Gene Expression Analysis: Cancer Classification using Clustering

BT19CSE052 Kaushal Lodd
BT19CSE092 Saarang Rajguru

Introduction:

Gene expression analysis is a set of techniques used to measure the activity of genes in a cell or organism. The activity of a gene is usually determined by the level of the mRNA transcripts it produces, which is a reflection of the amount of protein it codes for. By measuring the mRNA levels, researchers can determine which genes are being actively expressed and at what levels, providing valuable insights into cellular processes and biological mechanisms. It is used in a variety of research fields, including genetics, developmental biology, and disease research. By understanding the patterns of gene expression, researchers can gain insights into how cells and organisms respond to various stimuli, and how diseases alter normal gene expression patterns.

Gene expression analysis using clustering can help to identify different classes or subtypes of cancer based on their characteristic patterns of gene expression. By grouping together genes that have similar expression profiles, clustering algorithms can identify subgroups of cancer samples that are biologically distinct from each other. These subgroups can provide a way to classify cancers into different subtypes, based on their gene expression patterns.

Gene expression analysis using clustering can provide a more comprehensive understanding of the biology of cancer, beyond just individual gene expression levels. By identifying groups of genes that are co-regulated in cancer, clustering can help to identify novel targets for therapy, and provide a more accurate diagnosis for patients.

ALL (Acute Lymphoblastic Leukemia) and AML (Acute Myeloid Leukemia) are two different types of leukemia, a cancer of the blood and bone marrow.

ALL is a type of cancer that affects the lymphoid cells, which are a type of white blood cell involved in the immune response. In ALL, the bone marrow produces large numbers of immature lymphoid cells, which crowd out the normal blood-forming cells and can spread to other parts of the body.

AML, on the other hand, is a type of cancer that affects the myeloid cells, which are a type of white blood cell involved in the formation of red blood cells, platelets, and some types of immune cells. In AML, the bone marrow produces large numbers of immature myeloid cells, which crowd out the normal blood-forming cells and can spread to other parts of the body.

The two types of leukemia differ in their clinical presentation, treatment, and prognosis. ALL typically affects children and younger adults, and is generally considered more treatable than AML. AML, on the other hand, affects a broader age range and is generally considered more aggressive than ALL.

Gene expression analysis has been used to study the differences between ALL and AML, and has revealed distinct patterns of gene expression that are characteristic of each disease. These findings can help to guide the development of targeted therapies for each subtype of leukemia, and provide a more accurate diagnosis for patients.

Survey of existing work related to your project idea:

Cancer is a complex disease characterized by uncontrolled cell growth and the ability of cancer cells to invade normal tissues. Accurate classification of different types of cancer is crucial for effective diagnosis, treatment, and prognosis. Gene expression analysis, which measures the levels of gene expression in a sample, has emerged as a powerful tool for cancer classification, particularly for differentiating between different types of leukemia. In this literature survey, we will focus on the use of gene expression analysis and clustering methods for classifying Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL).

Gene expression analysis is based on the idea that different types of cancers have distinct patterns of gene expression, which can be used to classify cancers into different subtypes. Microarray technology has become widely used for gene expression analysis, allowing for the simultaneous measurement of the expression levels of thousands of genes in a single sample. The gene expression data obtained from microarrays can be used to develop classification algorithms that can differentiate between different types of cancer based on the expression levels of specific genes.

Clustering methods are widely used in gene expression analysis to classify samples into different subtypes based on their gene expression profiles. The most commonly used clustering method is the k-means algorithm, which groups samples into k clusters based on their similarity in gene expression. However, k-means can be sensitive to the initial conditions, and different runs of the algorithm can lead to different solutions. To address this issue, several modifications of the k-means algorithm have been proposed, including the deterministic k-means (DK-means) algorithm [3].

The first paper, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring" [1], focuses on the use of gene expression monitoring to classify cancers into different subtypes. The authors applied their method to a large dataset of microarray gene expression profiles and identified several new molecular classes of cancer, including AML and ALL. They then used this information to develop a class prediction algorithm based on gene expression monitoring. The algorithm was trained on the gene expression data from the AML and ALL samples and used to classify new samples into either AML or ALL. The results showed that the class prediction algorithm was highly accurate, with a sensitivity of 85% and a specificity of 94%.

The second paper, "Gene expression clustering using local neighborhood-based similarity measures" [2], presents a new method for gene expression clustering that takes into account local neighborhood information. The authors used this method to analyze a dataset of gene expression profiles from AML and ALL samples and showed that their approach was effective in clustering the samples into two distinct groups. The local neighborhood-based similarity measures were found to be more effective than traditional measures, such as Euclidean distance, in separating the AML and ALL samples. The results of this study demonstrate the importance of considering local neighborhood information in gene expression clustering and the potential for improving cancer classification accuracy.

The third paper, "DK-means: a deterministic k-means clustering algorithm for gene expression analysis" [3], introduces a deterministic k-means clustering algorithm for gene expression analysis, called DK-means. The authors applied their algorithm to a dataset of gene expression profiles from AML and ALL samples and found that it was effective in separating the samples into two distinct groups. The results showed that the DK-means algorithm was more accurate than traditional k-means algorithms in separating the AML and ALL samples, with a sensitivity of 95% and a specificity of 96%.

Overall, these three papers demonstrate the potential of clustering methods in gene expression analysis for cancer classification, specifically in differentiating between AML and ALL. They show that gene expression data can be used to accurately classify cancer samples into different subtypes and that clustering methods are effective in separating samples into distinct groups.

However, it is important to note that gene expression analysis is just one aspect of cancer classification and that a multi-disciplinary approach is required for a more comprehensive understanding of cancer subtypes. Additionally, the development of more advanced clustering methods and the integration of additional data sources, such as genomic and epigenetic data, will likely lead to further improvements in cancer classification based on gene expression analysis.

In conclusion, gene expression analysis is a valuable tool for cancer classification and clustering methods have proven to be effective in differentiating between AML and ALL. Further research in this area will likely lead to more accurate and comprehensive cancer classification methods based on gene expression data.

Can you use the basic graph theory in your project idea:

Basic graph theory can be used in the project of classifying AML and ALL using gene expression analysis and clustering methods in several ways.

1. Graph-based clustering: Gene expression data can be represented as a graph, where each sample is a node and the edges represent the similarity between the samples based on their gene expression profiles. Graph-based clustering algorithms, such as Spectral Clustering, can then be applied to this graph to cluster the samples into different groups.
2. Network analysis: Gene expression data can also be represented as a network, where genes are nodes and edges represent the relationship between the genes based on their co-expression patterns. Network analysis techniques, such as centrality measures, can be used to identify key genes or modules that are associated with AML and ALL and to investigate the relationships between the genes in the network.
3. Gene set analysis: Graph theory can be used to perform gene set analysis, which involves testing the hypothesis that a group of genes is differentially expressed between AML and ALL. This can be done by constructing a graph where each gene is a node and edges represent the similarity

between the genes based on their expression patterns. Clustering algorithms can then be applied to the graph to identify groups of genes that are differentially expressed between AML and ALL.

In conclusion, graph theory provides a useful framework for representing and analyzing gene expression data and can be used to enhance the accuracy and interpretability of gene expression-based cancer classification. By incorporating graph-based clustering, network analysis, and gene set analysis, we can gain deeper insights into the relationships between genes and gain a better understanding of the underlying biological mechanisms involved in the development of AML and ALL.

References:

[1] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.

[2] Jothi, R., Mohanty, S. K., & Ojha, A. (2021). Gene expression clustering using local neighborhood-based similarity measures. Computers & Electrical Engineering, 91, 107032.

[3] Jothi, R., Mohanty, S. K., & Ojha, A. (2019). DK-means: a deterministic k-means clustering algorithm for gene expression analysis. Pattern Analysis and Applications, 22, 649-667.