# A Hybrid Clustering Algorithm and Functional Study of Gene Expression in Lung Adenocarcinoma

**Mohammad Shabbir Hasan and Zhong-Hui Duan**
Department of Computer Science, University of Akron, Akron, Ohio, USA

**Abstract -** *DNA Microarray technology provides a convenient way to investigate expression levels of thousands of genes in a collection of related samples during different biological processes. Researchers from diverse disciplines such as computer science and biology have found it interesting as well as meaningful to group genes based on the similarity of their expression patterns. Hierarchical clustering and k-means clustering are commonly used algorithms to group genes with similar expression patterns. However, in spite of having some advantages such as producing tighter cluster than other algorithms, k-means clustering has some limitations also. The performance of k-means clustering algorithm largely depends on the selection of the value of k i.e., the number of clusters. In this research work, we proposed a new method to combine k-means clustering with hierarchical clustering to overcome the limitation. To test the algorithm, we used microarray data on lung adenocarcinoma, the most common type of non-small-cell cancers. We identified a number of representative genes from the group of normal tissue and from the group of KRAS mutation tissues. Genes for both of these groups were clustered using our proposed method. Finally we conducted functional investigation of the differentially expressed genes using Gene Ontology database to find changes in the enrichment of molecular functions of the genes contained in each cluster of both normal and KRAS positive groups. We discovered that our proposed method can group genes with similar expression pattern together and hence it can be used in future for clustering microarray data*.

**Keywords:** Gene Expression; Microarrays; K-means Clustering; Hierarchical Clustering; Gene Ontology

## 1 Introduction

In gene expression, gene products such as proteins or RNA are created from the inheritable information contained in a gene [1]. So far traditional molecular biology has focused on studying individual genes in isolation for determining gene functions. However it is not suitable for determining complex gene interactions as well as explaining the nature of complex biological processes. For this purpose, examining the expression pattern of a large number of genes in parallel is required [2]. DNA microarray technology which is one of the most important tools now-a-days for the analysis of gene expression patterns has made it possible to view thousands of genes expression levels in parallel [3]. This analysis is very useful to get information for diagnosis of different diseases and efficient algorithms are required to analyze DNA microarray datasets accurately. It is believed that a group of genes with similar gene expressions are likely to have related gene functions [4]. Hence identifying genes with similar expression patterns in different phases of the cell cycle or in different environmental conditions is an important task.

Clustering algorithms play an important role in gene analysis by separating a dataset of heterogeneous genes into homogeneous groups containing similar genes. It helps to analyze a group of genes instead of analyzing each one individually. After getting appropriate clusters, researchers can further investigated the clusters to find distinct pattern for each cluster as well as more information about functional similarities and gene interactions. A good number of algorithms have been developed for clustering DNA microarray data so far. These algorithms include k-means clustering [5], hierarchical clustering [6-8], self-organizing maps [9-11], support vector machines [12], Bayesian networks [13] and fuzzy logic approach [14]. Beside these algorithms, some algorithms use other genomic information along with gene expression data in order to improve clustering efficiency. Examples of such algorithms include [15] that use gene ontology data with gene expression data and [16-18] that clusters genes by using information of upstream regions of the coding sequences with gene expression profiles to get more biologically relevant clusters.

K-means clustering algorithm is computationally faster than hierarchical clustering and produces tighter clusters than hierarchical clustering. On the other hand, hierarchical clustering algorithm does not require the number of clusters to be known in advance and computes a complete hierarchy of clusters. Beside these advantages, however, both of these algorithms suffer from some limitations. The performance of *k*-means clustering depends on how effectively the initial number of clusters i.e. the value of *k* is determined. Moreover, these algorithms are computationally expensive which impede the wide use of these algorithms in gene expression data analysis [19-21]. To overcome these limitations, a combined hierarchical *k*-means clustering method has been proposed in [22] which firstly applies *k*-means algorithm in each cluster to determine *k* clusters and then feed those clusters to hierarchical clustering technique to shorten merging clusters time while generating a tree-like dendrogram. But still this algorithm suffers from the limitation of determining the initial value for *k*.

In this paper we present a new algorithm that combines both hierarchical clustering and *k*-means clustering. The goal is to take the advantages of both algorithms to overcome the limitations of *k*-means clustering algorithm. We use the result of hierarchical clustering to decide the initial number of clusters and then feed this information to *k*-means clustering to obtain the final clusters. In microarray data analysis, clustering genes to find out the biologically relevant groups based on their expression profiles is one of the basic techniques. Similarity in gene expression profiles indicates similarity in their gene functionalities also [23]. After getting the new clusters, we explore the change in enrichment of molecular functionalities of the genes of each cluster for normal tissue and adenocarcinoma lung cancer tissue by using Gene Ontology (GO) annotations.

## 2 Materials and Methods

Lung adenocarcinoma is the most frequent type of non-small-cell lung cancers (NSCLC) and it accounts for more than 50% of NSCLC and the percentage is increasing [24]. Recent studies have shown that activation of the EGFR, KRAS and ALK genes defines 3 different pathways which are responsible for a considerable fraction (30%–60%) of lung adenocarcinoma [25-29]. The dataset used in this research contains expression profiles for 246 samples where 20 samples belong to normal lung tissue. Out of 226 lung adenocarcinomas samples, 127 are with EGFR mutation, 20 with KRAS mutation, 11 with EML4-ALK fusion and 68 samples are with triple negative cases. Platform used for this dataset is GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. This dataset was collected from GEO database (accession number GSE31210). The dataset contains 54675 genes and out of the 246 samples, for this research work, we considered 40 samples that consist of 20 samples from normal tissue and 20 samples from KRAS positive tissues.

To determine the differentially expressed genes, we performed paired Student *t*-test and Bonferroni correction followed by the calculation of the value of fold change of the genes. In this study, after performing Bonferroni correction, we selected the genes as the most differentially expressed which have adjusted *p*-values $\leq 0.05$. In addition to that, we considered only those genes where the value of fold change (increase or decrease) is significant i.e. the average fold change between cancer and normal is greater than or equal to 2. Beside these preprocessing, we considered only those genes that are associated with molecular functions according to the Gene Ontology (GO). Figure 1 shows the flow diagram of the data preprocessing. After performing *t*-test, we obtained 21,880 genes which had significant *p*-value ($\leq 0.05$). We performed Bonferroni correction on these genes and found 1,988 genes which had a significant adjusted *p*-value ($\leq 0.05$). Adding the fold change criterion, we reduced the set of differentially expressed genes to 1,005. We then performed another step of filtering to keep only those genes that have Gene Ontology (GO) terms and responsible for molecular functions. Finally we came up with 464 genes in the dataset. The final dataset which is also termed as filtered dataset in this paper is given partially in Table 1 and the complete dataset is available in [30].

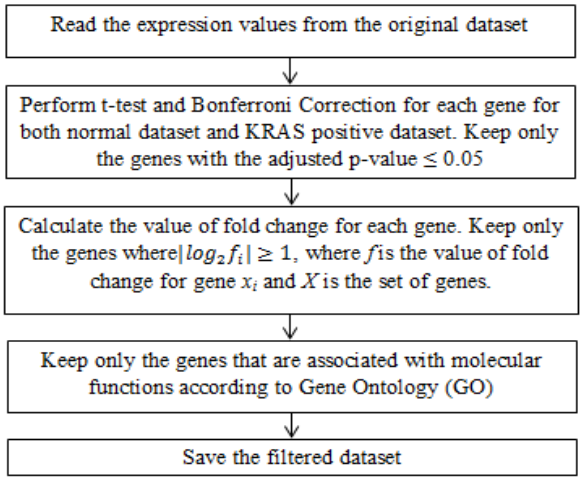


Figure 1: Flow diagram of data preprocessing

To overcome the limitation of k-means clustering algorithm, in our proposed method, we selected the value of *k* i.e. the number of clusters in a systematic way. Initially we use the hierarchical clustering algorithm for clustering the dataset and then check at which level the distance between two consecutive nodes of the hierarchy is the maximum and from this result the value of *k* is determined which is then used as the value of *k* for the *k*-means clustering. In both algorithms, Pearson correlation coefficient (*r*) is used as the similarity metric between two samples and 1-*r* is used as the distance metric.

Table 1: A brief overview of the final dataset

| Affymatrix ID | Gene Symbol | Samples | | |
|---|---|---|---|---|
| | | GSM 773551 | … | GSM 773784 |
| 1555579_s_at | PTPRM | 3441.22 | … | 3569.13 |
| 211986_at | AHNAK | 4395.68 | … | 7080.40 |
| 222392_x_at | PERP | 21707.73 | … | 11350.53 |
| 236715_x_at | UACA | 1303.01 | … | 1867.76 |
| 244704_at | NFYB | 124.08 | … | 277.49 |
| … | … | … | … | … |
| 211237_s_at | FGFR4 | 22.41 | … | 11.07 |
| 203980_at | FABP4 | 257.25 | … | 920.44 |
| 207302_at | SGCG | 47.09 | … | 9.61 |
| 210081_at | AGER | 241.63 | … | 2001.28 |
| 217046_s_at | AGER | 132.42 | … | 1016.05 |

# 3 Results and Discussions

Figure 2 shows the hierarchical clustering of normal tissue dataset. There are 463 interior nodes in the tree where each node is labeled based on the increasing order of its height. Therefore the root has its ID 463. To determine the number of clusters from the output of hierarchical clustering, we used a bar graph to show the difference of height between two consecutive interior nodes and it is shown in Figure 3.

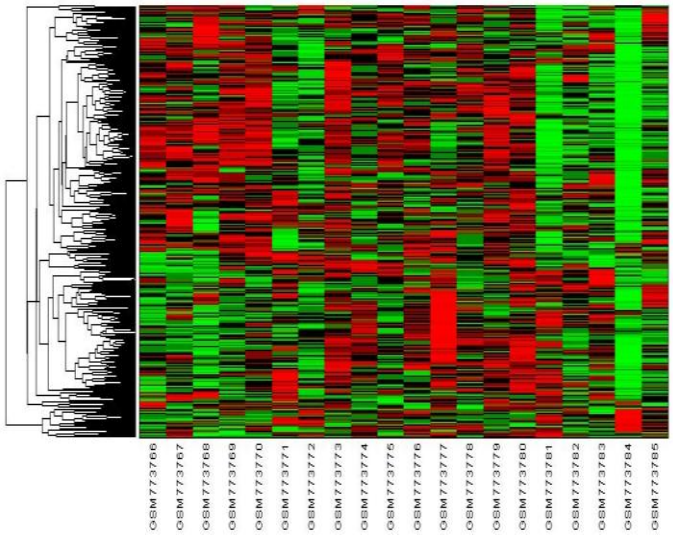


Figure 2: Hierarchical clustering of normal tissue dataset

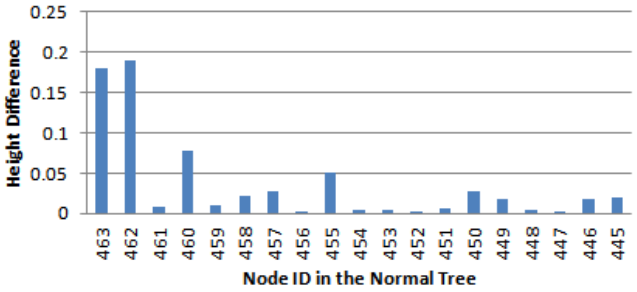


Figure 3: Height difference between two consecutive interior nodes in the hierarchical tree generated from the normal tissue. Since Pearson distance is used, the maximum height of the tree is 1.

From Figure 3 we can see that the difference is the maximum for node 461 and node 462. As there are total 463 nodes in the tree, node 461 is in level 3 from the top. So according to the approach we are discussing here, the total number of clusters for k-means clustering should be 4.

Similarly we can determine the number of clusters for the KRAS positive dataset. Figure 4 shows the hierarchical clustering of KRAS positive dataset and the height difference between two consecutive nodes is shown in Figure 5. The results indicate the number of clusters for KRAS positive dataset should be 4.

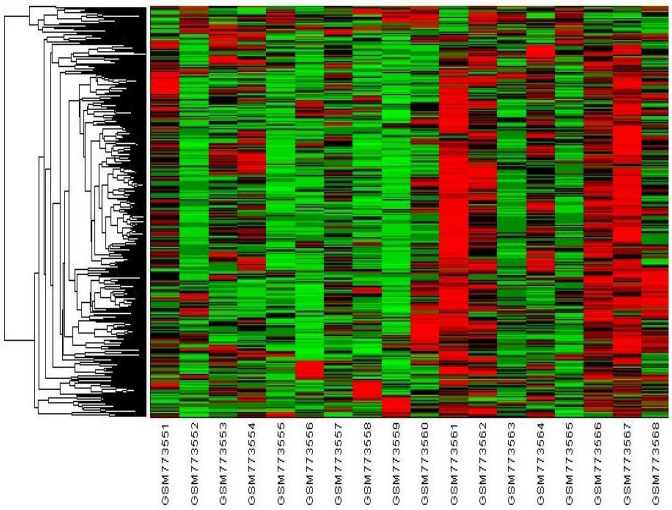


Figure 4: Hierarchical Clustering of KRAS positive dataset

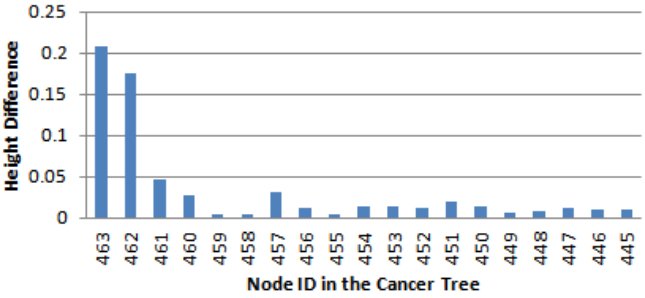


Figure 5: Height difference between two consecutive nodes in the hierarchical tree generated from KRAS positive dataset.

Clearly we see different clusters formed from normal tissue and cancer tissue. We explore their common features (genes) and explain the change of molecular function of the genes captured in the clusters of both normal tissue and KRAS positive datasets using Gene Ontology (GO) annotations. For comparing the molecular function of the clusters of normal tissue and KRAS positive tissues, we took one cluster from normal tissue dataset and one from KRAS positive dataset which have maximum number of common genes. Table 2 shows the clusters we have selected for comparing their molecular functions with the number of genes they have in common.

Table 2: List of the clusters to be compared for the alteration in molecular function

| Clusters to compare | | Number of genes in common |
|---|---|---|
| Normal Tissue | KRAS positive | |
| Cluster 1 | Cluster 1 | 20 |
| Cluster 2 | Cluster 3 | 52 |
| Cluster 3 | Cluster 4 | 46 |
| Cluster 4 | Cluster 2 | 69 |

We explain the molecular functions of the genes in each cluster using GO annotations and their relationship are represented using a directed acyclic graph (DAG) which is also termed as GO graph in this paper. To generate these graph, we used a web based tool Gene Ontology Enrichment Analysis Software Toolkit (GOEAST) [31]. This graph displays enriched Gene Ontology IDs (GOIDs) and their hierarchical relationships in Molecular Function GO categories. Here boxes represent GO terms, labeled by its GOID and term definition. Note that significantly enriched GO terms are marked yellow. The degree of color saturation of each node is positively correlated with the significance of enrichment of the corresponding GO term. Non-significant GO terms within the hierarchical tree are shown as white boxes. In this graph, edges stand for connections between different GO terms. Edges with red color stand for relationship between two enriched GO terms, black solid edges stand for relationship between enriched and un-enriched terms; black dashed edges stand for relationship between two un-enriched GO terms.

Figure 6 and 7 shows the GO graph for the cluster 1 of normal tissue dataset and cluster 1 of KRAS positive dataset respectively.

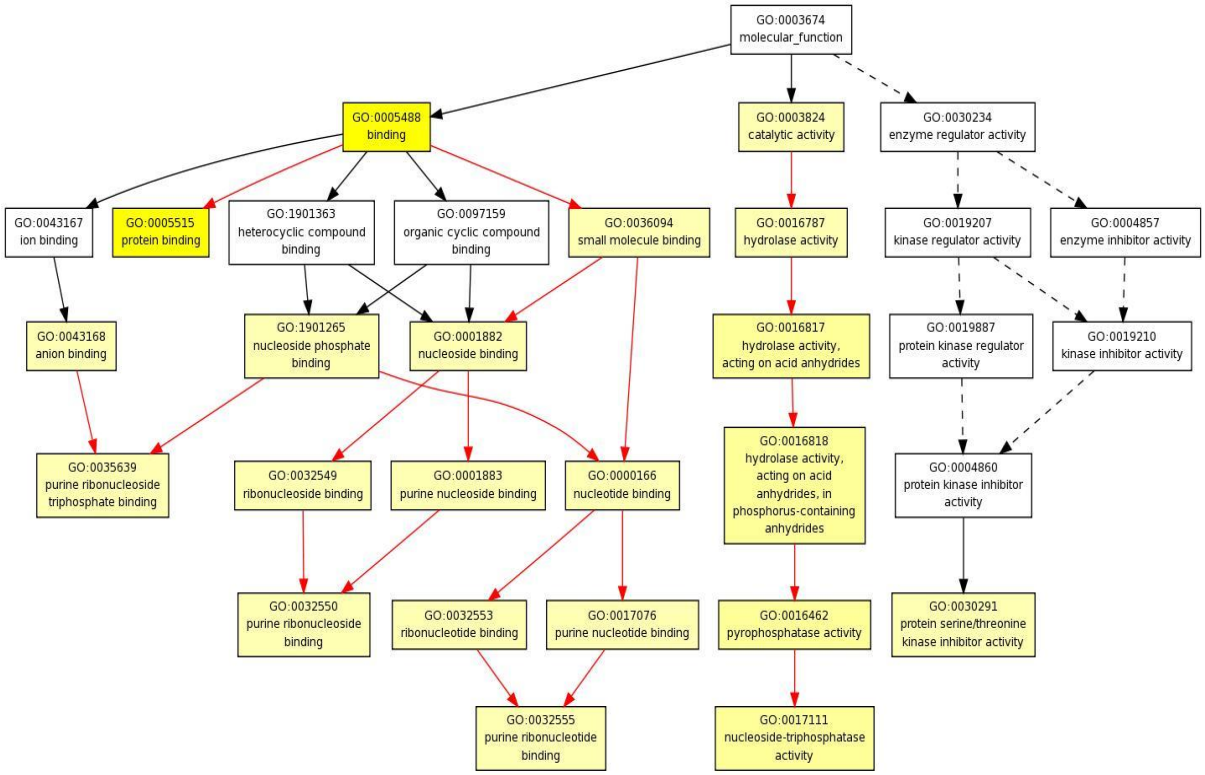


Figure 6: GO graph for cluster 1 of normal tissue data set.

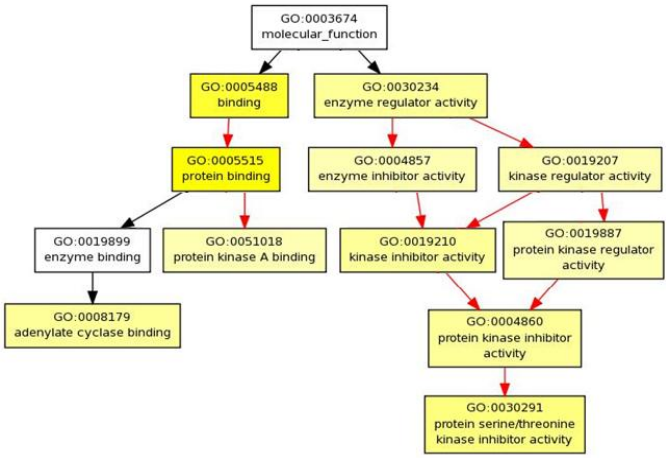


Figure 7: GO graph for cluster 1 of KRAS positive data set

In brief, from these two figures we see that, the significant GO terms GO: 0005488 (binding) and GO: 0005515 (protein binding) remain same in both clusters. GO terms such as GO: 0030234 (Enzyme Regulator Activity), GO: 0019207 (Kinase Regulator Activity), GO: 0019210 (Kinase Inhibitor Activity), GO: 0019887 (Protein Kinase regulator Activity) and GO: 0004860 (Protein Kinase Inhibitor Activity) which are un-enriched in normal tissue, become highly enriched in the KRAS positive tissues.

For better comparing the enrichment status of the two clusters, we used Multi-GOEAST which is an advanced version of GOEAST and it is helpful to identify the hidden correlation between the two clusters [31]. Figure 8 shows the comparative GO graph of the clusters discussed above.

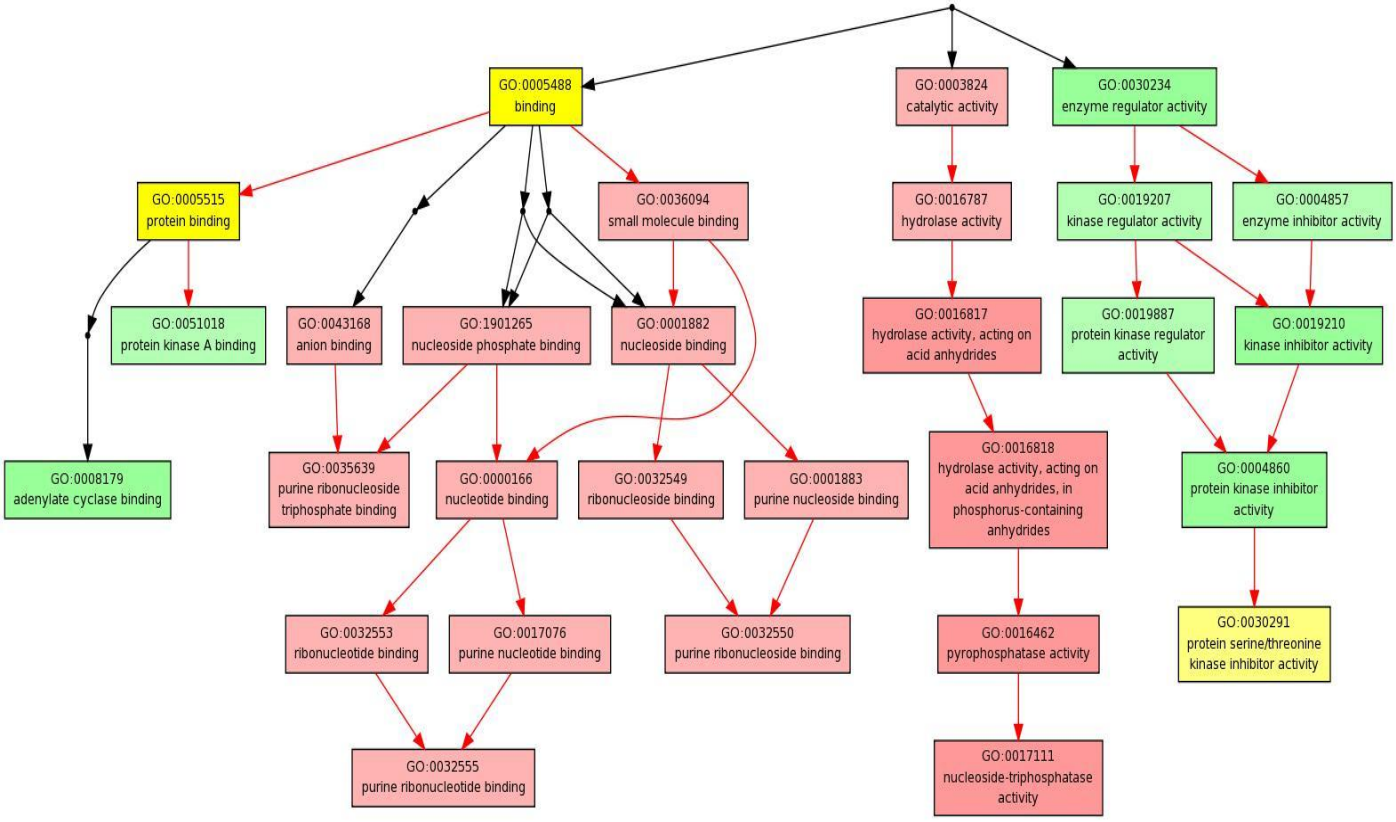


Figure 8: Comparative GO graph for comparing GO enrichment status of Cluster 1 of normal tissue dataset and Cluster 1 of KRAS positive dataset.

In the comparative GO graph, significantly enriched GO terms in both clusters are marked yellow, light yellow color indicates the GO terms which are enriched in both clusters. Nodes marked with coral pink indicate the GO terms which are enriched in normal tissue dataset but not in KRAS positive dataset. In addition to that, nodes with green color represent the GO terms which are un-enriched in normal tissue but enriched in KRAS positive tissues. Note that, the degree of color saturation of each node is positively correlated with the significance of enrichment of the corresponding GO term.

Table 3 lists the genes associated with the GO terms which are enriched in the cluster 1 of KRAS positive tissue dataset but not enriched in the cluster 1 of normal tissue dataset and these GO terms which are marked with green color in the comparative GO graph shown in Figure 8. We believe these are responsible for the alteration of the molecular activity in the cell and are linked to the development of KRAS lung cancer. Similarly we can generate and compare the GO enrichment graph for the rest of the clusters.

# 4 Conclusions

In this paper we proposed a combined clustering algorithm to cluster genes in a microarray dataset based on their expression levels. In the algorithm the number of clusters, i.e. the value of $k$ which is required for $k$-means clustering algorithm is determined from the output of hierarchical clustering. Using this systematic way of determining the value of $k$, this approach overcomes the limitation of $k$-means clustering. This proposed method of clustering takes the advantage of hierarchical clustering to get a complete hierarchy of clusters and uses this information to determine the number of clusters to be used in $k$-means clustering for producing tighter cluster.

In this study we examined 40 samples and 464 genes from the dataset of KRAS lung denocarcinoma which is one of the most frequent types of non-small-cell lung cancers. Out of the 40 samples, 20 were from normal tissue and 20 were from KRAS positive tissues. We applied $t$-test, Bonferroni correction and fold change cutoff to find the significantly differentially expressed genes and among them only the genes having GO terms and responsible for molecular functions were included in the final dataset.

After applying the clustering algorithms, we obtained 4 clusters for both normal tissue dataset and KRAS positive dataset. Hereafter, we examined the genes contained in each cluster with respect to their molecular functions based on Gene Ontology (GO) annotation to see what are the changes

in the enrichment of the molecular functions of the genes took place from normal tissues to KRAS positive tissues.

In summary, we presented a coherent approach to examine alterations of molecular activities in different environmental settings such as in cancer cells. Furthermore, the proposed clustering algorithm can be generalized for clustering other types of large datasets.

Table 3: GO Terms and pathways which are enriched in molecular functions of the genes of Cluster1 of KRAS positive tissue but un-enriched in the genes of Cluster1 of normal tissue dataset

| GO ID | GO Term | Associated Genes | Pathway |
|---|---|---|---|
| GO:0030234 | Enzyme Regulator Activity | TIMP3 | Matrix_Metalloproteinases |
| | | CDKN1C | G1_to_S_cell_cycle_Reactome |
| | | PAK1 | Integrin mediated_cell_adhesion_KEGG |
| | | ECT2 | ----- |
| | | RALGPS2 | ----- |
| | | SFN | Calcium_regulation_in_cardiac_cells |
| | | | Smooth_muscle_contraction |
| GO:0019207 | Kinase Regulator Activity | CDKN1C | G1_to_S_cell_cycle_Reactome |
| | | SFN | Calcium_regulation_in_cardiac_cells |
| | | | Smooth_muscle_contraction |
| GO:0004857 | Enzyme Inhibitor Activity | TIMP3 | Matrix_Metalloproteinases |
| | | CDKN1C | G1_to_S_cell_cycle_Reactome |
| | | SFN | Calcium_regulation_in_cardiac_cells |
| | | | Smooth_muscle_contraction |
| GO:0019887 | Protein Kinase Regulator Activity | CDKN1C | G1_to_S_cell_cycle_Reactome |
| | | SFN | Calcium_regulation_in_cardiac_cells |
| | | | Smooth_muscle_contraction |
| GO:0019210 | Kinase Inhibitor Activity | CDKN1C | G1_to_S_cell_cycle_Reactome |
| | | SFN | Calcium_regulation_in_cardiac_cells |
| | | | Smooth_muscle_contraction |
| GO:0004860 | Protein Kinase Inhibitor Activity | CDKN1C | G1_to_S_cell_cycle_Reactome |
| | | SFN | Calcium_regulation_in_cardiac_cells |
| | | | Smooth_muscle_contraction |
| GO:0051018 | Protein Kinase A Binding | AKAP12 | G_Protein_Signaling |
| GO:0008179 | Adenylate Cyclase Binding | AKAP12 | G_Protein_Signaling |

# 5 References


[1] L. Hunter, "Artificial intelligence and molecular biology," in Proceedings of the tenth national conference on Artificial intelligence, 1992, pp. 866-868.

[2] G. S. Michaels, D. B. Carr, M. Askenazi, S. Fuhrman, X. Wen, and R. Somogyi, "Cluster analysis and data visualization of large-scale gene expression data," in Pacific symposium on biocomputing, 1998, pp. 42-53.

[3] N. Speer, C. Spieth, and A. Zell, "A memetic co-clustering algorithm for gene expression profiles and biological annotation," in Evolutionary Computation, 2004. CEC2004. Congress on, 2004, pp. 1631-1638.

[4] D. W. Mount, "Sequence and genome analysis," New York: Cold Spring, 2004.

[5] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," Nature genetics, vol. 22, pp. 281-285, 1999.

[6] F. Luo, K. Tang, and L. Khan, "Hierarchical clustering of gene expression data," in Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on, 2003, pp. 328-335.

[7] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proceedings of the National Academy of Sciences, vol. 95, pp. 14863-14868, 1998.

[8] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, et al., "Large-scale temporal gene expression mapping of central nervous system development,"

Proceedings of the National Academy of Sciences, vol. 95, pp. 334-339, 1998.
[9] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, et al., "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," Proceedings of the National Academy of Sciences, vol. 96, pp. 2907-2912, 1999.
[10] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén, "Analysis of gene expression data using self-organizing maps," FEBS letters, vol. 451, pp. 142-146, 1999.
[11] J. He, A.-H. Tan, and C.-L. Tan, "Self-organizing neural networks for efficient clustering of gene expression data," in Neural Networks, 2003. Proceedings of the International Joint Conference on, 2003, pp. 1684-1689.
[12] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," Proceedings of the National Academy of Sciences, vol. 97, pp. 262-267, 2000.
[13] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," Journal of computational biology, vol. 7, pp. 601-620, 2000.
[14] P. J. Woolf and Y. Wang, "A fuzzy logic approach to analyzing gene expression data," Physiological Genomics, vol. 3, pp. 9-15, 2000.
[15] H. Wang, F. Azuaje, and O. Bodenreider, "An ontology-driven clustering method for supporting gene expression analysis," in Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on, 2005, pp. 389-394.
[16] I. Holmes and W. J. Bruno, "Finding regulatory elements using joint likelihoods for sequence and expression profile data," in Ismb, 2000, pp. 202-210.
[17] Y. Barash and N. Friedman, "Context-specific Bayesian clustering for gene expression data," Journal of Computational Biology, vol. 9, pp. 169-191, 2002.
[18] J. Kasturi, R. Acharya, and M. Ramanathan, "An information theoretic approach for analyzing temporal patterns of gene expression," Bioinformatics, vol. 19, pp. 449-458, 2003.
[19] G. Garai and B. Chaudhuri, "A novel genetic algorithm for automatic clustering," Pattern Recognition Letters, vol. 25, pp. 173-187, 2004.
[20] K. Ushizawa, C. B. Herath, K. Kaneyama, S. Shiojima, A. Hirasawa, T. Takahashi, et al., "cDNA microarray analysis of bovine embryo gene expression profiles during the pre-implantation period," Reproductive Biology and Endocrinology, vol. 2, p. 77, 2004.
[21] N. Bolshakova, F. Azuaje, and P. Cunningham, "An integrated tool for microarray data clustering and cluster validity assessment," Bioinformatics, vol. 21, pp. 451-455, 2005.
[22] T.-S. Chen, T.-H. Tsai, Y.-T. Chen, C.-C. Lin, R.-C. Chen, S.-Y. Li, et al., "A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray," in Intelligent Signal Processing and Communication Systems, 2005. ISPACS 2005. Proceedings of 2005 International Symposium on, 2005, pp. 405-408.
[23] F. Azuaje, J. Dopazo, and J. Wiley, Data analysis and visualization in genomics and proteomics: Wiley Online Library, 2005.
[24] H. Okayama, T. Kohno, Y. Ishii, Y. Shimada, K. Shiraishi, R. Iwakawa, et al., "Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas," Cancer research, vol. 72, pp. 100-111, 2012.
[25] W. Pao and N. Girard, "New driver mutations in non-small-cell lung cancer," The lancet oncology, vol. 12, pp. 175-180, 2011.
[26] R. S. Herbst, J. V. Heymach, and S. M. Lippman, "Lung Cancer," The New England Journal of Medicine, vol. 319, pp. 1367 - 1380, 2008.
[27] F. Janku, D. J. Stewart, and R. Kurzrock, "Targeted therapy in non-small-cell lung cancer—is it becoming a reality?," Nature Reviews Clinical Oncology, vol. 7, pp. 401-414, 2010.
[28] G. Bronte, S. Rizzo, L. La Paglia, V. Adamo, S. Siragusa, C. Ficorella, et al., "Driver mutations and differential sensitivity to targeted therapies: a new approach to the treatment of lung adenocarcinoma," Cancer treatment reviews, vol. 36, pp. S21-S29, 2010.
[29] D. E. Gerber and J. D. Minna, "ALK inhibition for non-small cell lung cancer: from discovery to therapy in record time," Cancer cell, vol. 18, pp. 548-551, 2010.
[30] (2013, 06/18/2013). Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31210
[31] Q. Zheng and X.-J. Wang, "GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis," Nucleic acids research, vol. 36, pp. W358-W363, 2008.