Abstract

Objective: The main objective of the research paper is to identify gene enrichment analysis of clustered genes from asthma microarray dataset that are clustered using k-means clustering algorithm. The enrichment analysis is used to assign biological meaning to some group of genes and in this paper the gene enrichment analysis is done using Gene Ontology terms. Method: The proposed research work consists of two-fold task: clustering gene expression profile using K-Means clustering algorithm; conducting Gene Enrichment Analysis using Gene Ontology (GO) to identify the enriched GO terms in each cluster with respect to specific set of Molecular Functions, Biological Processes and Cellular Components; Gene Ontology is used to provide external validation for the clusters to determine if the genes in a cluster belong to the specific functionalities. The most significant or enriched GO terms are extracted based on P-Value metric and the most enriched GO terms are visualized for genes in each cluster by using graph. Finding: The asthma microarray dataset contains 41,000 genes and out of that only 9,425 genes are considered for clustering process after preprocessing task. The series of preprocessing tasks certainly helpful to improve and predicts the accurate results. The experimental result of the research work produces the 3 sets of clusters and the first cluster set with 4 clusters have the number of genes 1720, 2636, 2458 and 2611 respectively. Similarly, the other two cluster sets numbers of genes are identified. The gene enrichment analysis is conducted for each cluster in the cluster set based on top most significant molecular functions GO terms enriched in each cluster. The most significant GO terms have been identified based on the P-value metric and associations among top most significant GO terms are visualized in each cluster.

Keywords: Analysis of Enriched Go Terms, Clustering in Bioinformatics, Gene Enrichment Analysis, Identify Enriched GO Terms of Genes, Microarray Gene Expression Data

1. Introduction

Data mining refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. Clustering is one of the tasks in data mining which is partitioning the points into natural groupscalled clusters, such that the points in inter-cluster are very similar, whereas the points in intra-clusters are as dissimilar as possible. In data mining, computer tools must be developed to allow the extraction of meaningful biological information. Biological data mining is a very important part of Bioinformatics.

Genome function annotation including the assignment of a function for a potential gene in the raw sequence is now the hot topic in bioinformatics. Bioinformatics is an interface between modern biology and informatics. Microarray is a group of small DNA spots attached to a solid surface which is in bioinformatics. It contains thousands of DNA spots, covering almost every gene in a genome and it is not possible to perform a meaningful microarray experiment without the bioinformatics involvement at every stage. In bioinformatics, Gene Ontology (GO) is to analyze microarray gene expression data by using the GO Database. GO is one of the most important tools for
representing and processing the gene information such as gene products and gene functions.

GO provides a controlled vocabulary for the description of cellular components, molecular functions and biological processes. GO terms that are associated with the microarray probes. Bioinformatics methods, using the biological knowledge accumulated in public databases as GO make it possible to systematically dissect large gene lists in an attempt to assemble a summary of the most enriched and pertinent biology. Enrichment analysis is to assign biological meaning to some group of genes. The gene-annotation enrichment analysis is a promising high-throughput strategy that increases the likelihood for investigators to identify biological processes most pertinent to the study.

Data mining approaches seem ideally suited for bioinformatics, since bioinformatics is data-rich but lacks a comprehensive theory of life's organization at the molecular level. However, data mining in bioinformatics is hampered by many facets of biological databases, including their size, number, diversity and the lack of a standard ontology to aid the querying of them as well as the heterogeneous data of the quality and provenance information they contain. Another problem is the range of levels of expertise present amongst potential users, so it can be difficult for the database curators to provide access mechanism appropriate to all.

Bioinformatics is the application and the development of data mining techniques to solve biological problems. Traditional genetics and molecular biology have been directed towards understanding the role of a particular gene or protein in an important biological process. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data.

Databases are essential for bioinformatics research and applications, so hundreds of databases in bioinformatics and inconsistent terminology and data formats from a variety of data sources must be efficiently managed.

In authors quoted that the current researchers have access to and can use not only DNA microarray data but also proteome and metabolome data. Thus, we are currently in a data rich environment, and it is difficult to understand these data comprehensively using only human cogitation. It is well known that the bioinformatics techniques described in this review are powerful tools for the analysis of such data in a variety of fields.

Clustering analysis of gene expressions is often the first step in microarray data analysis to discover co-expressed genes under different experimental conditions. GeneExpressionData is generally very huge in size and the ease of use for reusing patterns within this data, genes have to be grouped into clusters on the basis of similar features.

In various clustering methods, such as hierarchical clustering, k-means clustering, self-organized maps (SOMs) and the other methods, have been examined and used to elucidate fundamental and/or characteristic expression patterns. Clustering of genes with related function(s) and/or characteristics has been evident at several different levels.

K-means proposed K-means initializes the cluster means by randomly generating k points in the data space. This is typically done by generating a value uniformly at random within the range for each dimension. Each iteration of K-means consists of two steps: (1) cluster assignment, and (2) centroid update. In K-means, the separating boundary between clusters is linear.

Being a typical partitioning algorithm, the K-means algorithm works well only on data sets having isotropic clusters. The use of the K-means algorithm is often limited to numeric attributes.

In defines one key aspect of calculating GO term enrichment is the choice of a reference-term frequency. It is not clear what the appropriate reference-term frequency should be when calculating enrichment of ontology terms for which a background set is not defined. A natural background set is not available, however, when calculating enrichment using disease ontologies because these ontologies have not been used for manual annotation in a way the Gene Ontology has been used.

In proposed the biological relevance of a cluster can be verified based on the statistically significant Gene Ontology (GO) annotation database available at some websites. The p-value of a statistical significance test is used to find the probability of getting values of a test statistic. Statistical significance is evaluated for the genes in a cluster by computing p-values for each GO category. The clusters are sorted according to the decreasing significance level. The p-values are log-transformed (base 10) for better readability.

In refers, since genes with similar expression profiles may imply similarity among their functions in the biological activities, gene clusters may represent specific biological functions. The most used biological measure is the functional enrichment. A cluster of genes is said to be enriched for a functional category if the proportion
of genes within the cluster known to be in that category exceed the number that could reasonably be expected from random chance.

2. Methodology

The methodology of the research work is to analyze the severe asthma microarray dataset by remove empty spots, impute missing values and filter out the highly expressed genes; then the highly expressed genes are clustered using the partitional K-Means clustering algorithm. After clustering, the constructed clusters have been examined and analyzed to identify grouping of genes with similar functionalities in each cluster based on gene enrichment analysis using Gene Ontology. An experiment may compare gene expression in healthy cells versus diseased cells. Functional profiling can be used to elucidate the underlying cellular mechanisms associated with the diseased condition. This is also called term enrichment or term overrepresentation, as this work is testing whether a GO term is statistically enriched for the given set of genes.

The proposed methodology is used for analyzing microarray data using partitional k-means clustering algorithm and gene enrichment analysis. The proposed methodology consists of five major components as shown in Figure 1. The first component is to load asthma microarray dataset from GEO web browser. The second component focuses on preprocessing of inconsistent data in experimental dataset. The third component computes the clustering of highly expressed gene profiles using K-Means clustering algorithm. The fourth component is to load GO database from GO website. Finally, the fifth component is gene enrichment analysis that map the genes present in the clusters based similar on functionalities using Gene Ontology.

2.1 Preprocessing

The preprocessing phase consists of three processes which are used for extracting needed information from the dataset. Data mining techniques should be able to handle noise in data or incomplete information. The first process is remove empty spots from dataset such as the severe asthma microarray data contains 41,000 genes, with null values and empty genes. The second process is impute missing values such as imputes the continuity of time series and values for similar genes. The third process is remove genes with low expression profiles such as eliminate poorly expressed genes so that genes whose expressions remain constants or low expression profile can be eliminated for further analysis.

2.2 Clustering

The clustering phase uses K-Means clustering algorithm to produce clusters for severe asthma microarray dataset. One most important and useful exploratory analysis on the microarray data is to apply clustering techniques on the patient samples or tissues. Clustering algorithm divides a set of n objects into K partitions depending on some similarity/dissimilarity metric. If the number of data is bigger than the number of cluster, for each data, calculate the distance to all centroid and get the minimum distance. Genes with similar expression patterns or co-expressed genes should be put in a single cluster with similar cellular functions. K-Means is a non-hierarchical clustering algorithm which is a partitioning method. K-Means uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. The research work can control the details of the minimization using several optional input parameters to K-Means, including ones for the initial values of the cluster centroids, and for the maximum number of iterations. K-Means method is the most efficient in terms of execution time.

Each centroid is the component-wise means of the points in that cluster, after centering and normalizing those points to zero mean and unit standard deviation. Replicates which is number of times to repeat the clustering, each with a new set of initial cluster centroid positions. K-Means returns the solution with the lowest value. This paper can supply replicates implicitly by supplying a 3D array as the value for the start parameter. By default, K-Means uses squared Euclidean distances, which each centroid is the mean of the points in that cluster.

Figure 1. Methodology of proposed work.
The distance is calculated using the squared Euclidean distance formula as given below

$$d_{Ec}(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$  \hspace{1cm} (1)$$

In this equation distance measure examines the root of square difference between coordinates of a pair of objects $x$ and $y$.

### 2.2.1 K-Means Algorithm

The basic K-Means algorithm steps are defined below,

**Step 1**: Choose random $k$ points and set as cluster centers.

**Step 2**: Assign each object to the closest centroids clusters.

**Step 3**: When all objects have been assigned, re calculate the positions of the centroids.

**Step 4**: Go back to step 2 unless the centroids are not changing.

The simplest and most commonly algorithm is the K-Means algorithm. This algorithm partitions the data into $K$ clusters ($C_1, C_2, \ldots, C_K$), represented by their centers or means$^{19}$.

The K-Means algorithm has an advantage in comparison to other clustering methods which have non-linear complexity. The algorithm starts with an initial set of cluster centers, chosen at random or according to some heuristic procedure. In each iteration, each instance is assigned to its nearest cluster center according to the Euclidean distance between the two. Then the cluster centers are re-calculated.

Thus, the K-Means clustering algorithm has been successfully implemented and generated three sets of clusters for the dataset such as four clusters, six clusters and eight clusters.

### 2.3 Post-processing

Apart from the important role of preprocessing, together with the correct selection of the data mining technique which will really answer the target questions, there is an important job to be done between getting the results of the data mining techniques and using them to support decision-making which is to understand the results. Data post-processing is essential to visualize the extracted knowledge in such a way that user can interpret the knowledge easily. Generally, data post-processing includes knowledge filtering, evaluation, information visualization and knowledge integration techniques. In this phase, the generated clusters by K-Means clustering algorithm are analyzed to identify the enriched GO terms present in genes in each cluster.

#### 2.3.1 Mapping of Microarray Data with GO

In human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. Many Genome Projects interact with the Gene Ontology Consortium when annotating genes. Gene annotations for several organisms can be found at the Gene Ontology Website. These annotations are updated frequently and are usually curated by members of the genome group for each organism. NCBI also has a collective list of gene annotations that relate to their Entrez Gene database. These annotation files consist of large lists of genes and their associated Gene Ontology terms. These files follow the structure defined by the Gene Ontology Consortium. The file 'gene_association.goa_human' was obtained from the Gene Ontology Annotation site and used to get mapping of asthma microarray dataset with GO in the proposed research.

The clustered genes are involved with certain molecular function ('F'), biological process ('P') and are co-located in the same cellular component ('C') are analyzed. Human Genome Database (HGD) annotation is the master structure of the human genes containing all the information about the human data with human related genes shown in Table 1.

#### 2.3.1.1 Finding Biological Association Among Genes

Understanding the underlying molecular mechanisms in human diseases is important for diagnosis and treatment of complex conditions and has traditionally been

<table>
<thead>
<tr>
<th>Molecular Function</th>
<th>Unique GO terms associated to annotated genes</th>
<th>Gene-GO term association</th>
</tr>
</thead>
<tbody>
<tr>
<td>17,269</td>
<td>3,816</td>
<td>1,55,613</td>
</tr>
<tr>
<td>Cellular Component</td>
<td>18,313</td>
<td>1,440</td>
</tr>
<tr>
<td>Biological Process</td>
<td>17,202</td>
<td>10,792</td>
</tr>
</tbody>
</table>

Table 1. Annotated Genes using human genome database
done by establishing associations between disorder-genes and their associated diseases. This kind of analysis usually includes only the interaction of molecular components and shared genes.

2.3.1.2 Finding Probability of GO

In this process, find the most significant annotated terms by looking at the probabilities that the terms are counted by chance. For this process, use the hypergeometric probability distribution method. This method returns the p-value associated to each term, this paper can create a list of the most significant GO terms by ordering the p-values.

The functional enrichment of a group of genes is measured in terms of three structured, controlled vocabularies or ontologies, viz., associated biological processes, molecular functions and biological components. In other words, a hypergeometric test is applied to assess the significance of the overrepresentation of predefined gene sets within a list of differing genes derived from a certain experiment based on fold-change and/or p-value cutoffs.

The degree of functional enrichment for a given cluster \( n \) and functional category \( f \) can be quantitatively assessed by the hypergeometric data distribution \( P \).

\[
p = 1 - \sum_{i=1}^{k-1} \frac{\binom{f}{i} \binom{g-f}{n-i}}{\binom{g}{n}}
\]

where \( g \) is the size of genome, \( f \) is the number of genes of the genome in the considered category, \( n \) is the cluster size and \( k \) is the number of genes in the cluster which are in the same category. If this probability is sufficiently low, then the cluster is said to be enriched for that category. Usually the biological enrichment is evaluated according to the Gene Ontology categories.

2.3.1.3 Analyze Most Significant GO Terms

In the proposed research work, now build a sub-ontology that includes the ancestors of some most significant terms and visualize this using the graph. After that use the p-values, calculated before, to assign a color to the graph nodes. For this process, use the methods such as ancestors, descendants, relatives, and matrix which are to find out more about the terms that appear high on this list. Note that the terms of found ancestor appears as one of the ancestors. If first looks at the list for the next item, then will see many of the same ancestors.

3. Experimental Results

The proposed methodology has been successfully implemented and the result obtained in every phase is discussed below sections.

3.1 Dataset

The Severe Asthma: Bronchial Epithelial Cells Dataset is downloaded from National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) website and used for this research work. The Severe Asthma microarray data contains about 41,000 genes with their corresponding Severe Asthma samples about 108. The Human Genome Database is a scientific database of the molecular biology and genetics of the Homosapiens. This database includes a variety of genomic and biological information. The file ‘gene_association.goa_human’ was obtained from GO annotation site. The ‘gene_association.goa_human’ file contains the parameters namely HGDonnotation, HGDaspect, HGDgenes and HGDgo, which is used to map genes in severe asthma microarray dataset with Gene ontology in order to identify the association prevailed among genes in each cluster.

3.2 Result of Preprocessing

The asthma microarray dataset contains 41,000 genes and after applying preprocessing by removing empty spots and null values, it produces 30,935 genes out of 41,000 genes. Then, remove the low expression genes based on threshold value in that 9,425 genes are extracted out of 30,935 genes. The result of preprocessing is tabulated in Table 2.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>List of preprocessing genes in dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processing</td>
<td>Before Preprocessing</td>
</tr>
<tr>
<td>Removed empty spots and null values</td>
<td>41,000</td>
</tr>
<tr>
<td>Impute missing values</td>
<td>30,935</td>
</tr>
<tr>
<td>Remove low expression Genes</td>
<td>30,935</td>
</tr>
</tbody>
</table>
3.3 Result of K-Means Clustering Algorithm

In this work, K-Means clustering algorithm is applied for the asthma microarray dataset and the similarity measure is calculated using squared Euclidean distance, the output of k-means algorithm consists of clusters of similar objects. There are 3 set of clusters are generated using K-Means algorithm namely four clusters, six clusters and eight clusters. The total number of genes in each cluster of four clusters set is tabulated in Table 3 and the same is shown in Figure 2 as bar chart.

The Figure 3 (a), (b) and 4 (a), (b) shows the clustering and clustering centroids for four clusters set by applying K-Means algorithm for the dataset before and after applying preprocessing steps. When considering Figure 3 (a) and 3 (b), 4 (a) and 4 (b) contains more number of dense clusters and more centroids, because the K-Means clustering algorithm applied for the whole dataset.

The total number of genes in each cluster of six cluster set is tabulated in Table 4 and as same is shown in Figure 5 as bar chart.

The Figures 6(a), (b) and 7 (a), (b) shows the clustering and clustering centroids for six clusters set by

Table 3. List of genes in four clusters

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Number of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster – I</td>
<td>1720</td>
</tr>
<tr>
<td>Cluster – II</td>
<td>2636</td>
</tr>
<tr>
<td>Cluster – III</td>
<td>2458</td>
</tr>
<tr>
<td>Cluster – IV</td>
<td>2611</td>
</tr>
<tr>
<td>Total number of genes</td>
<td>9425</td>
</tr>
</tbody>
</table>

Figure 2. Bar chart for four clusters set.

Figure 3. (a) Clustering before preprocessing using four clusters. (b) Clustering centroids before preprocessing using four clusters.

Figure 4. (a) Clustering after preprocessing using four clusters. (b) Clustering centroids after preprocessing using four clusters.
Table 4. List of genes in six clusters

<table>
<thead>
<tr>
<th>Six Clusters</th>
<th>Number of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster – I</td>
<td>1387</td>
</tr>
<tr>
<td>Cluster – II</td>
<td>1963</td>
</tr>
<tr>
<td>Cluster – III</td>
<td>2097</td>
</tr>
<tr>
<td>Cluster – IV</td>
<td>1712</td>
</tr>
<tr>
<td>Cluster – V</td>
<td>1045</td>
</tr>
<tr>
<td>Cluster – VI</td>
<td>1221</td>
</tr>
<tr>
<td><strong>Total number of genes</strong></td>
<td><strong>9425</strong></td>
</tr>
</tbody>
</table>

Figure 5. Bar Chart for six clusters set.

Figure 6. (a) Clustering before preprocessing using six clusters. (b) Clustering centroids before preprocessing using six clusters.

Figure 7. (a) Clustering after preprocessing using six clusters. (b) Clustering centroids after preprocessing using six clusters.

applying K-Means algorithm for the dataset before and after applying preprocessing steps. When considering Figures 6 (a) and 6 (b), 7 (a) and 7 (b) contains more number of dense clusters and more centroids, because the K-Means clustering algorithm applied for the whole dataset.

The total number of genes in each cluster of eight cluster set is tabulated in Table 5 and as same is shown in Figure 8 bar chart.

The Figures 9 (a), (b) and 10 (a), (b) shows the clustering and clustering centroids for eight clusters set by applying K-Means algorithm for the dataset before and after applying preprocessing steps. When considering Figures 9 (a) and 9 (b), 10 (a) and 10 (b) contains more number of dense clusters and more centroids, because the K-Means clustering algorithm applied for the whole dataset.
Table 5. List of genes in eight clusters

<table>
<thead>
<tr>
<th>Eight Clusters</th>
<th>Number of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster – I</td>
<td>1410</td>
</tr>
<tr>
<td>Cluster – II</td>
<td>933</td>
</tr>
<tr>
<td>Cluster – III</td>
<td>1434</td>
</tr>
<tr>
<td>Cluster – IV</td>
<td>1329</td>
</tr>
<tr>
<td>Cluster – V</td>
<td>1367</td>
</tr>
<tr>
<td>Cluster – VI</td>
<td>834</td>
</tr>
<tr>
<td>Cluster – VII</td>
<td>1111</td>
</tr>
<tr>
<td>Cluster – VIII</td>
<td>1007</td>
</tr>
<tr>
<td><strong>Total number of genes</strong></td>
<td><strong>9425</strong></td>
</tr>
</tbody>
</table>

Figure 8. Bar chart for eight clusters set.

Figure 9. (a) Clustering before preprocessing using eight clusters. (b) Clustering centroids before preprocessing using eight clusters.

3.4 Result of Gene Enrichment Analysis

3.4.1 HGDannotation Structure

The gene_association.goa_human file was obtained from the GO annotation site. This file contains the parameters namely HGDannotation, HGDaspect, HGDgenes and HGDgo. HGDannotation is a master structure of human genes containing all the information about the human data with 982698 human related genes. The HGDaspect contains one of the three GO ontology division namely Biological Process (P), Cellular Component (C) and Molecular Function (F). The HGDgenes parameter contains all the HGDaspect corresponding genes. The HGDgo contains all the HGDgenes with corresponding GOid.

3.4.2 Mapped Genes based on GO Functions

The expressed genes contain different functionality. The HGDaspect column is retrieved from the master (HGDann) structure. The mapping is done based on the aspect value (Biological Process, Cellular Component.
and Molecular Function). Where ‘P’ denotes biological process, ‘C’ denotes cellular component and ‘F’ denotes molecular function. After mapping, it is identified that 189942 GO terms as biological processes, 145794 GO terms as cellular components and 155613 GO terms as molecular functions are associated with 9425 genes. Each cluster genes in three set of clusters is enriched GO terms identified, based on these three ontologies.

3.4.3 Identifying Biological Association among Genes based on P-Value

After identifying GO terms for each cluster in three sets of cluster, the most significant GO terms which are identified based on P-value to each cluster in three sets of clusters. Most significant GO terms are most essential GO terms which are actively participating in gene activity. The ten top most significant GO terms are identified for four clusters set, six clusters set and eight clusters set with respect to the three ontologies. The Figure 11 shows the ten top most significant GO terms, based on 1720 genes in first cluster of four clusters set with respect to molecular function category.

The Figure 12 shows the ten top most significant GO terms, based on 2636 genes in second cluster of four clusters set with respect to molecular function category.

The Figure 13 shows the ten top most significant GO terms, based on 2458 genes in third cluster of four clusters set with respect to molecular function category.

The Figure 14 shows the ten top most significant GO terms, based on 2611 genes in fourth cluster of four clusters set with respect to molecular function category.

Now build a sub-ontology that includes the ancestors of the ten most significant terms and visualize these relationships using the biograph function and the getmatrix method. For this use the P-Values which calculated before to assign a color to the graph nodes. In this work, an arbitrary color map is used, where bright red is the most significant and bright green is the least significant which shows in Figures 15, 16, 17 and 18 respectively.

Hence, the K-Means clustering algorithm has been implemented for severe asthma microarray dataset to group genes with similar functionalities using gene enrichment analysis based on GO.
4. Conclusion and Future Enhancement

The proposed work is based on clustering microarray dataset to group genes with similar functionalities based on enriched Gene Ontology (GO) terms appears in genes. The microarray severe asthma dataset was tested using K-Means clustering algorithm and found three sets of clusters such as four clusters set, six clusters set and eight clusters set. The significant GO terms to each cluster in three sets of clusters are identified using the Gene Enrichment Analysis based on Gene Ontology in which GO terms of genes are managed as molecular function, biological process and cellular component. The most enriched GO terms of genes in the experimental dataset are extracted based on P-Value and finally the enriched GO terms are built with sub-ontology in which the associations among these GO terms are analyzed.

This research work can be extended for future aspects are any other hierarchical clustering algorithms can be implemented and compared and this research work uses Euclidean distance similarity metric as a measure for clustering. This may be changed with some other distance metrics.

5. References


