ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 3; Issue 6; November-December-2015; Page No. 41-49

HIERARCHICAL CLUSTERING OF LUNG CANCER MICROARRAY DATASET USING RANDOM FOREST ALGORITHM BASED ON GO ANNOTATION

Dr.R.Porkodi[1], N.Poomani[2]

[1] Bharathiar University, Department of computer science, school of computer science and Engineering, Bharathiar University, Coimbatore-46.

porkodi_r76@yahoo.co.in

[2] Bharathiar University, Department of computer science, school of computer science and Engineering, Bharathiar University, Coimbatore-46.

poomanimsc@gmail.com

Abstract

Data mining refers to collecting or mining knowledge from large amounts of data. It is used in various medical applications like tumor classification, protein structure prediction, gene selection, cancer classification based on microarray data, clustering of gene expression data, and statistical model of protein-protein interaction. The emerging research area in bioinformatics is gene enrichment analysis using GO Terms present in the genes given in any dataset. Thus, GO Terms are important feature which considered for almost all researches in this field. In this paper the random forest algorithm to compute score for each gene based on gene ontology Terms which is downloaded using BioMart package. The random forest algorithm analyze extracting most significant GO Terms, extracting top most genes based on GO Terms enrichment and extracting genes with GO Terms mapping. Further, the gene expression profiles in the dataset are clustered based on top ranked GO Terms.

Key Words: Data Mining, Random Forest, Microarray, Gene Ontology Terms, Hierarchical Clustering

1. INTRODUCTION

Data mining is the process of extracting or mining knowledge from huge amounts of data. It is also known as knowledge mining from data. Data Mining is the method of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems. It consists of more than collection and managing data, such as analysis and prediction [1].

Data mining is the process of discovering meaningful, new correlation patterns and trends by shifting through large amount of data store in repositories, using pattern recognition techniques as well as statistical and mathematical techniques. The whole process of applying a computer-based methodology, including new techniques, for discovering knowledge from data is called data mining [2].

Bioinformatics was coined by Hwa Lim in the late 1980s, and popularized in the 1990s through its relationship with the human genome project. It is an art and science concerned with the use of computing in biological research areas such as genomics, transcriptomics, proteomics, genetics, and evolution. Bioinformatics is the application of computer science and information technology in the field of biology and medicine [3].

Data mining approaches are perfectly suitable for bioinformatics and ongoing the analysis of biological data, the presence of biological signals despite high data noises, and the gap between data collection and knowledge extraction have collectively created new and exciting opportunities for data mining researchers in this field. The extensive availability of open-access biological databases has created both challenges and opportunities for developing novel knowledge discovery and data mining methods specific to molecular biology.

Microarray is used for gene expression analysis. It comprises of a tiny membrane or glass slide having samples of many regularly arranged genes. Microarray analysis can detect thousands of genes in a small sample along with the expression of those

Corresponding author: Dr.R.Porkodi

genes. Microarray Datasets are often characterized by high-dimensions and small samples.

This paper provides the detailed analysis of top ranking go terms with gene annotations such as biological process, molecular function and cellular component. This paper also shows the details of genes associated with go terms and overlapping of genes between go terms, and provides hierarchical clustering of lung cancer dataset samples based on gene expression associated with the go term. The paper is organized as follows: Review of literature is presented in section 2. In section 3, methodology of this research is discussed. The result of lung cancer discussion is presented in section 4 and the paper is concluded in section 5.

2. RELATED WORK

Microarrays that consist of ordered sets of DNA fixed to solid surfaces provide pharmaceutical firms with a means to identify drug targets. In the future, the emerging technology promises to help physicians decide the most effective drug treatments for individual patients. Gene expression microarray data are available from multiple public and commercial databases. The most widely used array repositories are Gene Expression Omnibus (GEO) Profiles, and Array Express. The GEO Profiles database provides graphic gene expression profiles derived from microarray experiments stored at NCBI's GEO microarray resource [4].

Turkheimer et al [5] identified that microarray experiment can measure the expression levels of tens of thousands of genes simultaneously. However, they can be very expensive, when it comes to data analysis; there is a recurring problem of high dimension in the number of genes and only a small number of cases. This is a characteristic shared by spectroscopic data, which additionally have high correlations between neighboring frequencies; analogously for microarray data, there is evidence of correlation of expression of genes residing closely to one another on the chromosome.

V.S. Tseng [6] describes every cell in our body contains a number of genes that specify the unique features of different types of cells. The gene expression of cells can be obtained by DNA microarray technology which is capable of showing simultaneous expressions of tens of thousands of genes. This technology is widely used to distinguish between normal and cancerous tissue samples and support clinical cancer diagnosis. Microarray learning data samples are typically gathered from often less than one hundred of patients, while the number of genes in each sample is usually more than thousands of genes.

Aniruddha Datta [7] analyzed gene is expressed if its DNA has been transcribed to RNA, and gene expression is the number of transcriptions of the DNA of the gene. This is known as transcription level gene expression. Microarrays measure the gene expression. It is one of the most popular methods to compare the expressions of a set of genes from a cell maintained in a particular test condition to the same set of genes from a reference cell maintained under normal condition. This process starts from extraction of RNA from the cells. These RNA molecules are reverse transcribed in to cDNA molecules. The cDNA from the test cell is grown in test condition and the cDNA from the reference cell is grown in normal condition.

Cutler et al. [8] presented either categorical (i.e., classification) or continuous (i.e., regression) response variables, and either categorical or continuous predictor variables, worked by growing an ensemble of regression trees based on binary recursive partitioning, where the predictor space at each tree node was partitioned based on binary splits on a subset of randomly selected predictors. At each binary split, the response data were grouped into two descendant nodes to maximize homogeneity, and the best binary split was selected. The response data for each tree were obtained through bootstrap sampling of original observations in the training set.

Tierney L [9] discussed that the future research are using random forest for the selection of potentially large sets of genes that include correlated genes, and improving the computational efficiency of these approaches; in the present work, they have used parallelization of the "embarrassingly parallelizable" tasks using MPI with the Rmpi and Snow packages for R. In a broader context, further work is warranted on the stability properties and biological relevance of this and other gene-selection approaches, because the multiplicity problem casts doubts on the biological interpretability of most results based on a single run of one gene selection approach.

Sara Alvarez [10] describes the comprehensive evaluation of random forest for classification problems with microarray data, including an assessment of the effects of changes in its parameters and we show it to be an excellent performer even in multi-class problems, and without any need to fine-tune parameters or pre-select relevant genes. We then propose a new method for gene selection in classification problems (for both two-class and multi-class problems) that uses random forest; the main advantage of this method is that it returns very small sets of genes that retain a high predictive accuracy, and is competitive with existing methods of gene selection.

Díaz-Uriarte R [11] defines the random forest has been used extensively in the biomedical domain because it is well suited for microarray data. Features will not be deleted based on one decision or one tree, but many trees will decide and confirm elimination of features. Another positive characteristic of random forest is that it is applicable to very high dimensional data with a low number of observations, a large amount of noise and high correlated variables.

A random forest (RF) [12] is one of the most popular ensemble learning methods and has very broad applications in data mining and machine learning. Prediction is often a primary goal of genomic data analyses. For example, one often needs to predict disease status such as tumor subtype using genomic markers. Random forest is a particularly appropriate tool and has been broadly used to predict clinical outcomes under various high-throughput genomic platforms.

Han-Yu Chuang et al [13] describes the gene selection method is a latest improvement in investigational molecular biology which can produce quantitative expression magnitudes for large number of genes in a single, cellular mRNA sample. All these gene expression magnitudes outline a collective profile of the sample, which can be utilized to distinguish samples from dissimilar classes such as tissue types or treatments.

3. METHODOLOGY

Microarray lung cancer dataset is analyzed using random forest classification algorithm and hierarchical clustering. The framework of the proposed research is shown in Figure. 1. Framework consists of following phases, Preprocessing, Analyzing GO Terms and Hierarchical clustering of gene expression profiles using GO Terms.

The proposed research work is to analyse the lung cancer microarray dataset using GOexpress package in bioconductor using random forest algorithm in order to identify the significant associations among lung cancer genes based on Gene Ontology (GO) Terms. Further, the gene expression profiles in the dataset are clustered based on top ranked GO Terms. The Gene Ontology terms are analysed using BioMart database and random forest algorithm. GO contains a table ranking all GO Terms related to genes in the expression dataset based on the average ability of their related genes to cluster the samples according to the predefined grouping factor.

The GO Consortium is developing three ontologies: molecular function, biological process, and cellular component, to describe attributes of gene products or gene product groups. Molecular function describes what a gene product does at the biochemical level. Biological process describes a broad biological objective. Cellular component describes the location of a gene product, within cellular structures and within macromolecular complexes.

Definitions for GO Terms are being provided as part of the development of the ontologies, Gene Ontology analysis consists of detailed analysis of top ranking GO Terms with gene annotations such as biological process, molecular function and cellular component. And also shows the details of genes associated with GO Terms and overlapping of genes between GO Terms, and provides hierarchical clustering of lung cancer dataset samples based on gene expression associated with the GO Terms.

RANDOM FOREST ALGORITHM:

Random forest algorithm is a one of the family in classification methods based on several decision trees. The main characteristic of this group of classifiers is that their components grow like a tree in a random mode. It is selected to perform both gene selection and classification of the microarray data. Improved random forest gene selection has performed better in terms of selecting the smallest subset as well as biggest subset of informative genes with lowest out of bag error rates through gene selection.

The random forest algorithm is applied to the subset of lung cancer dataset in which 1000 genes and 58 samples with 40 Adenocarcinoma and Squamous Cell Carcinoma category are considered for further analysis. This algorithm computes score for each gene based on gene ontology Terms which is downloaded using BioMart package. The random forest algorithm analyze the subset of genes based on decision tree and produce different result slots such as extracting most significant GO Terms, extracting top most genes based on GO Terms enrichment and extracting genes with GO Terms mapping.

The GOexpress R package emerged from a set of plotting function analysis across various complex multifactorial transcriptomics from both microarray and RNA-sequence platforms [14]. Functions were repeatedly used to visualise the expression profile of genes across groups of samples, to annotate technical gene identifiers from both microarray and RNA-sequence platforms. GOexpress offers an extendable set of data driven plotting functions readily applicable to the output of widely used analytic packages estimating gene expression.

The GOexpress is one of the packages in R. It is used to analyse and visualise the expression profile of genes across groups of samples, to annotate technical gene identifiers from both microarray and RNA-seq platforms (probe sets, Ensemble gene identifiers) with their associated gene name, and to evaluate the clustering of samples based on genes participating in a common cellular function or location (i.e. Gene Ontology).

GOexpress offers an extendable set of data-driven plotting functions readily applicable to the output of widely used analytic packages estimating (differential) gene expression. GOexpress is a software package developed based on real experimental datasets to ease the visualization and interpretation of multifactorial transcriptomics data by bioinformaticians and biologists, while striving to keep it simple and quick analysis.

This package contains methods to visualise the expression profile of genes from a microarray or RNA-seq experiment and offers a supervised clustering approach to identify GO Terms containing genes with expression levels that best classify two or more predefined groups of samples. Annotations for the genes present in the expression dataset may be obtained from Ensembl through the BioMart package, if not provided by the user. The random forest framework is used to evaluate the capacity of each gene to cluster samples according to the factor of interest. Finally, GO Terms are scored by averaging the rank or score of their respective gene sets to cluster the samples.

GEO (Gene Expression Omnibus) is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. The GEO dataset gives the accession number for different data sets the obtained accession number is submitted to GEO accession viewer.

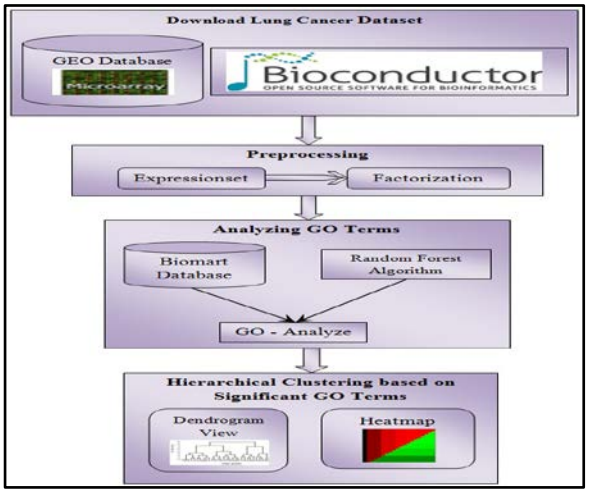

Figure 1: Methodology for proposed work

HIERARCHICAL CLUSTERING BASED ON GO TERMS:

Hierarchical clustering consists of successive joining together objects or group of objects based upon the measure of similarity or distance between the objects, The most common hierarchical methods are called bottom-up methods, starting with each object forming a cluster of size is one, At each step, the closest two clusters are joined until all objects are in a single cluster. Many hierarchical clustering methods have an interesting property that the nested Sequence of clusters can be graphically represented with a tree, called a dendogram usually; each join in a dendogram is plotted at a height equal to the dissimilarity between the two clusters which are joined. Hierarchical clustering analysis is one of the most powerful methods for the exploratory analysis

of gene-expression data. It does not need prior knowledge of the data set and provides the structure for the whole data set.

Many visualization tools are available that are of great assistance in interpreting the results of microarray experiments. Heatmap consist of small cells, each consisting of a color, which represent relative expression values. Heatmaps are often generated from hierarchical cluster analyses of both samples and genes. Often the rows represent genes of similar expression values, whereas the columns indicate different biological samples. Heatmaps offer a quick overview of clusters of genes that show similar expression values.

4. RESULTS AND DISCUSSION

The lung cancer dataset is downloaded from National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) website. This experimental dataset, The Non Small Cell Lung Cancer contains two subtypes: Adenocarcinoma (AC) and squamous Cell Carcinoma (SCC) consist of 54,675 are features and 58 samples, 40 are adenocarcinoma (AC) and 18 are squamous cell carcinoma (SCC).

4.1. RESULT OF ANALYZING THE GENES IN EXPERIMENTAL DATASET USING RANDOM FOREST ALGORITHM:

After preprocessing the subset of the experimental dataset is analyzed using random forest algorithm. This algorithm work based on decision tree and uses GO annotation information as knowledge score to find out the different results such as enriched GO Terms, top genes that induces lung cancer, mapping of genes with significant GO Terms, and overlapping of GOTerms among genes and presented in this section.

A. Top ranked GO Terms: - The top most GO Terms in the ontologies such as biological process, molecular function and cellular component obtained by random forest algorithm as shown in Figure.2. The top most GO Terms are identified based on average score and average rank.

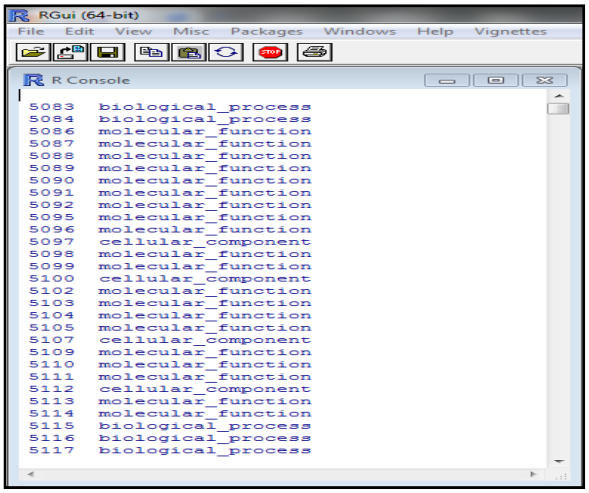

Figure 2: GO terms

B. Top Ranked Genes: -The top most genes from subset of the experimental dataset are extracted and displayed based on score and rank as shown in Figure.3. This figure has score, rank, external_ gene_ name and description of the top most genes.

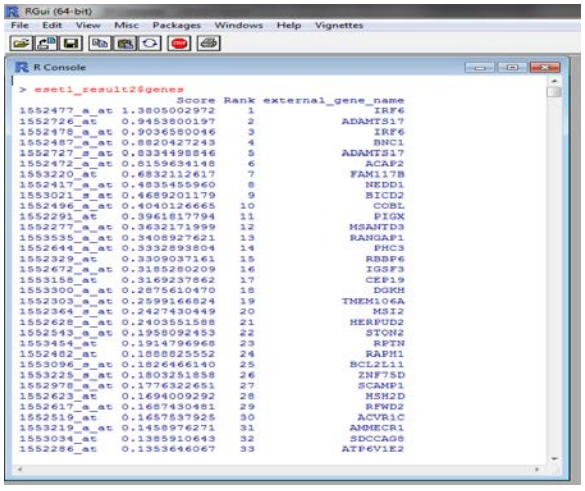

Figure 3: Ranked genes

C. Gene-GO Mapping: - The Gene to Gene Ontology mapping is obtained by mapping the genes with it's corresponding GO terms using GO annotation dataset for homosapiens organism downloaded using BioMart package in R. The Gene_ GO mapping for genes present in the dataset is shown in Figure.4.

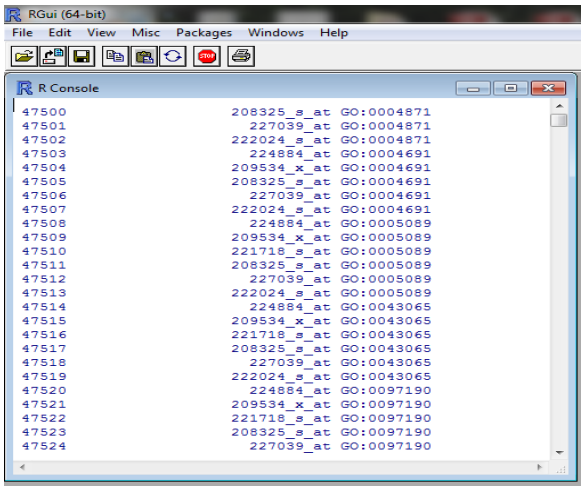

Figure 4: GO mapping

D. Top Ranked Go Terms In Three Ontology Categories:- The overall result produced by random forest algorithm has been classified into three different ontologies namely biological process, molecular function and cellular component. The top 6 ranked biological processes involved in lung cancer dataset is shown in Figure. 5.

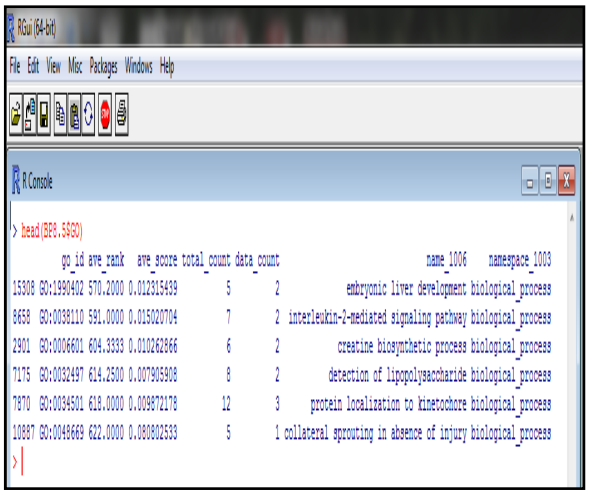

Figure 5: Biological process

Similarly, the top ranked molecular functions are extracted and displayed as shown in Figure 6 and Figure.7 displays the top ranked cellular components where all the biological processes and molecular functions related to lung cancer disease are taking place.

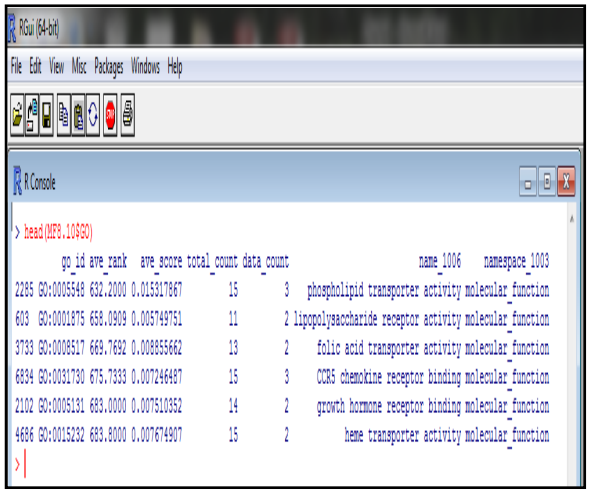

Figure 6: Molecular function

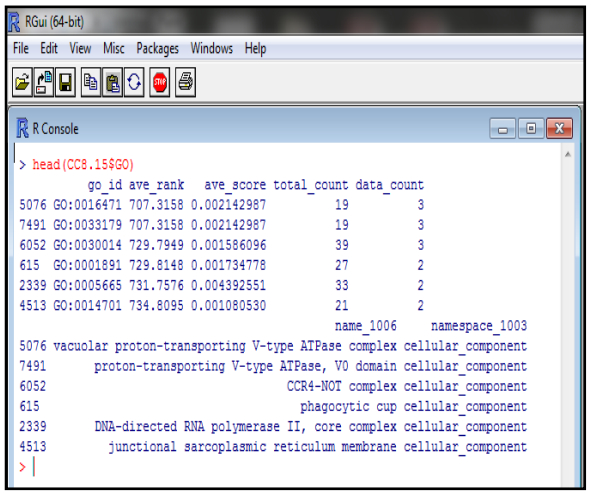

Figure 7: Cellular component

E. Overlapping Genes Between Go Terms:- The overlapping of genes between go terms are identified based on three ontology categories, overlapping of genes using top ranked molecular functions, or top ranked biological processes, or top ranked cellular components. Overlapping of genes between go terms are represented using Venn diagrams. The Figure.8 shows the Venn diagram of genes associated with the five top ranked biological processes and Figure.9 shows the genes associated with five top ranked molecular functions; finally Figure.10 shows five top ranked cellular components.

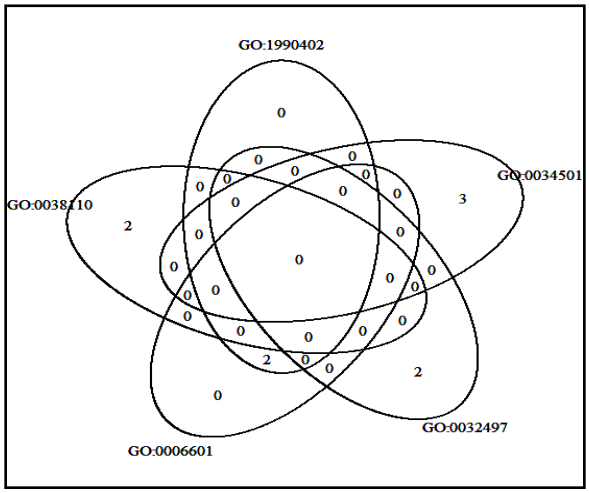

Figure 8: Biological process

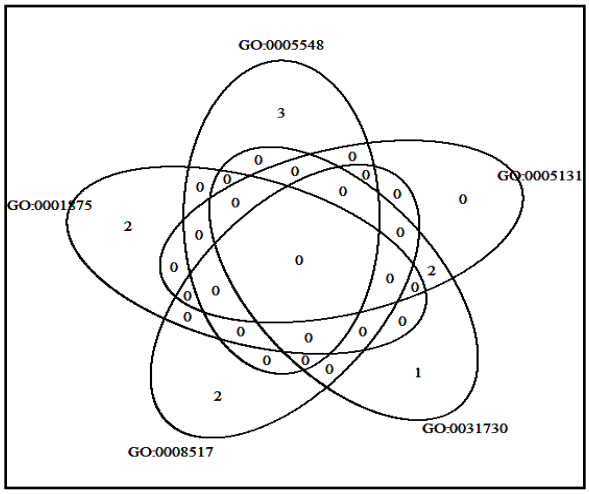

Figure 9: Molecular function

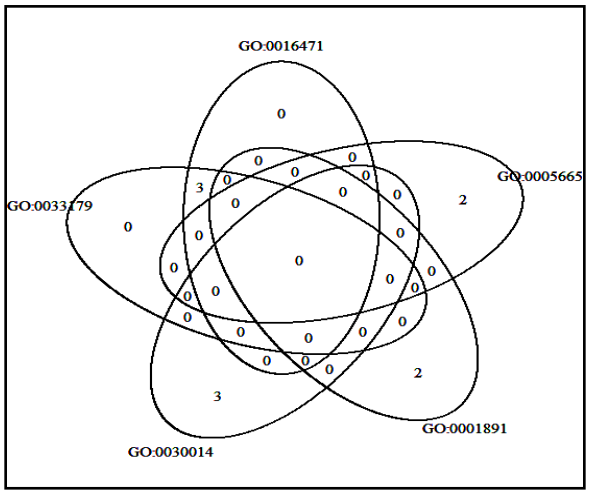

Figure 10: Cellular component

4.2. RESULT OF HIERARCHICAL CLUSTERING:

The hierarchical clustering has been implemented successfully in R using two functions namely heatmap_GO () and cluster_GO (). Both functions produce the set of hierarchical clusters for the experimental dataset. The hierarchical clustering found clusters based on top ranked GO terms. The subset of experimental dataset has been clustered using heatmap_GO () based on the top ranked GO Term '*GO: 0032497*' as shown in Figure.11 ,This figure shows that the samples in the experimental dataset clustered based on three genes namely *TLR4, SCARB1* and *LY96.*

The graphical representation of the resulting hierarchy is a tree-structured graph called a dendrogram. The expression data of the lung cancer samples of genes is associated with top ranking *GO: 0032497* with description, detection of lipopolysaccharide, distance is measured.

The hierarchical clustering of genes based on gene expression profile associated with a top ranked Gene Ontology Terms. The hierarchical clustering of samples using GO Term *GO: 0032497* belong to molecular function with definition detection of lipopolysaccharide visualized using heatmap and dendogram.

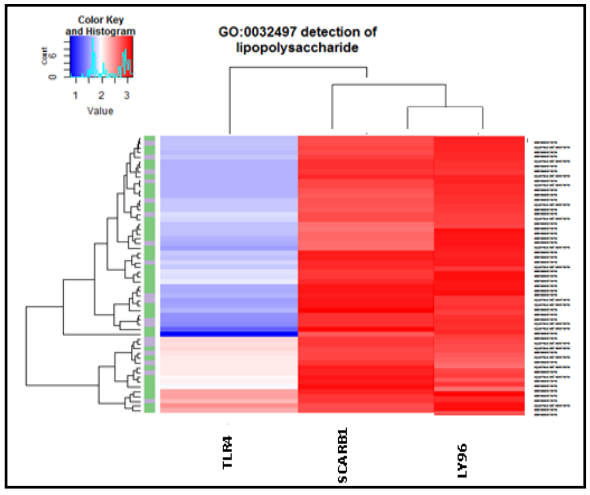

Figure 11: Hierarchical clustering based on biological process 'Detection of Lipopolysaccharide'

The identified clusters are visualized using dendrogram by cluster_GO function in R. The clusters identified by heatmap () function based on top ranked GO term '*GO: 0032497*' are visualized as dendrogram shown in Figure.12. The Figure.13 presents the dendrogram view of cluster based on the top ranked cellular component 'vocuolar proton trsansporting V-Type ATPase complex'. The difference between heatmap and dendrogram view

of clusters is that the first one the cluster is visualized by considering the gene expression profile of genes which are occurs in the same cluster.

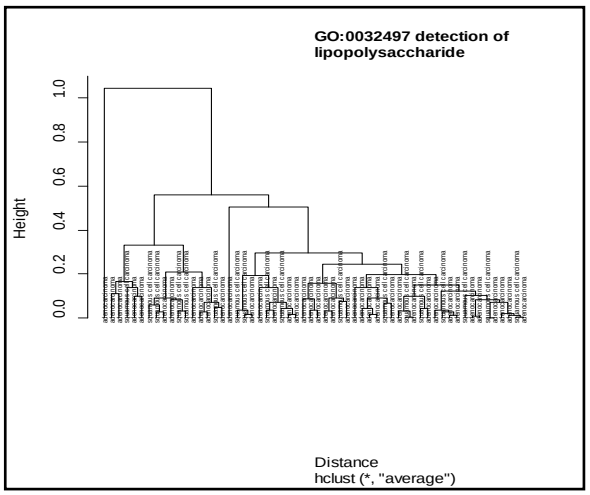

Figure 12: Biological cluster

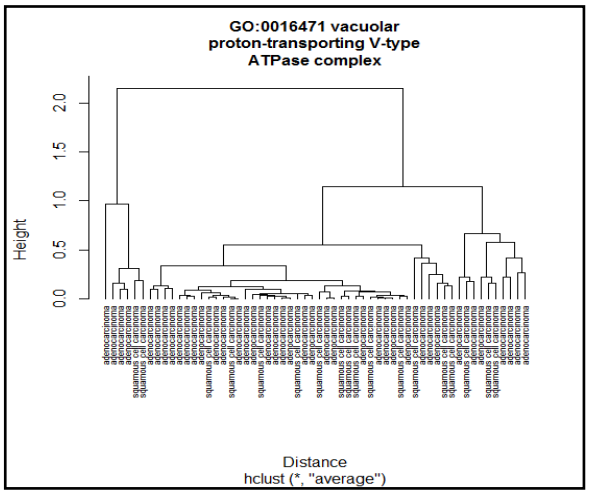

Figure 13: Top ranked cellular

The clusters identified by heatmap () function based on top ranked GO term molecular function '*GO: 0005548*' is visualized as shown in Figure.14

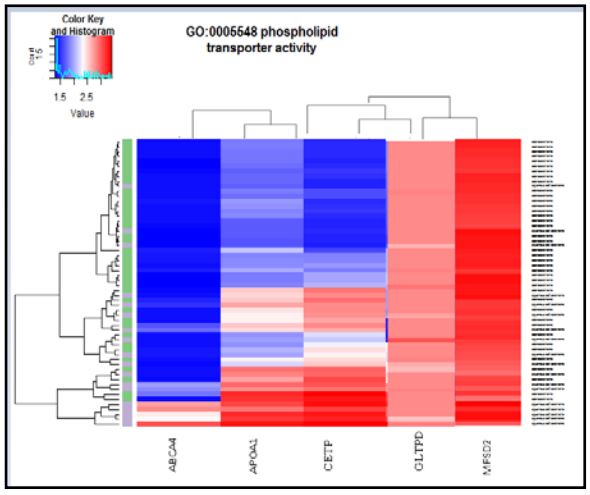

Figure 14: Molecular based cluster_GO

5. CONCLUSION:

The random forest algorithm has successfully computed score and rank for all GO Terms present in each gene given in the dataset based on decision tree. The GO Terms for genes in the experimental dataset were extracted using BioMart packages in R. The random forest algorithm have produced the results such as extracting most significant GO Terms, most significant genes in the experimental dataset, genes with its GO Terms and overlapping of genes based on GO Terms. The result of random forest algorithm has been used to cluster the gene expression and profiles in the experimental dataset successfully. This research work can be extended in future by incorporating other classification algorithms to analyze GO Terms present in the experimental data and the result to be produced may be compared with existing random forest algorithm result.

REFERENCES:

1. Wolfson, O., Sistla, P., Chamberlain, S. and Yesha, Y. "Updating and Querying Databases that Track Mobile Units," Distributed and Parallel Databases 1999.
2. Canasai Kruengkrai , Chuleerat Jaruskulchai, "A Parallel Learning Algorithm for Text Classification," The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002),Canada,July 2002.
3. J. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," artificiall intelligence, 4011-61, 1989.

4. Yu, H. L., Ma, W. L., & Zheng, W. Y. Gene expression database GEO for data mining. Chinese Journal of Biotechnology, 2007.
5. V.S. Tseng, H.H. Yu, Microarray data classification by multi-information based gene scoring integrated with Gene Ontology, Int.J. Data mining .Bioinformatics. Volume 5, issue, 2011.
6. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol, 2006
7. Aniruddha Datta, E. R. D., Ed. Introduction to genomic signal processing with Control, CRC Press, 2007.
8. Cutler, A., Cutler, D.R., Stevens, J.R. In: Zhang, C., Ma, Y. (Eds.), Ensemble Machine Learning: "Methods and Applications. Springer Science Business Media", LLC, 2012.
9. Tierney L, Rossini AJ, Li N, Sevcikova H: SNOW: "Simple Network of Workstations" Tech. rep, 2004.
10. Chin YL: "Effective Gene Selection Techniques for Classification of Gene Expression Data" University of Malaysia, 2005.
11. Díaz-Uriarte R, De Andres SA: "Gene selection and classification of microarray data "using random forest. BMC Bioinformatics, 2006.
12. L. Breiman, Random forests, Machine Learning, 2001.
13. Han-Yu Chuang, Hong fang Liu Brown, S. McMunn- Coffran, C. Cheng-Yan Kao, D. F. Hsu, "Identifying significant genes from microarray data," Fourth IEEE Symposium on Bioinformatics and Bioengineering, 2004.
14. Kevin Rue-Albrecht (2014). GOexpress: Visualise microarray and RNAseq data using gene ontology annotations. R package version 1.2.2.