Gene Expression Data Analysis using Rough K-Means Clustering Method

P. Rajalakshmi
Department of Computer Science
Periyar University, Salem
rajalakshmimsc92@gmail.com

K. Thangavel
Department of Computer Science
Periyar University, Salem
drktvelu@yahoo.com

E. N. Sathishkumar
Department of Computer Science
Periyar University, Salem
drensathish@gmail.com

P. S. Raja
Department of Computer Science
Periyar University, Salem
Psraja5@gmail.com

Abstract - Data mining is the process of finding patterns in large datasets. Cluster analysis is an important part of the data mining community. The traditional clustering algorithm is slow in convergence and sensitive to the initial value in large datasets. Data clustering plays an important role in many disciplines including data mining, machine learning, pattern recognition and other fields. Cluster analysis is a popular data analysis and data mining technology. High quality and fast clustering algorithms play a vital role for users to navigate, effectively organize and summarize the data. Data mining is the process of finding patterns in large datasets. Cluster analysis is an important part of the data mining community. The traditional clustering algorithm is slow in convergence and sensitive to the initial value in large datasets. Data clustering plays an important role in many disciplines including data mining, machine learning, pattern recognition and other fields. Cluster analysis is a popular data analysis and data mining technology. High quality and fast clustering algorithms play a vital role for users to navigate, effectively organize and summarize the data. In this research work, K-Means clustering algorithm and Rough K-means clustering algorithm have been studied. Rough K-Means algorithm deals with lower and upper boundary approximations. In this paper, four datasets have taken from National Center for Biotechnology and Information (NCBI) Gene Expression Omnibus datasets. The entire datasets contain missing values and empty space or undefined values, these all are handled by filtering methods such as Gene Variance filter, Gene Low absolute variance filter, Gene Entropy filter. Then pre-processed genes have been given as an input to K-Means and Rough K-Means clustering algorithms to cluster the similar kind of genes. Comparative analyses have performed and it observed that the Rough K means algorithm selects the highly expressed genes. The Rough K Means algorithm will generate lower and upper approximations according to the mathematical property of Rough Set Theory. Then the lower and boundary values have given as an input to the Quick Reduct algorithm to select the genes. Gene Ontology weighting methods such as Biological process, Molecular Function, and Cellular components use these selected genes for finding the Biological significance. The query gene connects the entire network.

Keywords – Rough K-Means, Quick Reduct, Gene Entropy Filter, Gene Variance Filter, Gene Low Val Filter

1. INTRODUCTION

Knowledge Discovery in Databases is a process of deriving hidden knowledge from data, KDD consist of several phases like Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern Evaluation, Knowledge Representation. Data mining is an important phase of knowledge discovery. Data mining is a technique, which is used to find new hidden and useful patterns of knowledge from the data. There are several datamining techniques such as Concept Description, Association, Classification, Prediction, Clustering and Sequence Discovery to find useful patterns (E. N. Sathishkumar, K. Thangavel, & T. Chandrasekhar, 2013). The Knowledge Data Discovery includes Data selection, Data Cleaning/Preprocessing, Data Reduction, Data mining and Interpretation of the results. This process is interactive, because it requires user participation, and iterative because it allows for going back to a previous phase and then proceeding with the knowledge discovery process.

The data mining process has been applying in the most different areas. Recently, the data mining approach has been used for the gene expression data analysis obtained by the microarray approach for the classification and clustering task. The analysis of data type presents certain particularities comparing with other databases,

considering that the number of samples is small (in the order of dozens of samples) and with a high number of attributes (that correspond to each one of the genes that has its expressed pattern in each one of the samples, typically in order of thousands of genes). (Ashima Gawar, 2014)

Data Preprocessing is a data mining technique that involves the transformation of raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepare raw data for further processing.

Data cleaning is a process, which deals with missing values, smoothing the noisy data, or resolving the inconsistencies in the data. Data with different representations have put together and conflicts within the data are resolved. Data transformation is a process of converting data from one form to another using normalization, aggregation, and generalization. Data reduction obtains the reduced representation of the data that is much smaller in volume but provides the same analytical results. Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals (K.Anitha & P. Venkatesan, 2013).

Clustering has been using in a number of applications such as biology, engineering, medicine and data mining. Cluster analysis of gene expression data has proved to be a useful tool for identified the co-expressed genes. Clustering gene experiments results are usually present in the form of a data matrix in which rows represent genes and column represent samples or conditions (Daxin Jiang, Chun Tang, & Zhang., 2004). Each entry in the matrix is a measure of the expression level of a particular gene under specific conditions. Analysis of these data sets reveals genes of relationships between genes (Erfaneh Naghieh & Yong hong Peng). Co-expressed genes can be group into clusters based on their expression patterns of gene based clustering and sample based clustering. In gene based clustering, the genes are treated as the objects, while the samples are the features. In sample based clustering, the samples can be partition into homogeneous groups where the genes are regarded as features and the samples as objects (K. Thangavel, T. Chandrasekhar, E.Elayaraja, & E.N. Sathishkumar, 2013), (C. Velayutham & K.Thangavel, 2011).

In this Paper, K-Means clustering algorithm and Rough K-means clustering algorithm have studied. Rough K-Means algorithm deals with lower and upper boundary approximations. In this work, four datasets have taken from National Center for Biotechnology and Information Gene Expression Omnibus datasets. The entire datasets contain missing values and empty space or undefined values, these all are handled by filtering methods such as Gene Variance filter, Gene Low absolute variance filter, Gene Entropy filter. Then pre-processed genes have given an input to K-Means and Rough K-Means clustering algorithms to cluster the similar kind of genes. Comparative analyses have performed and it observed that the Rough K- means algorithm selects the highly expressed genes. The Rough K Means algorithm will generate lower and upper approximations according to the mathematical property of Rough Set Theory. Then the lower and boundary values have given as an input to the Quick Reduct algorithm to select the genes. Gene Ontology weighting methods such as Biological process, Molecular Function, and Cellular components uses these selected genes for finding the Biological significance. The query gene connects the entire network.

In Quick Reduct algorithm, we remove the attributes so that the set we get after reduction provides the same prediction of the decision feature as the original set which is achieved by comparing equivalence relations generated by sets of attributes. The attribute selected for the first time is to be included in the Reduct set in the Quick Reduct algorithm is the degree of dependency of that attribute which is not equal to zero. The algorithm tries to find out a minimal Reduct without generating all possible subsets. Initially, we take an empty set and add in the empty set R those attributes that will result in the greatest increase in dependency value one by one until we get the maximum possible value for the dataset. (Nisha Singh, Khushboo Guliani, & Prashant Prabhat, 2013).

Datasets downloaded from National Center for Biotechnology Information online repository from GEO (Gene Expression Omnibus). In this works, four types of datasets have used, such as Leukemia, Breast Cancer, Sickle cell anemia disease, and Liver datasets. Filtering methods used in the Preprocessing. Three types of filtering methods have used, such as Gene variance filter, Gene entropy filter, Low absolute variance filter. The Rough K-Means algorithm has used for clustering. All datasets have been divided into two clusters (K=2). Quick Reduct algorithm has used for feature selection.

Section 2 describes the literature survey. Section 3 describes Preprocessing technique based on filtering method for gene expression datasets. Section 4 contains the overview of clustering. The K-Means and Rough K-Means algorithms have studied and implemented. Section 5 describes the Quick Reduct algorithm. Section 6 presents Biological Significance used Gene Mania Gene Ontology results such as Biological process, molecular significance, Cellular Components. Section 7 provides the comparative analysis. Section 8 presents Conclusion.

2. LITERATURE REVIEW

2.1. List of Related works

The analysis of data type presents certain particularities comparing with other databases, considering that the number of samples is small (in the order of dozens of samples) (Helyane Bronoski & Julio Caser Nievola, 2007) and with a high number of attributes. (K.Anitha & P. Venkatesan, 2013), Co-expressed genes can be grouped into clusters based on their expression patterns of gene based clustering and sample based clustering. In gene based clustering, the genes treated as the objects, while the samples are the features. In sample based clustering, the samples can be partition into homogeneous groups where the genes are regarded as features and the samples as objects.

Nisha Singh et.al, had proposed several of type of preprocessing filtering methods (Nisha Singh, Khushboo Guliani, & Prashant Prabhat, 2013). Gene entropy filter, Gene variance filter, Gene low absolute value filter.

In Clustering gene experiments results are usually presented in the form of a data matrix in which rows represent genes and column represent samples or conditions (T. Chandrasekhar, K. Thangavel, & E.N. Sathishkumar, 2013). There are several data mining functions such as Concept Description, Association Rules, Classification, Prediction, Clustering and Sequence Discovery to find useful patterns (Lavanya & Rani, 2011). Each entry in the matrix is a measure of the expression level of a particular gene under specific conditions. Analysis of these data sets reveals genes of relationships between genes.

Feature selection is an important operation in processing the data stored in gene microarrays (Tomasz Latkowski & Stanislaw Osowki, 2015). (K. Thangavel, T. Chandrasekhar, E.Elayaraja, & E.N. Sathishkumar, 2013), Co-expressed genes can be group into clusters based on their expression patterns of gene based clustering and sample based clustering. In gene based clustering, the genes are treated as the objects, while the samples are the features.

In sample based clustering, the samples can be partitioned into homogeneous groups where the genes are regarded as features and the samples as objects https://www.techopedia.com/definition/14650/data preprocessing, Data preprocessing is a data mining technique that involves the transformation of raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepare raw data for further processing steps in data mining.

C. Velayutham, K.Thangavel had proposed Quick Reduct algorithm (C. Velayutham & K.Thangavel, 2011). In Quick Reduct algorithm, first they remove the attributes so that the set get after reduction provides the same prediction of the decision feature as the original set which is achieved by comparing equivalence relations generated by sets of attributes. The attribute selected for the first time is to be included in the Reduct set in the Quick Reduct algorithm is the degree of dependency of that attribute which is not equal to zero. The algorithm tries to find out a minimal Reduct without generating all possible subsets. Initially, they take an empty set and add in the empty set R those attributes that will result in the greatest increase in dependency value one by one until it gets the maximum possible value for the dataset.

Microarray experiments contain 103 to 104 genes, and this number is expected to reach the order of 106 (Daxin Jiang, Chun Tang, & Zhang., 2004). One of the characteristics of gene expression data is that it is meaningful to cluster both genes and samples. Quick Reduct algorithm searches for a minimal subset without exhaustively generating all possible subsets. The search begins with an empty subset, dependency value is add iteratively. This process continues until the search produces its maximum possible dependency value for that dataset ($\gamma_C(D)$). This type of search does not guarantee a minimal subset and may only discover a local minimum. (Pawlak, 2002), the Rough set theory is a new mathematical approach to intelligent data analysis and data mining.

3. PREPROCESSING TECHNIQUES

3.1. Preamble

A very large proportion of gene expression datasets contain continuous variables. It needs to discretize the continuous data of a dataset before applying data mining methods based on Filtering. One solution to this problem is to partition numeric variables into a number of sub-ranges and treat each such sub-range as a category. This process of partitioning continuous variables into categories is termed as discretization.

One of the interesting features of a microarray experiment is the fact that the group information on a large number of genes. These issues will affect biologist in many ways and we face a lot of problems while go for convergence. So, the dimensionality reduction of gene expression datasets should be consider. One of the characteristics of gene expression data is that it is meaningful to reduce the dimension in both genes and samples, but this work performs only the gene based clustering. The gene selection has one different approach such as Filtering Approach. There are three types of filtering methods are used in this paper.

3.2. Filtering Methods

There are three types of filtering methods such as Gene entropy filter, Gene variance filter, and Gene low absolute value filter.

3.2.1 Gene Entropy Filter

The effectiveness of the gene has calculated by using entropy filter method. Entropy measures the uncertainty of a random variable. For the measurement of the interdependency of two random genes X and Y, we used a direct function: Mask, Fdata, FNames, = gene entropy filter (Data, Names)

3.2.2 Gene Variance Filter

Gene profiling experiments have the gene that exhibits little variation in the profile and are generally not of interest in the experiment. These genes are commonly removed from the data, For the measurement of interdependency of two random genes X and Y, we used a direct function: Mask, FData, FNames = gene varfilter (Data, Names)

3.2.3 Gene Low Val Filter

Gene expression profile experiments have data where the absolute values are very low. The quality of this type of data is often bad due to large quantization errors or simply poor spot hybridization (Lavanya & Rani, 2011). For the measurement of the interdependency of two random genes X and Y, we used a direct function: Mask, FData, FNames = genelowvalfilter (Data, Names).

Table 1. Filtering results for Preprocessing

Gene Data Sets Name	Original Data sets	Gene Variance Filter	Low Absolute Expression Value Filter	Gene Entropy Filter
Platelets	54,675	49,207	4,063	1,807
Leukemia	22,283	20,055	18,049	13,807
Breast Cancer	47,316	42,583	38,325	11,440
Liver	45,281	38,753	31,678	11,453

In this Section, Gene variance filter method, Gene Low value filter method and Gene Entropy filter method have been studied and applied to all the four data sets to filter the genes for further processing. The preprocessing results are listed in the table1 and figure1 The filtered gens has clustered by using K-Means and Rough K-Means method in the next Section in order to identify the highly expressed genes.

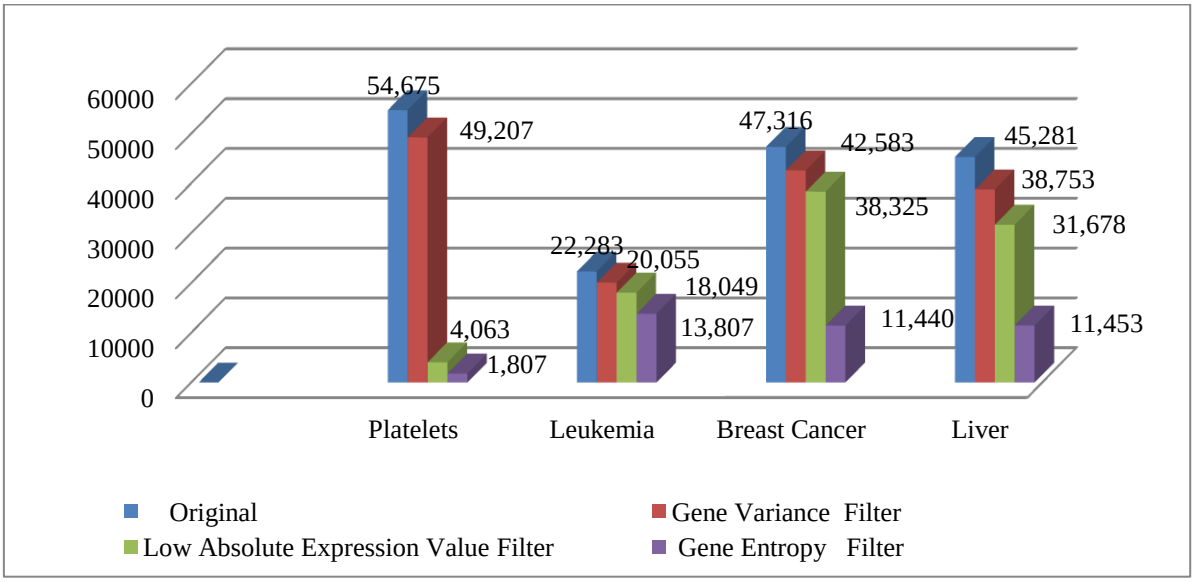

Figure 1. Accuracy for preprocessing Techniques

4. CLUSTERING METHODS

4. 1. Clustering Overview

Clustering is a process of grouping similar objects. In other words, similar objects are the group in one cluster and dissimilar objects are the group in another cluster. This clustering concept has introduced and studied by several researchers. It is the process of grouping data objects into a set of disjoint classes, called clusters, so the "objects within the same class have high similarity to each other, while objects in separate classes are more dissimilar". Clustering is an example of un-supervised classification. Unsupervised means that clustering does not rely on predefined classes and training examples while classifying the data objects. Microarray experiments contain 103 to 104 genes, and this number is expected to reach the order of 106. One of the characteristics of gene expression data is that it is meaningful to cluster both genes and samples. (Sauravjoyti Sarmah & Dhruba K. Bhattacharyya, 2010)Clustering gene expression data can be categorized into three groups and they listed below:

4.1.1 Gene Based Clustering

In this type of clustering, genes are treated as the objects, while samples as the features. The purpose of gene-based clustering is to group together co-expressed genes that indicate co-function and co-regulation.

4.1.2 Sample Based Clustering

In this type of clustering, samples are the objects and genes are features. Within a gene expression matrix, there is usually particular macroscopic phenotype of samples related to some diseases or drug effects, such as diseases sample, normal samples or drug treated sample. The goal of sample based clustering is to find the phenotype structures or sub-structure of the sample.

4.1.3 Subspace Clustering

In this type of clustering, the job is to find subsets of objects such that the objects appear as a cluster in a subspace formed by a subset of the features. In subspace clustering, the subsets of features for various subspace clusters can be different. Two subspace clusters can share some common objects and features, and some objects may not belong to any subspace cluster. In this Section, gene-based clustering has been used to group co-expressed genes (https://www.techopedia.com/definition/14650/data-preprocessing.).

4. 2. K-Means Clustering

Among the various clustering algorithms, K-Means is one of the most popular methods used for data analysis due to its good computational performance. It is well known that K-Means might converge to a local optimum,

and its result depends on the initialization process. The K-Means algorithm used to cluster the objects into K partitions using the distance between the objects. Here, K is the number of clusters and provided by the user. The idea is to choose the random cluster centroids, one for each cluster. These centroids are preferred to be as far as possible from each other. Starting points affect the clustering process and results. After that, each point will be taken into consideration to calculate similarity with all cluster centroids through a distance measure, and it will be assign to the most similar cluster, the nearest cluster centroid. When this assignment process is over, a new centroid will be calculated for each cluster using the points in it. For each cluster, the mean value will be calculated for the coordinates of all the points in that cluster and set as the coordinates of the new centroid.

Once the algorithm gets the K new centroid that changes their location step by step until no more changes are made. When the centroids do not move anymore or no more errors exist in the clusters, the algorithm shows the cluster has reached the optimum. Finally, this algorithm aims at minimizing an objective function. K-Means is one of the most commonly used clustering methods and has a wide application in microarray studies.

Algorithm 1: K-Means Clustering Algorithm

Input : Set of sample patterns $\{x1, x2, \ldots, x_m\}$, $x_i \in R^n$

Output : Set of code vectors of quantization $z_1, z_2, \ldots, z_K$, which are centroid of the clusters $\{C_1, C_2, \ldots, C_K\}$.

Step 1: Choose K initial cluster centers $z_1, z_2, \ldots, z_K$ randomly from the m patterns $\{x_1, x_2, \ldots, x_m\}$ where $K < m$.

Step 2: Assign pattern x_i to cluster C_j, where $i = 1, 2, \ldots, m$ and $j \in \{1, 2, \ldots, K\}$,

if and only if $\|x_j - z_j\| < \|x_j - z_p\|$, $p = 1, 2, \ldots, K$ and $j \neq p$.

Ties are resolved arbitrarily. And compute cluster centroid for each point x_i as follows,

$z_i = (1/n_i) \sum x_j$,

$i = 1, 2, \ldots, K$.

$x_j \in C_i$, Where n_i is the number of elements belongs to cluster C_i.

Step 3: Assign each pattern x_i to cluster C_j, where $i = 1, 2, \ldots, m$ and $j \in \{1, 2, \ldots, K\}$

if and only if $\|x_j - z_j\| < \|x_j - z_p\|$, $p = 1, 2, \ldots, K$ and $j \neq p$,

where $\| \bullet \|$ is an Euclidean metric norm.

Ties are resolved arbitrarily, without changing the cluster centroid z_j, $j = 1, 2, \ldots, K$

Step 4: Stop.

4.2.1 Limitations of K-Means clustering

a) Since K-Means clustering starts with random seed points, the end result will not be the same and will depend on the initial random vector.
b) K-Means clustering needs the number of clusters from the uses and forces all the genes/samples to fit on those defined number of clusters.
c) Does not work well with non-globular clusters. Non-globular clusters are those whose boundaries are not well defined.

4.2.2 Experimental Dataset

Sickle Cell Platelets, Leukemia, Breast cancer and Liver gene expression data sets are considered in this Section for analysis of the K-Means and Rough K-Means algorithms. These data sets are available in NCBI (http://www.ncbi.nlm.nih.gov/projects, Series: GSE11524, GSE9476, GSE65517, GSE39549).

4. 3. Performance of K-Means Algorithm

In this section, we analyse the performance of the K-Means algorithm four types gene expression data set. They entire gene expression dataset is clustered by the proper grouping of similar genes or co-expressed genes by applying K-Means algorithm. One has to specify the number of clusters well in advance. The optimal number of clusters is difficult to determine, because it may depend on different sets of genes under investigation. In this study, the number of clusters is chosen to be two (K = 2).

Table 2. Results of K-Means Clustering Algorithm

S. No	K –Value	Data Set Names	No. of Clusters	No. of Genes
1	K = 2	Sickle Cell Platelets	Cluster 1	423
			Cluster 2	1384
2	K = 2	Leukemia	Cluster 1	6201
			Cluster 2	7606
3	K=2	Breast Cancer	Cluster 1	11312
			Cluster 2	128
4	K = 2	Liver	Cluster 1	8573
			Cluster 2	2880

The results of K-Means clustering for K = 2 are tabulated in Table 2 along with the number of genes in each cluster.

First, the dataset is divided into two different clusters, which are uses to differentiate between normal and highly expressed genes. K-Means algorithm chooses two genes as random cluster center when K = 2, one for each cluster. In this case, we obtained two gene clusters; 6201 genes are placed in cluster 1 and 7606 genes assigned in cluster 2 for Leukemia datasets. Then we obtained two gene clusters; 423 genes placed in cluster 1 and 1384 genes assigned in cluster 2 for Sickle cell platelets datasets. We obtained two gene clusters; 8573 are placed in cluster 1 and 2880 genes assigned to cluster 2 for Liver datasets and we obtained two gene clusters; 11312 genes are placed in cluster 1 and 128 genes are assigned in cluster 2 for Breast cancer datasets

4. 4. Rough K-Means Clustering algorithm

The Rough set theory is a new mathematical approach to intelligent data analysis and data mining (T. Chandrasekhar, K. Thangavel, & E.N. Sathishkumar, 2013). This section describes rough clustering technique and its performance analysis. A large number of genes grouped into a smaller number of clusters in order to simplify modelling and Decision making process. A decision maker can then develop guidelines and models for a group instead of individual genes. The conventional Rough K-Means clustering techniques mandate that a gene must belong to precisely one cluster. Such a requirement found to be too restrictive in many data mining applications. In practice, a gene may display characteristics of different clusters. In such cases, a gene should belong to more than one cluster and as a result, cluster boundaries necessarily overlap. Rough set representations of clusters, using algorithms such as Rough K-Means, make it possible for a gene to belong to multiple clusters with a degree of membership between 0 and 1.

Rough set is a mathematical process used to deal with uncertainty (Pawlak, 2002). When we have insufficient knowledge of precisely define clusters, we use rough sets; here, a cluster represented by a rough set based on a lower approximation and an upper approximation. The lower approximation is a subset of the upper approximation. The members of the lower approximation belong certainly to the cluster therefore, they cannot belong to any other cluster. The data objects in an upper approximation may belong to the cluster. Since their membership is uncertain, they must be a member of an upper approximation of at least another cluster.

In Rough K-means, a cluster is described by two hard approximations. Lower and upper approximation or respectively a lower approximation and a boundary region. Hence, an object has two bivalent members ship degrees to a cluster k, one for its lower approximation and one for its boundary. The Rough K-Means algorithm is described hereunder.

4. 5. Performance of Rough K-Means Algorithm

In this section, Rough K-means algorithm has used for clustering. There are two clusters, k = 2 is applied to each dataset, such as Leukemia data set, sickle cell platelets data set, breast cancer data set, Liver data set. The entire gene expression data set is divide into two groups of similar genes. The result of Rough K-means clustering for k = 2 tabulated in table 3 along with the number of genes in each clusters. Lower and boundary region of clusters have taken in this work.The data sets has taken from the NCBI repository. Breast cancer, sickle cell, Leukemia data sets are the homo sapience organism. The liver data set is the mus muscles organism.

The performance of the Rough K-Means Algorithm is the most popular data clustering techniques has analysed with microarray expression data set. A large number of genes can group into a smaller number of clusters to simplify gene selection process. We can select a high class discriminated genes from co-expressed gene group instead of individual genes. In this section, we studied and implemented the clustering approaches, such as Rough K-Means for exploring gene expression data.

Algorithm 2: Rough K-Means Clustering Algorithm

Input : Data Set with Set of sample patterns {$x1$, $x2$, …,x_m}.

Output : Lower approximation $\underline{U}(K)$ and Upper approximation $\overline{U}(K)$ of K Clusters.

Step1: Randomly assign each data object to one lower approximation$\underline{U}(K)$. By definition (property 2) the data object also belongs to upper approximation $\overline{U}K$ of the same Cluster.

Step 2: Compute the Cluster Centroid C_j

If $\underline{U}(K) \neq \emptyset$ and $\overline{U}(K) - \underline{U}(K) = \emptyset$

$$C_j = \sum_{x \in \underline{U}(K)} \frac{x_i}{|\underline{U}(K)|}$$

Else If $\underline{U}(K) = \emptyset$ and $\overline{U}(K) - \underline{U}(K) \neq \emptyset$

$$C_j = \sum_{x \in \overline{U}(K) - \underline{U}(K)} \frac{x_i}{|\overline{U}(K) - \underline{U}(K)|}$$

Else

$$C_j = W_{lower} \times \sum_{x \in \underline{U}(K)} \frac{x_i}{|\underline{U}(K)|} + W_{upper} \times \sum_{x \in \overline{U}(K) - \underline{U}(K)} \frac{x_i}{|\overline{U}(K) - \underline{U}(K)|}$$

Step 3: Assign each object to the lower approximation $\underline{U}(K)$ or upper approximation $\overline{U}(K)$ of cluster i respectively. For each object vector x, let d(X, C_j) be the distance between itself and the centroid of cluster C_j.

$d(X, C_j) = min_{1 \leq j \leq K} d(X, C_j)$.

The ratio d(X, C_i) / d(X, C_j), 1≤ i, j ≤ K is used to determine the membership of x as follow: If d(X, C_i) / d(X, C_j) ≤ epsilon, for any pair (i, j), the $x \in \overline{U}(C_i)$ and $x \in \overline{U}(C_j)$ and x will not be a part of any lower approximation. Otherwise, $x \in \underline{U}(C_i)$, such that d(X, C_i) is the minimum of 1 ≤ i ≤ K. In addition $x \in \overline{U}(C_i)$.

Step 4: Repeat Steps 2 and 3 until convergence.

In this Section, the K-Means clustering algorithm and Rough K-Means algorithm have been studied and implemented for all the four data sets described in the earlier Section. The genes obtained in each cluster has further given as input the Quick Reduct algorithm to identify the highly expressed genes in the next Section.

5. FEATURE SELECTION

Feature selection is an important operation in processing the data of gene microarrays. The most relevant genes (treated as the features) increase our understanding of the mechanism of disease formation and allow predicting the potential danger affected by such disease. The feature selection methods allow us to identify a small number of important genes that can use as biomarkers of the appropriate disease. In this Section, we use Quick Reduct algorithm for feature selection methods have examined and integrated into the final system (Helyane Bronoski & Julio Caser Nievola, 2007).

Table 3. Clustering results for Rough K-means algorithm

Datasets	Number of Clusters	Lower/Boundary	Number of genes
Sickle cell diseases	Cluster 1	Lower	419
		Boundary	1644
	Cluster 2	Lower	163
		Boundary	1388
Leukemia	Cluster 1	Lower	9125
		Boundary	11790
	Cluster 2	Lower	2017

Breast Cancer	Cluster 1	Boundary	4682
		Lower	11346
		Boundary	11404
	Cluster 2	Lower	36
		Boundary	94
Liver	Cluster 1	Lower	2893
		Boundary	11005
	Cluster 2	Lower	4138
		Boundary	12250

5.1. Overview of Quick Reduct algorithm

Quick Reduct algorithm searches for a minimal subset without exhaustively generating all possible subsets. The search begins with an empty subset, dependency value is add iteratively. This process continues until the search produces its maximum possible dependency value for that dataset (ɣc(D)). This type of search does not guarantee a minimal subset and may only discover a local minimum (Sunnyvale & Schena M., 2000).

The Quick Reduct algorithm starts with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric until this produces its maximum possible value for the dataset.

Algorithm 3: Quick Reduct

Input : C, the set of all conditional features and D the set of decision features.
Output : Selected Features Subset

Step 1: R ←{}
Step 2: **Do**
Step 3: T ← R
Step 4: ∀ X ∈ (C-R)
Step 5: **if** $\gamma_{R\cup\{x\}}(D) > \gamma_T(D)$
Where $\gamma R(D) = \text{card}(POS_R(D)) / \text{card}(U)$
Step 6: T ← R∪{x}
Step 7: R ←T
Step 8: **until** $\gamma_R(D) = = \gamma_C(D)$
Step 9: **return R**

Quick Reduct Algorithm to find the minimum subset of features, Rmin = {X |X ∈Rall, ∀Y ∈Rall,|X| ≤ |Y|}. Stepwise execution of Quick Reduct Algorithm has given below in the figure 2 (Nisha Singh, Khushboo Guliani, & Prashant Prabhat, 2013).

Table 4. Quick Reduct Algorithm Results

Dataset Name	Number of Cluster K=2	K-Means Quick Reduct Selected genes Id	Rough K-Means Quick Reduct Selected genes Id
Sickle Platelets	Cluster 1	1007_s_at	1552309_a_at, 1556588_at 1564155_x_at, 201279_s_at, 1007_s_at, 1494_f_at, 1552321_a_at, 1552400_a_at 117_at
	Cluster 2	1053_at, 1552999_a_at, 1553553_at	
Leukemia	Cluster 1	117_at, 201231_s_at, 208903_at	1053_at, 201321_s_at, 202281_at, 204435_at, 213475_s_at 117_at, 202281_at 204436_at, 213476_x_at, 201392_at 117_at, 202478_at, 205495_at,221558_s_at
	Cluster 2	1007_s_at, 202277_at, 204434_at	

Breast Cancer	Cluster 1	ILMN_1343048, ILMN_1343373	ILMN_1651210, ILMN_1651296 ILMN_1651210, ILMN_1651296
	Cluster 2	ILMN_1343049, ILMN_1343311, ILMN_1343331	ILMN-1651221, ILMN_1659086
Liver	Cluster 1	ILMN_1243094	ILMN_2454720, ILMN_2429159, ILMN_1243094
	Cluster 2	ILMN_3140158, ILMN_3093150	ILMN_1243094, ILMN_1240307, ILMN_2754227

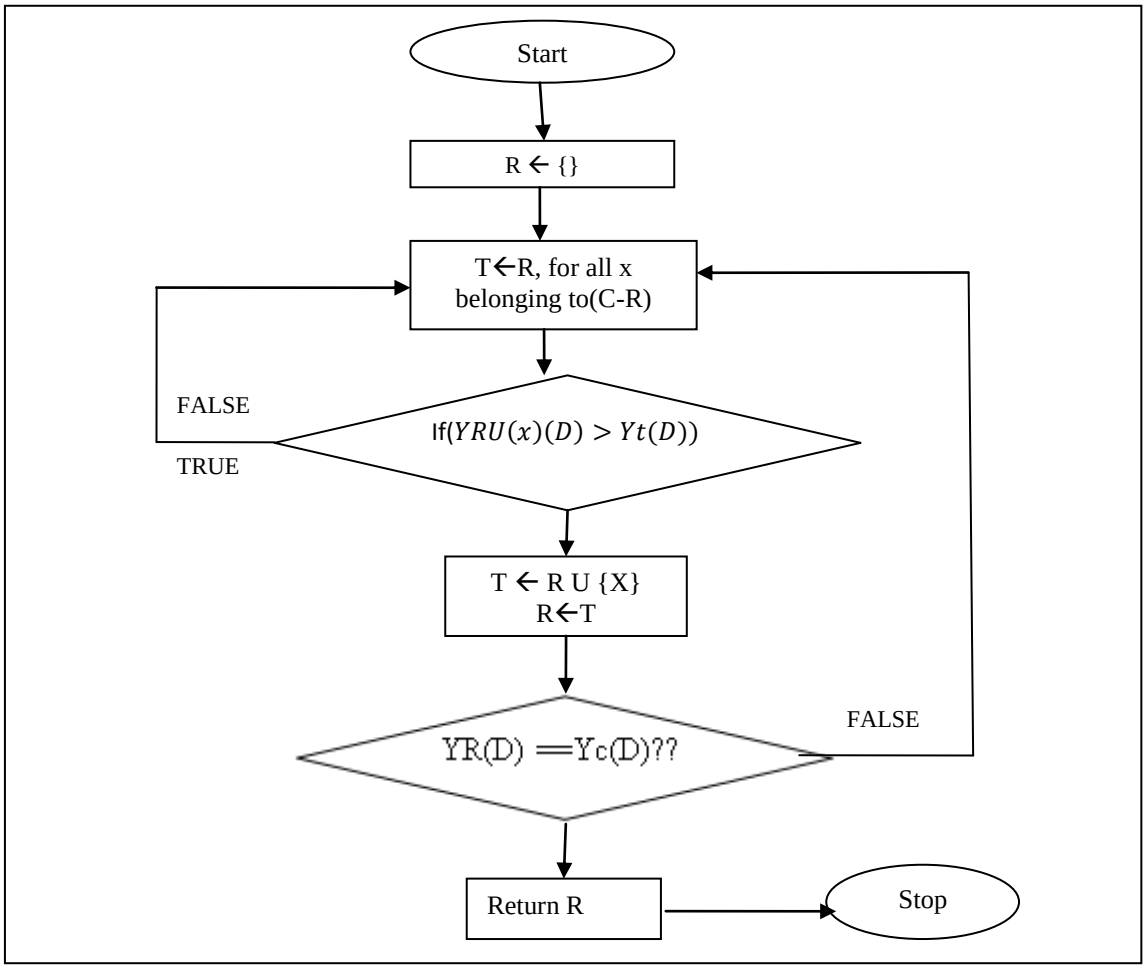

Figure 2. Stepwise Execution of Quick Reduct Algorithm

The reduced genes after applying Quick reduct are shown in the table4, has given as input to Gene Mania in-order study the biological significance of the genes of the four data sets.

6. BIOLOGICAL SIGNIFICANCE

6. 1. Preamble

Gene expression profiles generated by microarrays can help us to understand the cellular mechanism of the biological process. Once a list of differentially expressed genes has generated the next task to determine the biological significance of the genes in that list. The interactions between biological processes are very complicated. One biological process may require involvement of hundreds of genes. One gene may also be involved in many biological processes. Many genes have been studied and their biological processes have been found; however, there are still a lot of genes within biological processes whose involvement is unknown, even in well-studied organisms. In DNA microarray techniques, Gene Ontology and Gene MANIA annotations can assist in predicting biological significance for unknown genes. Gene Ontology covers three domains.

6.1.1 Biological Process

Given input gene list has related to Gene Ontology biological process. The default value of our input gene list contains less than 5 genes. The biological process operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units. Like as cells, tissues, organs, and organisms.

6.1.2 Cellular Component:

Cellular component is the parts of a cell or its extra cellular environment. Given input gene list has related to the GO cellular components. It attempts to determine whether an observed level of annotation for a group of genes is significant within the context of annotations for all genes within the genome or not.

6.1.3 Molecular function

This input genes list related to the GO Molecular function. The elemental activities of a gene product at the molecular level, such as binding or catalysis.

6. 2. Gene MANIA

Gene MANIA works best if most of the input genes are functionally related. If our query list consists of 6 or more genes, Gene MANIA will calculate gene list-specific weights. If our query list has less than 6 genes, Gene MANIA will make gene function predictions based on Gene Ontology annotations patterns.

The gene MANIA Cytoscape app is handling larger gene lists than the website. The Cytoscape app allows for using other Cytoscape features and apps with Gene MANIA data. Gene MANIA searches many large, easily available biological data sets to find related genes.

6.2.1 Gene MANIA network categories:

Gene Mania searches many large, easily available biological data sets to find related genes. Gene Mania includes protein-protein, protein-DNA, gene and protein expression data, protein domains pathways reactions and phenotypic screening profiles.

6.2.2 Co-expression (Gene expression data)

Two genes have linked if their expression levels are similar across conditions in a gene expression study. Most of these data have collected from the Gene Expression Omnibus (GEO).

6.2.3 Physical Interaction (Protein-protein interaction data)

Two gene products linked if they have found to interact in a protein-protein interaction study. These data collected from primary studies found in protein interaction databases, including BioGRID and Pathway Commons.

6.2.4 Genetic interaction (Genetic interaction data)

Two genes are functionally associated if the effects of perturbing one gene were found to be modified by perturbations to a second gene. These data are collected from primary studies and BioGRID.

6.2.5 Shared protein domains (Protein domain data):

Two gene products are like if they have the same protein domain. These data are collected from domain databases, such as InterPro, SMART, and Pfam.

6.2.6 Co-localization

Genes expressed in the same tissue, or proteins found in the same location. Two genes are like if they are both expressed in the same tissue or if their gene products are both identify in the same cellular location.

6.2.7 Pathway

(Pathway data): Two gene products are like if they participate in the same reaction within a pathway. These data are collected from various source databases, such as Reactome and BioCyc, via Pathway Commons.

6.2.8 Predicted

Predicted functional relationships between genes, often protein interactions. A major source of predicted data is mapping known functional relationships from another organism via orthology. For instance, two proteins are predicted to interact if their orthology has known to interact with another organism. In these cases, network names describe the original data source of experimentally measured interactions and which organism the interactions are mapped.

6. 3. Gene MANIA Gene Ontology Enrichment Analysis

Gene MANIA performs Gene Ontology term enrichment of the query list along with the returned gene list. If users query list consists of 6 or more genes, Gene MANIA will calculate gene list-specific weights. If users query list has less than 6 genes, Gene MANIA will make gene function predictions based on GO annotations patterns. If we enter a query gene list, such as 'DDR1', 'TIMD4', 'NEXN', 'DAB2' and 'C4', Gene MANIA will output connections between query genes and three GO networks. In this network, five query genes are represented by the largest black circles. The graph shows the local neighbourhood around the query genes as well as the top predictions. The combined network is constructed from co-expression, co-localization, pathways, genetic and physical interactions, and shared protein domains.

6. 4. Biological Significance for K-Means Clustering

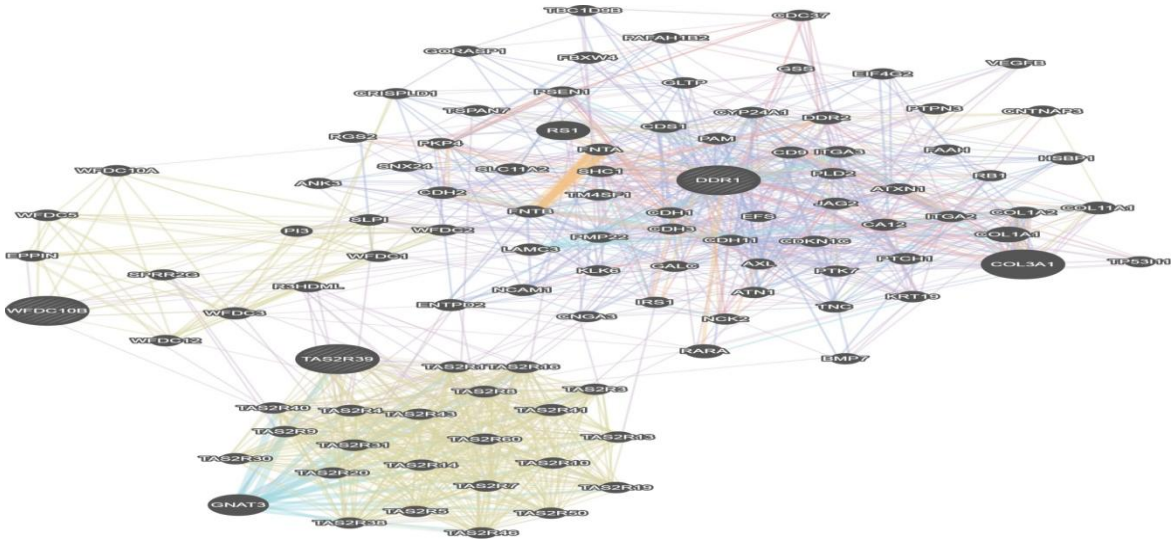

Figure 3. Biological process for Platelets datasets

Figure 3 illustrates biological process networks of query genes using Gene MANIA. According to this database, 20 genes have found as the local neighborhood around the query genes. The following distribution characterized the types of interactions extracted from the BP network: consolidated-pathways 3.02%, co-expression 24.98%, genetic interactions 1.76%, co-localization 12.82%, Pathway 3.02%, physical interactions 32.31%, predicted 20.37%, shared protein domains 0.93%. DDR1 is one of the major participated genes in the top ranked consolidated pathway

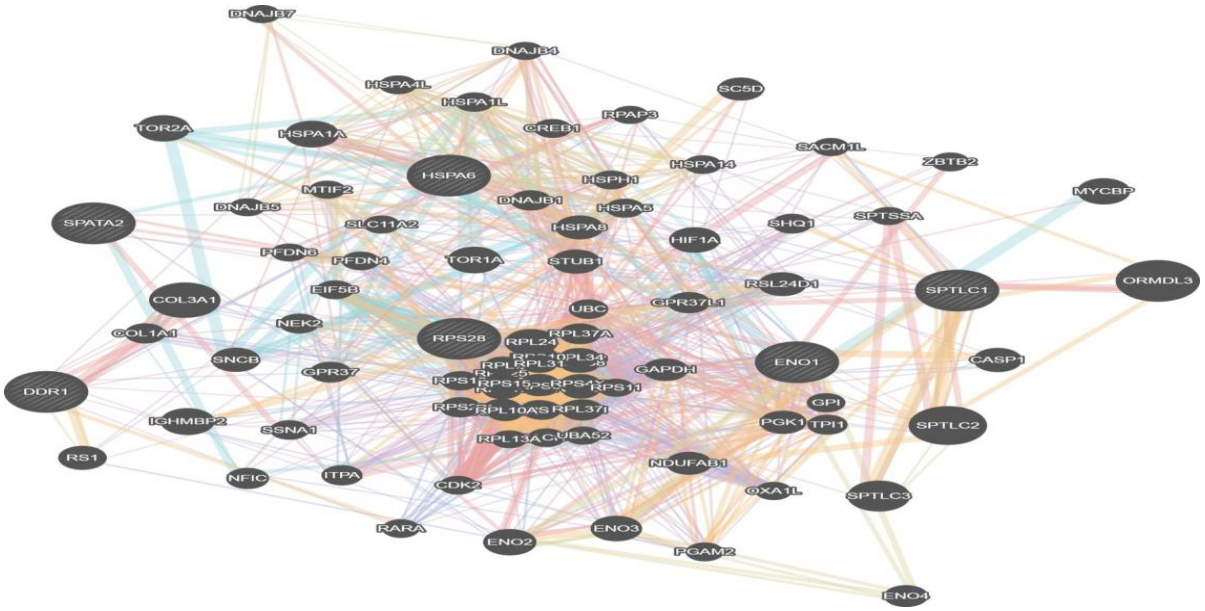

Figure 4. Biological process for Leukemia datasets

Figure 4 illustrates biological process networks of query genes using Gene MANIA. According to this database, 20 genes have found as local neighbourhood around the query genes. The following distribution characterized the types of interactions extracted from the BP network: consolidated-pathways 3.02%, co-expression 24.98%, genetic interactions 1.76%, co-localization 12.82%, Pathway 3.02%, physical interactions 32.31%, predicted 20.37%, shared protein domains 0.93%. RPS28 is one of the major participated genes in the top ranked consolidated pathway

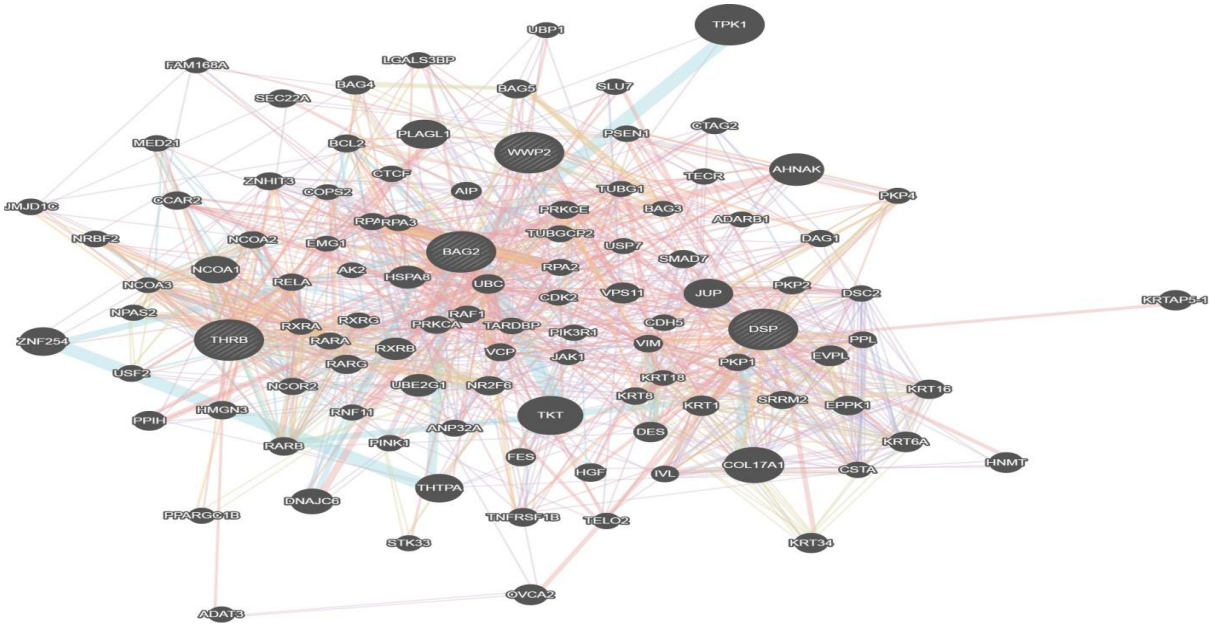

Figure 5. Biological process for Breast cancer datasets

Figure 5 illustrates biological process networks of query genes using Gene MANIA. According to this database, 20 genes have found as local neighbourhood around the query genes. The following distribution characterized the types of interactions extracted from the BP network: consolidated-pathways 3.02%, co-expression 24.98%, genetic interactions 1.76%, co-localization 12.82%, Pathway 3.02%, physical interactions 32.31%, predicted 20.37%, shared protein domains 0.93%. WWP2 is one of the major participated genes in the top ranked consolidated pathway

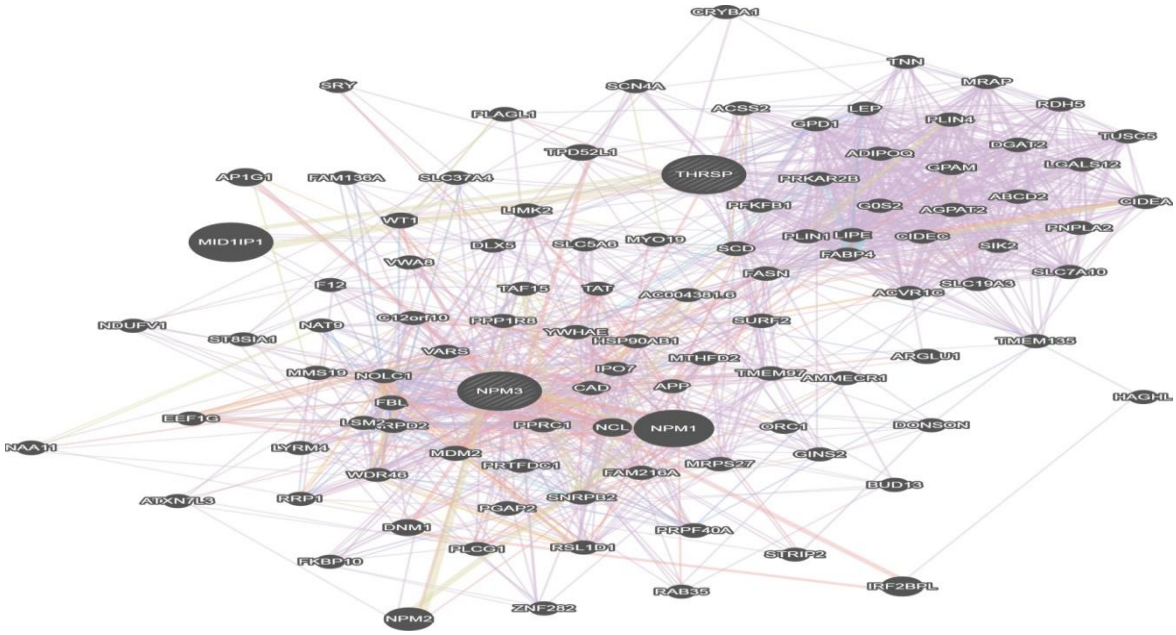

Figure 6. Biological process for Liver datasets

Figure 6 illustrates biological process networks of query genes using Gene MANIA. According to this database, 20 genes have found as local neighborhood around the query genes. The following distribution characterized the types of interactions extracted from the BP network: consolidated-pathways 3.02%, co-expression 24.98%, genetic interactions 1.76%, co-localization 12.82%, Pathway 3.02%, physical interactions 32.31%, predicted 20.37%, shared protein domains 0.93%. NPM3 is one of the major participated genes in the top ranked consolidated pathway

6. 5. Biological Significance for Rough K-Means Clustering

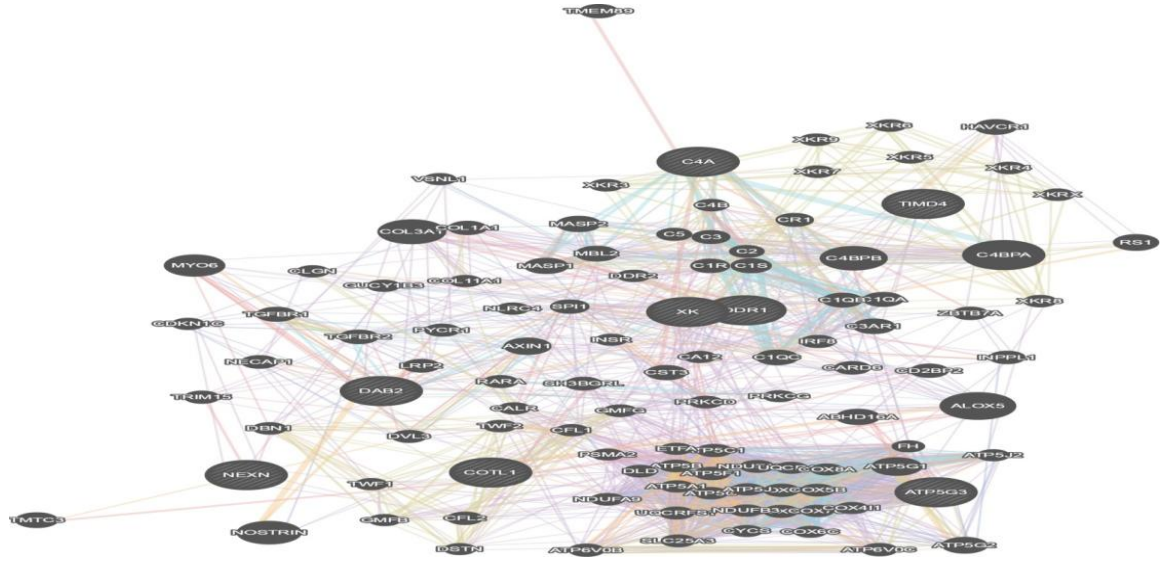

Figure 7. Biological Process for Platelets

Figure 7 illustrates biological process networks of query genes using Gene MANIA. According to this database, 20 genes have found as local neighborhood around the query genes. The following distribution characterized the types of interactions extracted from the BP network: consolidated-pathways 4.35%, co-expression 13.50%, genetic interactions 1.40%, co-localization 6.17%, Pathway 4.35%, physical interactions 67.64%, predicted 6.35%, shared protein domains 0.59%. DDR1 is one of the major participated genes in the top ranked consolidated pathway.

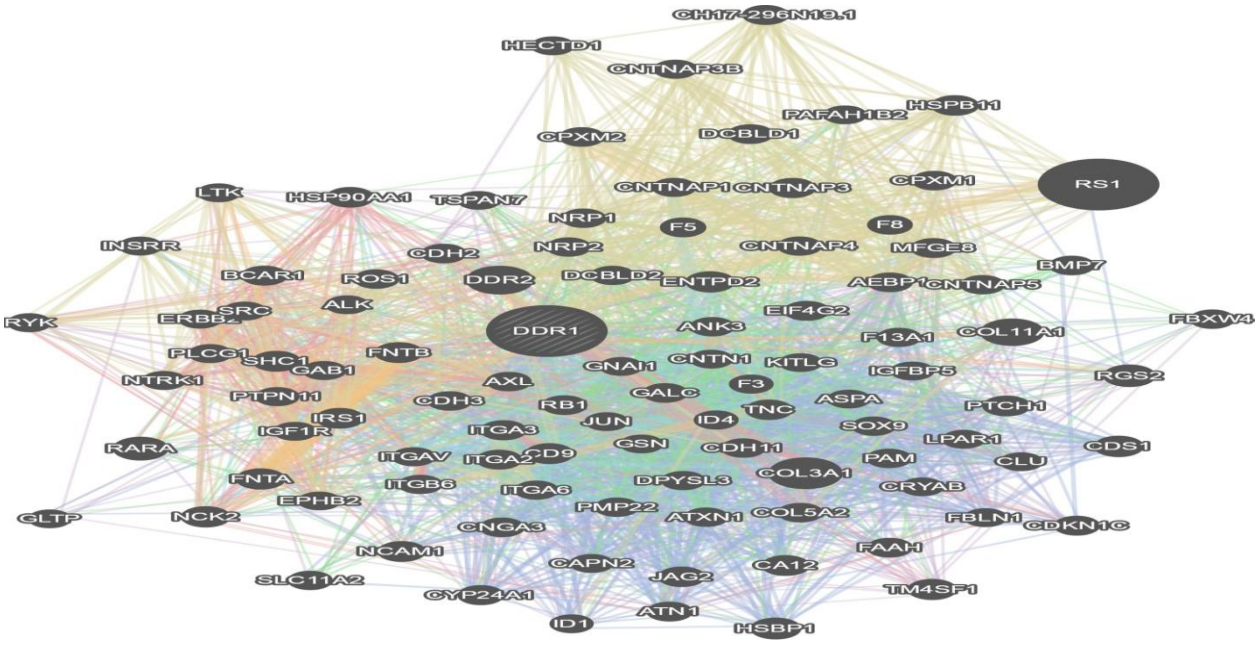

Figure 8. Biological Process for Leukemia datasets

Figure 8 illustrates biological process networks of query genes using Gene MANIA. According to this database, 20 genes have found as local neighbourhood around the query genes. The following distribution characterized the types of interactions extracted from the BP network: consolidated-pathways 4.35%, co-expression 13.50%, genetic interactions 1.40%, co-localization 6.17%, Pathway 4.35%, physical interactions 67.64%, predicted 6.35%, shared protein domains 0.59%. DDR1 is one of the major participated genes in the top ranked consolidated pathway.

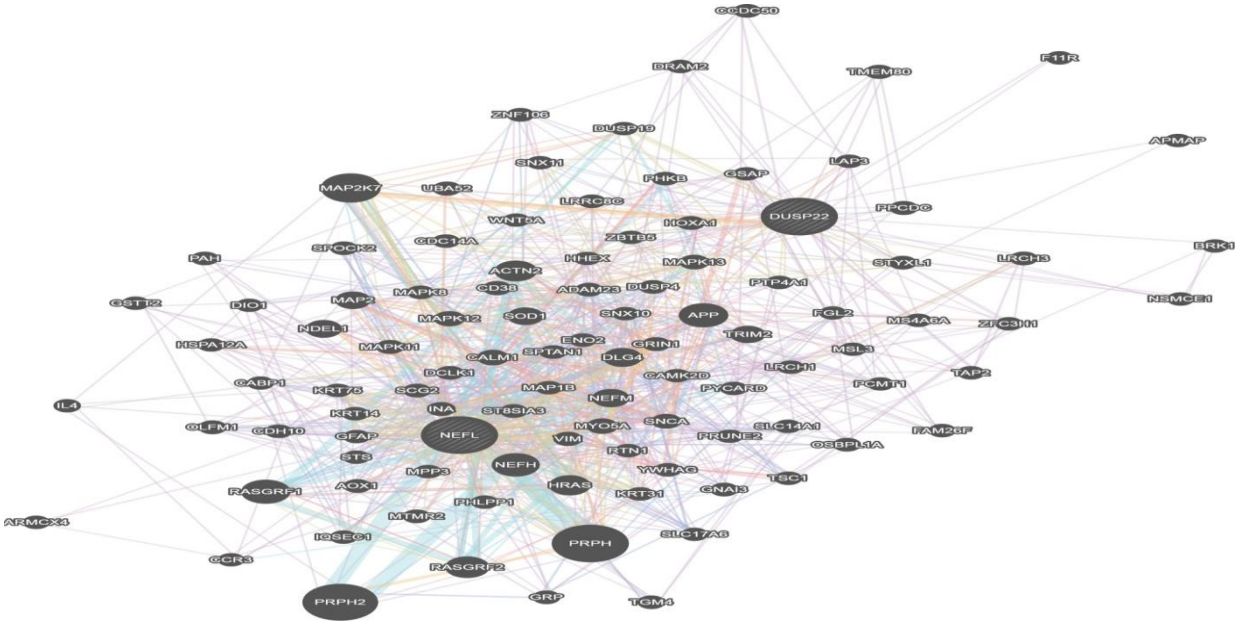

Figure 9. Biological Process for Breast Cancer

Figure 9 illustrates biological process networks of query genes using Gene MANIA. According to this database, 20 genes have found as local neighbourhood around the query genes. The following distribution characterized the types of interactions extracted from the BP network: consolidated-pathways 4.35%, co-expression 13.50%, genetic interactions 1.40%, co-localization 6.17%, Pathway 4.35%, physical interactions 67.64%, predicted 6.35%, shared protein domains 0.59%. NEFL is one of the major participated genes in the top ranked consolidated pathway.

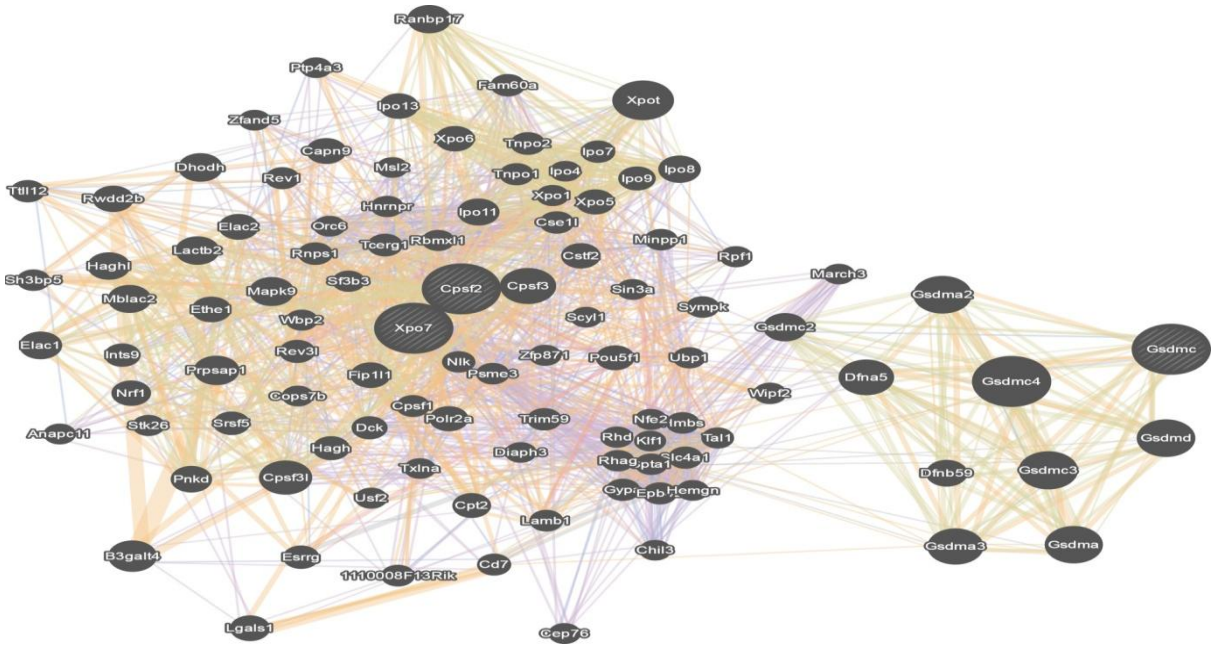

Figure 10. Biological Processes for Liver

Figure 10 illustrates biological process networks of query genes using Gene MANIA. According to this database, 20 genes have found as local neighbourhood around the query genes. The following distribution characterized the types of interactions extracted from the BP network: consolidated-pathways 4.35%, co-expression 13.50%, genetic interactions 1.40%, co-localization 6.17%, Pathway 4.35%, physical interactions 67.64%, predicted 6.35%, shared protein domains 0.59%. cpsf2 is one of the major participated genes in the top ranked consolidated pathway.

In this Section, the genes have selected using Quick Reduct algorithm from the different clusters have been given as input and identified the biological significance of the genes of the four data sets.

7. EXPERIMENTAL RESULTS AND ANALYSIS

To evaluate the classification accuracy of Quick Reduct algorithm using Rough K-Means clustering techniques. The performance of Rough K-Means is analysed through the confusion matrix these approaches were implemented. The gene expression datasets are use for our experimental results. They are four types of Datasets have used in this algorithm. Data flow diagram of this experimental analysis is given below.

7.1. Experimental Datasets

We use four gene expression datasets: leukemia, breast cancer, Sickle cell platelets and Liver datasets, which are available on the website: http://www.ncbi.nlm.nih.gov/gds/. This is Gene Expression Omnibus (GEO) online repository of gene expression data. This database stores curated gene expression Data sets, as well as original Series and Platform records in the Gene Expression Omnibus repository. Enter search terms to locate experiments of interest. Dataset records contain additional resources including cluster tools and different expression queries. These biomedical applications are also challenging problems to the machine learning and data mining community. As the file, formats of these original raw data are different from common ones used in most of machine learning software. The gene number and class contained in four datasets are listed in Table 5.

Table 5. Summary Of The Four Types Of Gene Expression Datasets And Samples

Dataset	Gene	Class	Samples
Sickle cell Platelets	1,807	Sickle cell disease/Control	30 (18/12)
Leukemia	13,807	Leukemia/Normal	64 (26/38)

Breast Cancer	11,440	Metaheustic Breast cancer/ healthy control /gram negative sepsis/ Tuberculosis	13(4/3/3/3)
Liver	11,453	Base line/normal diet/high diet	51 (3/24/24)

7.2. Proposed Model

The data flow diagram of this proposed work is shown in the figure 11.

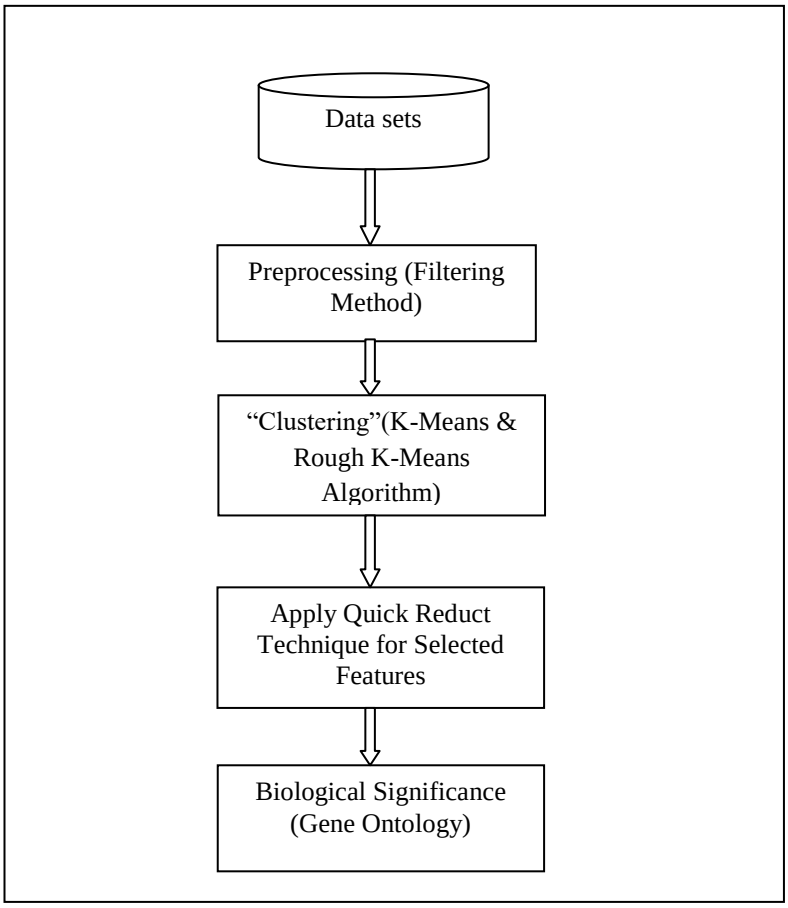

Figure 11. Data Flow Diagram For Proposed Model

7.3. Evaluation Methods

An attribute is discrete if it has a relatively small (finite) number of possible values while a continuous attribute is consider that had a very large number of possible values (infinite). Usually, the discretization process consists of two steps. First, the number of discrete intervals needs to choose. Even though there are discretization methods, which determine this number in the discretization process and this step done usually by the user either by some heuristic techniques or by running the discretization technique for different number of intervals. Second, the cut points must be determined, which does a discretization algorithm itself often did. The goal of discretization is to reduce the number of possible values a continuous attribute takes by partitioning them into a number of intervals, which done by K-Means discretization algorithm.

Here the discretized datasets consist of two types of attributes; they are conditional (C) attributes and decision attribute (D). For this, apply Quick Reduct Algorithm to find minimal subset of features, $Rmin = \{X \mid X \in Rall, \forall Y \in Rall, |X| \leq |Y|\}$. The selected genes has listed in Table 6.

By applying Quick Reduct Algorithm, in Leukemia gene dataset, gene 848 and are identified, whereas in Sickle cell platelets dataset gene 4 and 3252 are identified, in breast cancer dataset gene 2,18 and 1126 are identified, finally in liver dataset gene 1,6 are identified.

Table 6. Genes Selected By Quick Reduct Algorithm

Datasets Name	No. of Clusters K=2	Identified Attributes(Genes)	
		K-Means	Rough K-Means
Sickle Cell Platelets	Cluster 1	#1	#48,#6022,#3100
	Cluster 2	#2,#530,#949	#1,#13,#58,#107
Leukemia Cancer	Cluster 1	#4,#760	#3,#850,#4,#1810
	Cluster 2	#2,#1806,#3962	#2007,#5023,#4
Breast Cancer	Cluster 1	#1,#90	#3,#2,#18
	Cluster 2	#2,#28,#48,#85	#5,#1126
Liver	Cluster 1	#1	#2
	Cluster 2	#6,#59	#1,#6

It has observed from the biological significance of the genes selected from K-Means and Rough K-Means algorithms with Quick reduct are comparable.

8. CONCLUSION

Disease associated gene identification is one of the most important areas of medical research today. Many current methods for disease associated gene identification are based on microarray data. The rough computing model has proposed for improving the gene selection method in a simple and efficient way. In this paper, four gene datasets, such us Leukemia, Liver, Breast Cancer, and Platelet shave been used. Then we compare with Rough K-means and K-means algorithm then we get the best results in Rough K-means algorithm. K-means selected informative genes are DDR1, DDR1, WWP2, NPM3 selected from Leukemia, breast cancer, Liver, Platelets datasets. All datasets Informative genes are used to the biological significance of selected genes such as Biological Process is originated using Gene MANIA. Rough K-Means selected informative genes RFC2, TRAP1, TUBGCP2, GZMA, TRAB2, and DDR1 have selected from Leukemia gene expression dataset by applying the proposed rough set model. Then another informative genes are DUSP22, NEFL and LOC64220 are selected from Breast cancer gene expression dataset by applying proposed rough set model. Then another informative genescpsf2, xpo7, gsdmc3 and ILMN are selected from Liver gene expression dataset by applying the proposed rough set model, other informative genes TIMD4, NEXN, C4, XK, and DAB2 are selected from Liver gene expression dataset by applying proposed rough set model. The Biological Significance of selected genes such as Biological Processing originated using Gene MANIA. The proposed rough Computing method gives the best results for gene selection method.

REFERENCES

Ashima Gawar. (2014). Performance Analysis of Quickreduct, Quick Relative Reduct Algorithm and a New Proposed Algorithm. Global Journals Inc(USA), 14(4).

C. Velayutham, & K.Thangavel. (2011). Unsupervised Quick Reduct Algorithm using Rough Set Theory. Proceeding on International Conference, Emerging Trends in Computing (ICETC-2011).

D. Lavanya, & Dr.K.Usha Rani. (2011, Nov). Analysis of feature selection with classification: breast cancer datasets. IJCSE, 2, 5.

Daxin Jiang, Chun Tang, & Zhang., A. (n.d.). Cluster Analysis for Gene Expression Data: A Survey.

E. N. Sathishkumar, K. Thangavel, & T. Chandrasekhar. (2013). A New Hybrid K-Mean-Quick Reduct Algorithm for Gene Selection. International Journal of Computer, Electrical, Automation, Control and Information Engineering, 7, 2.

Erfaneh Naghieh, E., & Yong hong Peng. (n.d.). Microarray Gene Expression Data Mining: Clustering Analysis Review. Bradford.: Department of Computing,University of Bradford.

Helyane Bronoski, & Julio Caser Nievola. (n.d.). Feature Selection as a Preprocessing step for classification in gene expression data.

https://www.techopedia.com/definition/14650/data-preprocessing.

K. Thangavel, T. Chandrasekhar, E.Elayaraja, & E.N. Sathishkumar. Unsupervised Gene Expression Data Using Enhanced Clustering Method.

K.Anitha, & P. Venkatesan. (2013, August). Feature Selection By Rough Quick Reduct Algorithm. IJIRSET, 2(8).

Nisha Singh, Khushboo Guliani, & Prashant Prabhat. (2013, May). Comparison of Different Filtering Approach on Gene Expression Data for Clustering., 2.

Pawlak, Z. Rough Sets andIntelligentDataAnalysis. InformationSciences, 147, 1–12.

Sauravjoyti Sarmah, & Dhruba K. Bhattacharyya. (2010, May). An Effective Technique for Clustering Incremental Gene expression data. IJCSI International Journal of Computer Science Issues, 7(3), 3.

Sunnyvale, & Schena M. (2000). Microarray biochip technology. CA:Eaton Publishing.

T. Chandrasekhar, K. Thangavel, & E.N. Sathishkumar. (2013). Simultaneous Clustering and Feature selection method for gene expression data. World Academy of Science, 7, p. 4.

T. Chandrasekhar, K.Thangavel, & E.N.Sathishkumar. (2011, June). Verdict Accuracy of Quick Reduct Algorithm for Gene Expression Data. International Journal of Computational Intelligence and Informatics, 1, 1.

Tomasz Latkowski, & Stanislaw Osowki. (2015). Data mining for feature selection in gene expression autism data. 2015.