our model consistently and significantly outperforms all counterparts on two gene expression |

| | | | |
|---|---|---|---|
| | | genes, representing the entire set of genes, and infer the remaining genes, called target genes, using a computational model. There are several shallow and deep regression models in literature to estimate the expressions of target genes from the landmark genes. proposed method outperforms the shallow and deep regression models for gene expression inference and alternative multitask learning algorithms on two large-scale datasets regardless of the network architecture. | datasets with various base network architectures. We also visualized the role of landmark genes in estimating the expressions of target genes, providing better insights about the knowledge learned by our regression model |
| J Child Orthop. 2007 Mar; 1(1): 63–68 | Acute lymphoblastic leukemia. Current problems in pediatric and adolescent health care, 32(2), 40-49, | Studies on musculoskeletal manifestations (MSM) of childhood acute lymphoblastic leukemia (ALL) have yielded variable findings with regard to their clinical impact. We investigated the significance for differential diagnosis, treatment and outcome of musculoskeletal complaints as presenting symptoms of ALL, and their | MSM occur mostly in children with BCP ALL who present with less involvement of extramedullary organs, low peripheral blood blasts and white blood cells counts. These findings highlight the importance of including ALL in the differential diagnosis of MSM even in the presence of an apparently normal peripheral blood count. Our study also suggests that MSM are caused |

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 380 – 391

| | | | |
|---|---|---|---|
| | | correlation with leukemia immunophenotypes, for which data is lacking. | by leukemic cells with enhanced biological propensity to remain relatively confined within the intramedullary bone-marrow space. |

## 3. Objectives

Considering the existent research gap in ALL studies, this research work is undertaken to understand fourfold primary.

- Identifying the differentially expressed genes from the microarray dataset
- Constructing and analysing a complex biomolecular network out of these
- Computing different network parameters using a network visualization software
- Based on that information, evaluating potential drug targets among the set of genes.

## 4. Materials and Methods

Data clustering plays an important role in effective analysis of gene expression. Although DNA microarray technology facilitates expression monitoring, several challenges arise when dealing with gene expression datasets. Some of these challenges are the enormous number of genes, the dimensionality of the data, and the change of data over time. The genetic groups which are biologically interlinked can be identified through clustering. This project aims to clarify the steps to apply clustering analysis of genes involved in a published dataset. The methodology for this project includes the selection of the dataset representation, the selection of gene datasets, Similarity Matrix Selection, the selection of clustering algorithm, and analysis tool.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 380 – 391

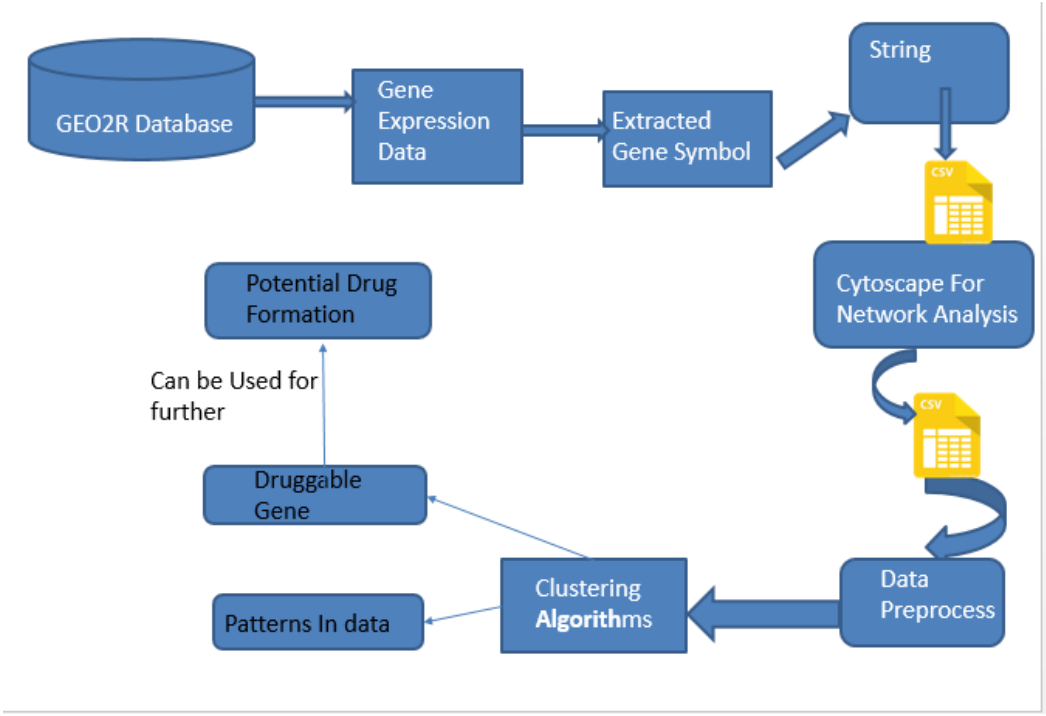



**Fig 2 (Workflow for finding DEgenes using Unsupervised ML)**

Microarray data from ALL-related microRNA and mRNA expression profiles were retrieved from the National Center for Biotechnology Information (NCBI) GEO[13] database of species—Homo sapiens. We downloaded the microRNA expression microarray dataset GSE4072 using platform on Homo sapiens organism.

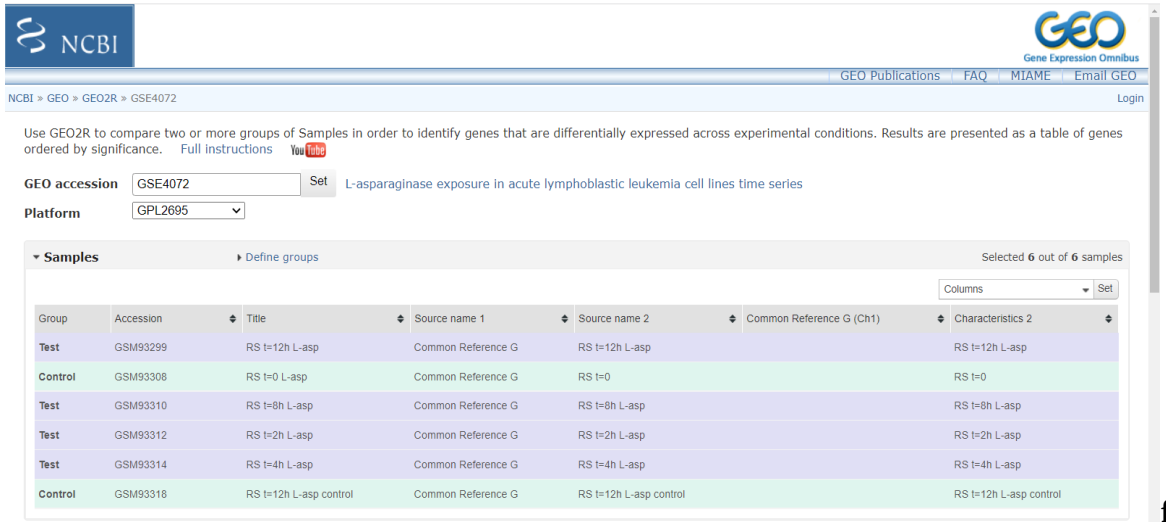



**fig 2 (https://www.ncbi.nlm.nih.gov) (Defined Groups Control and TEST)**

Clustering is an unsupervised learning technique which classify objects in groups with respect to their similar characteristics. Cluster analysis is traditionally used in phylogenetic research and has been adopted to microarray analysis as well. Traditionally there are various

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 380 – 391

clustering algorithm like k-means, hierarchical, SOM etc. Cluster analysis may be used as data reduction method in which observations can be represented by mean of the observations in particular cluster.

In silico investigation of L-asparaginase exposure on ALL cell lines might be helpful. cDNA microarray dataset from NCBI (National Center for Biotechnology Information) with GEO accession number GSE4072, "L-asparaginase Exposure in Acute Lymphoblastic Leukemia Cell Lines Time Series." Since ALL cell line RS4; 11 has an LC50 (the quantity of L-asparaginase fatal to 50% of the cells) less than 0.003 IU/mL, [14] it is chosen as our experimental cell line and is proven to be sensitive to L asparaginase [1].

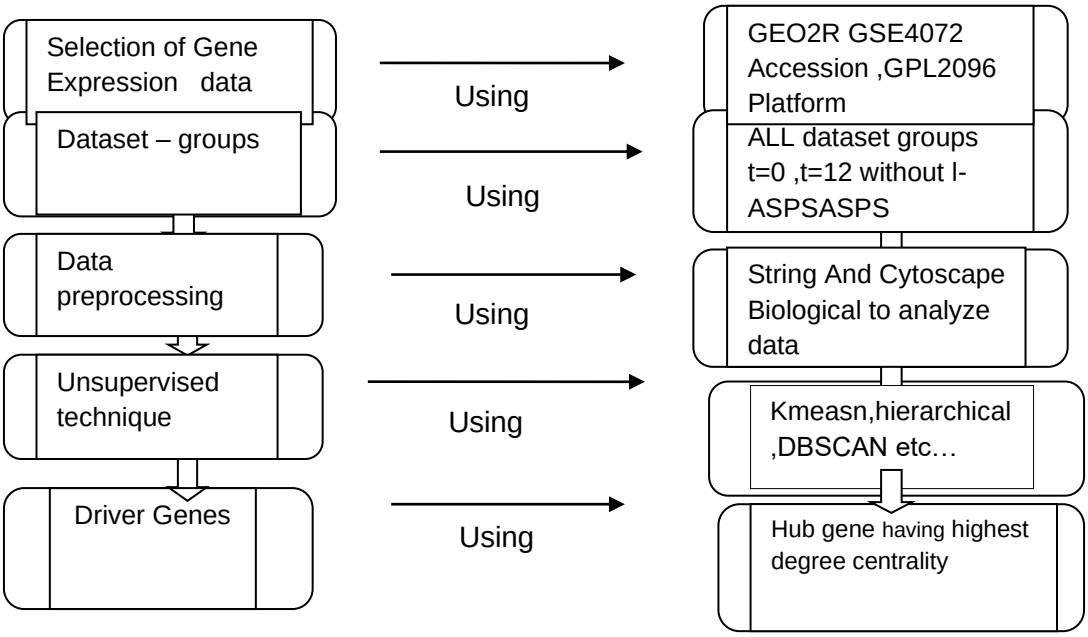



**Fig 3 (Selected required subset of gene expression)**

As shown in fig 3 gene which is top genes in GEO2R data are selected, Using the set of differentially expressed genes, we construct and analyze a complex bimolecular network. Using a network visualization software called STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database (http://string-db.org) and as organism Homo sapiens was selected. Following, a biomolecular network was constructed out of it with a minimum required interaction score, set to high confidence (0.700) for eliminating weaker interactions. More nodes (gene factors) were added for more intermolecular interactions and better molecular visualization. In the 'Evidence' section, only co-expression was selected to get genes that were co-expressed in the same or other species (transferred by homology).to find the protein interaction of expressed gens [16].

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 380 – 391

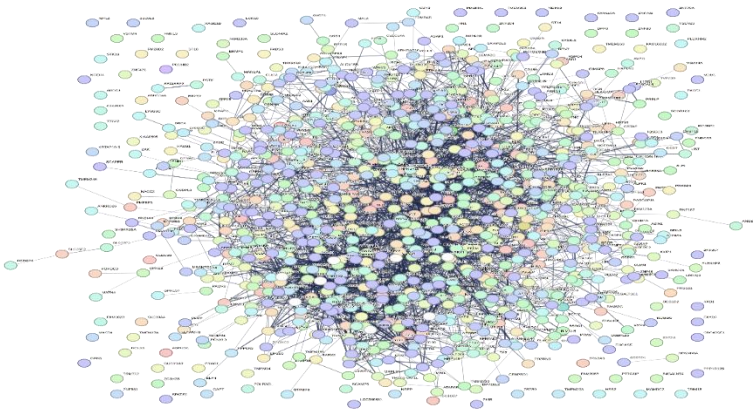


**Fig 4 (String Network Of Selected Data)**
**(https://version-11-5.string)**

We analyze different network parameters and based on that information again the is fed to another biological tool named CYTOSCPE as shown in fig 5 to find the how the proteins are strongly related according to their degree centrality and other network parameter a data of network and the different Machine learning clustering techniques are applied to group the genes with similarexpression and analyzed different Unsupervised techniques .

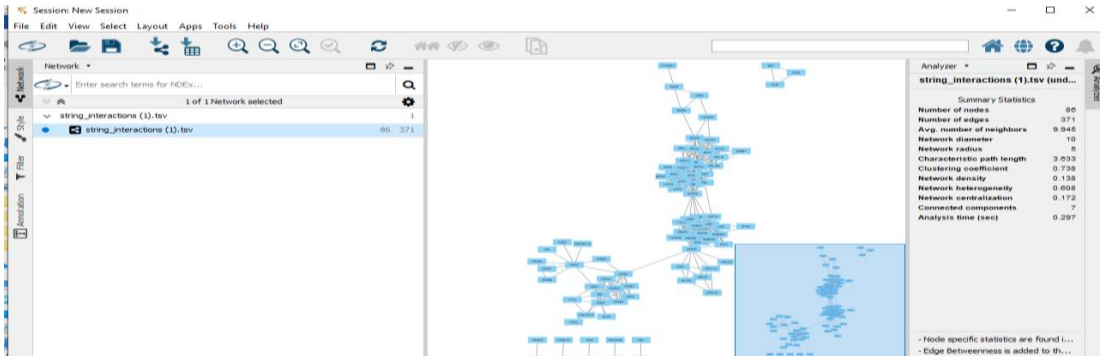



**Fig 5 (Analyzed Cytoscape Network)**

## 5. Result

The clustering of gene expression data has been proven to be useful in making known the natural structure inherent in gene expression data, understanding gene functions, cellular processes, and subtypes of cells, mining useful information from noisy data, and understanding gene regulation.Differntially expressed genes which we got in the retrieved dataset met the criteria of highest degree and Betweenjness centrality and n the review we got two differentially expressed genes named LSM3 and HSPA8 and further can be used for the drug design and it depends which genes the user gets sonce this is time series data .In 2020 GEO2R doesn't more than 250 rows of data .restricted to top 250 only which we need to analyze .But now a days it is allowing thousands of data due to this user can use large amount of data and analyze it for the biologist this in silico analysis of gene expression will reduce the cost and time the biologist which they need for in vitro analysis of cell line .

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 380 – 391

## 6. Conclusion and Future Scope

Analyzing data using biological tools like STRING and CYTOSCAPE to find the network data is time consuming when the data is too large since machine learning lays vital role in analyzing the data in very less amount of time .We created a simple Unsupervised Machine Learning module using very popular programming language Python which gives a analysis and accuracy of different clustering methods on the Acute Lymphoblastic Leukemia Dataset . The same can be used by biologist to find whether a particular sample of data has driver genes for ALL or not that too also in less amount of time. We used these tools just for the data production one can directly apply clustering methods on the cancer data . The study is restrict in-silico analysis to only one ALL cell line i.e. RS4;11 which is L-asparaginase sensitive. Analyzing differentially expressed genes in many ALL cell lines( irrespective of it being sensitive / intermediate or resistant to L-asparaginase) and clinical samples will give a comprehensive genome-wide view of ALL cells to L-asparagine ,which can be fruitful to evaluate druggable gene targets in ALL cells.

## 7. References

[1] Zhang, S., Wang, Y., Gu, Y., Zhu, J., Ci, C., Guo, Z., Chen, C., Wei, Y., Lv, W., Liu, H. and Zhang, D., 2018. Specific breast cancer prognosis-subtype distinctions based on DNA methylation patterns. Molecular oncology, 12(7), pp.1047-1060.

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7133071/

[3] Chi, C., Ye, Y., Chen, B. and Huang, H., 2021. Bipartite graph-based approach for clustering of cell lines by gene expression–drug response associations. Bioinformatics, 37(17), pp.2617-2626.

[4] Silverman, E.K., Schmidt, H.H., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., Balligand, J.L., Benincasa, G., Capasso, G., Conte, F. and Di Costanzo, A., 2020. Molecular networks in Network Medicine: Development and applications. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 12(6), p.e1489.

[5] Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A. and Hoffman, M.M., 2019. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion, 50, pp.71-91.

[5] Sealfon, R.S., Wong, A.K. and Troyanskaya, O.G., 2021. Machine learning methods to model multicellular complexity and tissue specificity. Nature Reviews Materials, 6(8), pp.717-729.

[6] Xu, W., Liu, X., Leng, F. and Li, W., 2020. Blood-based multi-tissue gene expression inference with Bayesian ridge regression. Bioinformatics, 36(12), pp.3788-3794.

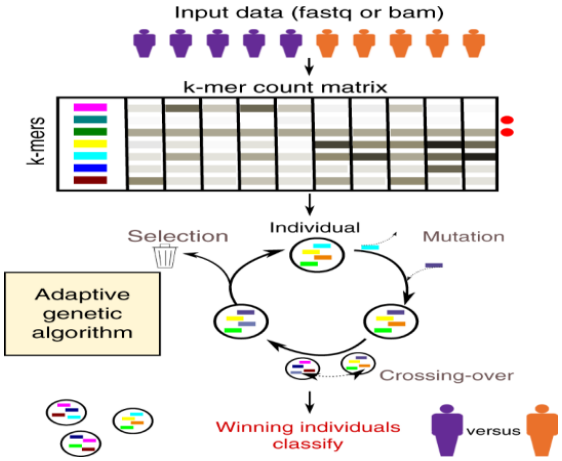
[7] Way, G.P., Natoli, T., Adeboye, A., Litichevskiy, L., Yang, A., Lu, X., Caicedo, J.C., Cimini, B.A., Karhohs, K., Logan, D.J. and Rohban, M.H., 2022. Morphology and gene expression profiling provide complementary information for mapping cell state. bioRxiv, pp.2021-10.

[8] Dizaji, K.G., Chen, W. and Huang, H., 2021. Deep large-scale multitask learning network for gene expression inference. Journal of Computational Biology, 28(5), pp.485-500.

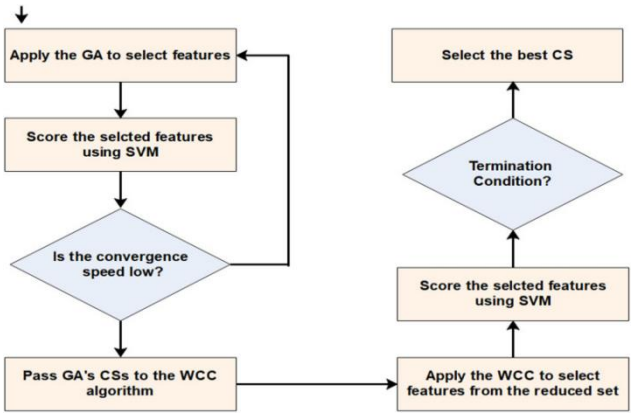
[9] Dincer, A.B., Celik, S., Hiranuma, N. and Lee, S.I., 2018. DeepProfile: Deep

learning of cancer molecular profiles for precision medicine. BioRxiv, p.278739.
[10] Battineni, G., Sagaro, G.G., Chinatalapudi, N. and Amenta, F., 2020. Applications of machine learning predictive models in the chronic disease diagnosis. Journal of personalized medicine, 10(2), p.21.
[11] Raza, K., 2019. Fuzzy logic based approaches for gene regulatory network inference. Artificial intelligence in medicine, 97, pp.189-203.
[12] Almugren, N. and Alshamlan, H., 2019. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE access, 7, pp.78533-78548.
[13]https://www.ncbi.nlm.nih.gov/geo/geo2r/?
[14]https://www.cancer.gov/types/leukemia
[15]https://pubmed.ncbi.nlm.nih.gov/1566536
[16] https://string-db.org/
[17] Ching-Hon Pui, et al. (2004).Acute lymphoblastic leukemia, Journal of medicine, 350(15), 1535-1548, Available at: https://www.nejm.org/doi/full/10.1056/NEJ Mra023001.
[18] Ching-Hon Pui, et al. (2004).Acute lymphoblastic leukemia, Journal of medicine, 350(15), 1535-1548, Available at: https://www.nejm.org/doi/full/10.1056/NEJ Mra023001.
[19] Dieter Hoelzer,et al.(2002).Acute lymphoblastic leukemia, ASH Education Program Book ,Available at : https://ashpublications.org/hematology/artic le/2002/1/162/18610/Acute-LymphoblasticLeukemia.
[20] Scott A. Armstrong & A. Thomas Look. (2005). Molecular genetics of acute lymphoblastic leukemia. Journal of Clinical Oncology, 23(26), 6306-6315, Available at: https://ascopubs.org/doi/10.1200/JCO.2005. 05.047.
[21] Pui, C. H. (2000). Acute lymphoblastic leukemia in children. Current opinion in Oncology, 12(1), 3-12, Available at: https://journals.lww.com/cooncology/Abstract/2000/01000/Acute_lymphoblastic_leukemia_in_children.2.aspx.
[22] Elias J. Jabbour. (2005). Adult acute lymphoblastic leukemia. In Mayo Clinic Proceedings, 80(11), 1517-1527, Available at: https://mdanderson.elsevierpure.com/en/publications/adult-acute-lymphoblasticleukemia.
[23] Chan, K. W. (2002). Acute lymphoblastic leukemia. Current problems in pediatric and adolescent health care, 32(2), 40-49,
[24] Pirim H, Ekşioğlu B, Perkins AD, Yüceer Ç. Clustering of high throughput gene expression data. *Comput Oper Res.* 2012;39(12):3046–61. [PMC free article] [PubMed] [Google Scholar]
[25]Centrality-Measures Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks-2008
[26] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2733090/
[27] Clustering cancer gene expression data by projective clustering ensemble-2017
[28] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5325197/

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 380 – 391




(Source: https://media.springernature.com/)




(Source: https://media.springernature.com/)

**Appendix 2: Machine learning algorithms method**

**Appendix 3: Machine learning applications**